CGIAR | Independent Advisory and Evaluation Service

Applying the CGIAR Quality of Research for Development Framework to Process and Performance Evaluations

CGIAR EVALUATION GUIDELINES 2022

Citation

Independent Advisory and Evaluation Service (IAES). (2022). CGIAR Evaluation Guidelines. Applying the CGIAR Quality of Research for Development Framework to Process and Performance Evaluations. (Beta version) Rome: Independent Advisory and Evaluation Service.

Design and layout

Khania Curtis Design and Creative Direction

Name of Guidelines:

Applying the CGIAR Quality of Research for Development Framework to Process and Performance Evaluations (Beta version): CGIAR Evaluation Guidelines 2022

Purpose:

To harmonize the overall approach for evaluating research and science quality in CGIAR and similar contexts; to facilitate a common understanding of the Quality of Science (QoS) evaluation criterion and the methods.

Audiences:

The primary audience is evaluators, evaluation managers and commissioners involved in evaluating Quality of Science in CGIAR. The secondary audience is those responsible for providing inputs into evaluations in the CGIAR system. Users in other agricultural research-for-development (AR4D) contexts may find this document useful too.

Framework and Policy Reference:

This guidance supports the CGIAR Evaluation Framework and the CGIAR Evaluation Policy (2022) and should be read in conjunction with other evaluation-related guidelines. The framework and policy were guided by the Quality of Research for Development (QoR4D) framework (2017, revised in 2020).

Contact:

For queries and feedback about learning from the roll-out and application of this **Beta version** contact the Evaluation Function within the Independent Advisory and Evaluation Service (IAES) of CGIAR at IAES-Evaluations@cgiar.org

Acknowledgments

The Evaluation function of the Independent Advisory and Evaluation Services (IAES) co-developed this guidelines document in consultation with many stakeholders (see Annex 1). From IAES, Svetlana Negroustoueva, Evaluation Function Lead, and Stefania Sellitti, evaluation consultant, led the process under the overall direction of Allison Grove Smith, Director of IAES. These guidelines draw on the practical experiences of subject matter experts involved in the TOR design and evaluation of CGIAR Research Programs, specifically Paolo Sarfatti, Jillian Lenne, and John Morton. Rich input was provided by participants in the validation meeting and through the EvalForward Community of Practice discussion, and CGIAR MEL CoP, with thanks to these groups' respective organizers for providing space for the co-design process. The authors would also like to acknowledge peer reviews conducted by Rachel Bedoiun, Donna Podems, and Marina Apgar, and contributions by Sara Vaca, evaluation and visualization consultant. The work also benefitted from the contributions of ISDC members Holger Meinke and Andrew Ash, as well as the IAES Evaluation Reference Group members Zenda Ofir and Guy Poppy. We further acknowledge the work of Science Metrix, Elsevier: the Technical note made a deep contribution to the associated thinking. The broader IAES team in Rome helped ensure smooth editorial and design support, technical assistance, and administrative processes related to the development of these guidelines.

Soliciting Input and Feedback

The IAES would like to receive feedback about learning from the roll-out and application of this Beta version of the Guidelines and invites users to contact IAES/Evaluation (IAES-Evaluation@cgiar.org) as custodian of the document. The next version will continue building on direct experiences and evolving industry standards on how to evaluate science quality to best ensure effective and cost-efficient evaluations, including in CGIAR.

Contents

List of Text Boxes

Figures

Tables

Abbreviations

AR4D	Agricultural Research for Development
CIFOR	Center for International Forestry Research
CRP	CGIAR Research Program (prior to 2022)
DAC	Development Assistance Committee
DCM	Data Collection Matrix
EA	Evaluability Assessment
EF	Evaluation Framework
EP	Evaluation Policy
ERG	Evaluation Reference Group
FAO	Food and Agriculture Organization of the UN
FGD	Focus Group Discussion
GDI	Gender, Diversity, and Inclusion
ICARDA	International Center for Agricultural Research in the Dry Areas
IUCN	International Union for Conservation of Nature
KII	Key Informant Interviews
M&E	Monitoring and Evaluation
MARLO	Managing Agricultural Research for Learning and Outcomes
MEL/IA	Monitoring, Evaluation and Learning/Impact Assessment
MELCOP	Monitoring, Evaluation and Learning Community of Practice
OECD	Organisation for Economic Co-operation and Development
PCU	Project Coordination Unit
PPU	Performance Portfolio Unit
PRMF	Performance and Results Management Framework
PRMS	Performance and Results Management System
QA	Quality Assurance
QoR4D	Quality of Research for Development
QoS	Quality of Science
SDGs	Sustainable Development Goals
SMEs	Subject-matter experts
ToC	Theory of Change
ToR	Terms of Reference
TRA	Technical Reporting Arrangement

1 Introduction

THESE EVALUATION GUIDELINES ARE INTENDED TO HARMONIZE THE OVERALL APPROACH FOR EVALUATING RESEARCH AND SCIENCE QUALITY IN CGIAR. THEY ARE NEEDED TO GUARANTEE A COMMON UNDERSTANDING OF THE EVALUATION CRITERION AND THE AVAILABLE METHODS AMONG SUBJECT-MATTER EXPERTS AND EVALUATORS CONDUCTING EVALUATIONS.

AS SUCH, THE GUIDELINES PROVIDE A MENU OF METHODS FOR EVALUATORS AND TRANSPARENCY FOR USERS OF EVALUATIONS, INCLUDING MONITORING PROFESSIONALS, TO INTERPRET FINDINGS AND OPERATIONALIZE A QUALITY OF SCIENCE (QOS) EVALUATION CRITERION.

Rigorous and independent process and performance evaluations play a critical role in helping CGIAR inform the design and implementation of its research and innovations. Evaluations provide actionable evidence for management and governance decisions, facilitate learning and ensure accountability to funders and other stakeholder groups. CGIAR's Evaluation Framework (EF) and Evaluation Policy (EP) (2022) articulate how process and performance evaluations support CGIAR to deliver its mission and implement its 2030 Research and Innovation Strategy (CGIAR 2030 Strategy).[1]

CGIAR's core business is delivering research for development. This guideline for evaluators and subject-matter experts explains the drivers, parameters, and methods to evaluate CGIAR research and science quality.

These evaluation guidelines build on the Quality of Research for Development (QoR4D) frame of reference (2020) and were informed by the Research Excellence Framework (REF), and the Research Quality Plus (RQ+) Assessment Instrument. The CGIAR Evaluation Framework and Policy (2022) and CGIAR's Performance and Results Management Framework 2022-2030 established the foundation for this guidance in CGIAR context.

1.1 A Changing Context for Evaluations in CGIAR

CGIAR is a global research partnership for a food-secure future with a mission to deliver science and innovation to transform food, land, and water systems in a climate crisis. As part of the One CGIAR reform process,[2] the CGIAR 2030 Strategy strives for transformative change across three action areas (Systems Transformation, Resilient Agrifood Systems, and Genetic Innovation), and five Sustainable Development Goal (SDG)-focused impact areas, delivered through more than 30 regional and global initiatives.

The CGIAR 2030 Strategy[3] uses the following definitions:

Research – Generation and communication of data, information and knowledge on an empirical basis.

Science – Rigorous theory-based research.

The Performance and Results Management Framework (PRMF) supports the CGIAR 2030 strategy implementation. It provides the basis for CGIAR accountability, learning, communication and resource mobilization, and serves as the basis for the Technical Reporting Arrangement (2022) (Figure 2).

[1] The CGIAR Evaluation Policy is to be revised from time to time. CGIAR's Independent Advisory and Evaluation Service may recommend amendments as appropriate to the System Council and Board for approval. Future revisions may, for instance, consider if the system transformation agenda of CGIAR may be best served through additional evaluation criteria.

[2] One CGIAR is a process undertaken by CGIAR which started in 2019 to strengthen its partnerships, knowledge, assets and global presence, aiming for greater integration and impact in the face of the interdependent challenges facing today's world. https://www.cgiar.org/food-security-impact/one-cgiar/

[3] CGIAR 2030 Research and Innovation Strategy

1.2 Deploying Evaluation Criteria for Evaluating Research and Science Quality

Evaluating CGIAR research and science quality is important for three main reasons: (1) to provide accountability for public and private investment in research that generates international public goods; (2) to inform funders about the quality both of the scientific processes and the scientific outputs of research conducted; (3) to provide evidence about how CGIAR science contributes towards the goals of the organization, including as part of a wider effort, to contribute to accelerated progress toward the SDGs.

For the specific case of CGIAR, its core business rests in research for development (R4D). The CGIAR 2030 Strategy defines this as both research and science. CGIAR funders focus on development outcomes. Consequently, the evaluation of CGIAR interventions must respond to both the QoR4D – research oriented to deliver development outcomes – and OECD/DAC – development orientation – frameworks. The standard development assistance OECD/DAC evaluation criteria are insufficient to evaluate the core business of the CGIAR. In the evaluation planning and approach, the special characteristics of R4D need to be considered. In particular, the unpredictable and risky nature of research and the long time it takes to witness outcomes is considered throughout this document.

The methods and criteria for evaluating CGIAR respond to **two foundational frameworks** crucial to CGIAR and its stakeholders, which form the basis for evaluating research and science quality (Figure 1).

Figure 1: *Two foundational frameworks for process and performance evaluations in CGIAR*

Quality of Research for Development Frame of Reference (QoR4D) (2020)

Entities across CGIAR engaged to co-develop the QoR4D framework that establishes the connective elements that link the CGIAR research effort with development outcomes.

The QoR4D framework prompted system-wide agreement on the nature and evaluation of CGIAR's research quality and the system-wide agreement on the nature and assessment of the Quality of science, a concept broadened beyond scientific credibility to include the likelihood of achieving development outcomes. QoR4D framework consists of four key elements: relevance, effectiveness, legitimacy and credibility.

QoR4D was designed to assist CGIAR in research and portfolio design, monitoring, evaluation, and performance management.

OECD Development Assistance Committee (OECD/DAC) (2019)

A global standard, the 2019 OECD Development Assistance Committee (OECD/DAC) provides a normative framework used to determine the merit or worth of an intervention, they serve as the basis upon which evaluative judgments are made.

To use of the six OECD/DAC evaluation criteria (relevance, effectiveness, coherence, efficiency, impact, and sustainability) OECD recommends the two principles: (1) The criteria should be applied thoughtfully to support high-quality, useful evaluation; and (2) The use of the criteria depends on the purpose of the evaluation. OECD encouraged reading the definitions in conjunction with other principles and guidance on how to conduct evaluations in ways that will be useful and of high quality in a particular context of the intervention being evaluated.

Responsive to both the QoR4D and OECD/DAC frameworks, CGIAR's seven evaluation criteria frame and provide structure to the substantive focus of evaluation questions (Annex 2; CGIAR Evaluation Policy 2022).

A Quality of Science (QoS) evaluation criterion has been used in CGIAR evaluations for at least the past decade (see 2012 evaluation policy)[4]. Guided by the QoR4D framework, the QoS evaluation criterion was applied in the 2020 evaluative reviews of the 12 CGIAR Research Programs (CRPs) coupled with the effectiveness criterion. Mixed methods (quantitative and qualitative) were applied to address both evaluation criteria.[5]

[4] (retired) CGIAR Policy for Independent External Evaluation (2012) replaced by 2022 version.

[5] For additional detail consult the CRP full reports and annexes for individual evaluative reviews of CRPs and/or read the four case studies in the technical note co-developed by Science-Metrix and CAS/Evaluation (2022).

1.3 Rationale and Objectives of the Guidelines

The two foundational frameworks, QoR4D and OECD/DAC, have synergies and complementarities, as well as overlapping terminologies and concepts. These synergies and overlaps provide **the rationale** for this guideline, which sets out to establish how to evaluate research and science that is conducted for development and contributes to SDGs.

The guidelines focus on operationalizing the Quality of Science (QoS) evaluation criterion and deploying the seven CGIAR evaluation criteria in combination to make evaluative judgments about research-for-development activities. The targeted evaluation of QoS is essential to verify whether the science produced is credibly (robust) and legitimately (fair and ethical) produced through multi-stakeholder cooperation and aligns with the needs of people on the ground.

The main objectives of this Evaluators' Guide: Applying the CGIAR Quality of Research for Development Framework to Process & Performance Evaluation are to:

- Facilitate a common understanding of the QoS evaluation criterion including in relation to other evaluation criteria.
- Outline a common approach to evaluating research and science and provide a menu of methods based on a critical review of their strengths and challenges.
- Cross-reference between ISDC ex-ante measures and CGIAR Evaluation Policy measures for midline/ex-post evaluation.
- Underline the roles and responsibilities to facilitate evaluating QoS in CGIAR at different levels in line with Evaluation Framework principles of 'Measurability' and 'Mutual Accountability' and underscoring Pillar 2 (Holistic and Consistent CGIAR-wide approach to monitoring and learning) of the CGIAR Evaluation Policy (Figure 2).

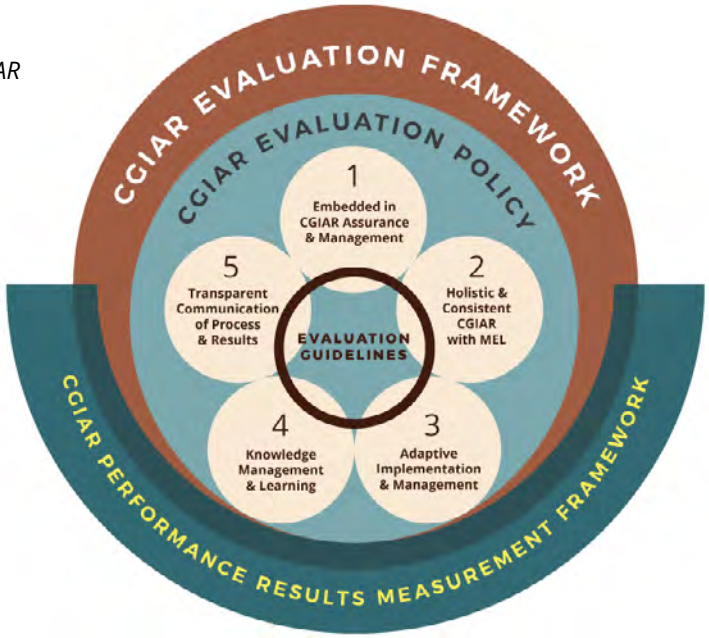


Figure 2: *Five pillars of the CGIAR Evaluation Policy*

1.4 Audience and Users

These guidelines are intended primarily for evaluators, evaluation managers and commissioners involved in process and performance evaluations of research and science in CGIAR. They will also be useful for those responsible for providing inputs into evaluations in the CGIAR system, for example, research managers and MEL professionals.

The described approaches are sufficiently broad to be used across the many contexts within as well as outside CGIAR. They are suitable to be adapted as needed by CGIAR partners and like-minded agencies conducting research for development.

Familiarity with the CGIAR Evaluation Framework and Evaluation Policy is a prerequisite to understand and use these guidelines.[7] The guidelines are launched at the outset of the CGIAR business cycle and Technical Reporting Arrangement (TRA, 2022) towards conducting process and performance evaluations of the new CGIAR portfolio. To consider all QoR4D elements, a performance and process evaluation of an R4D intervention must be sufficiently resourced to address, at a minimum, the relevance, effectiveness, and Quality of Science (legitimacy and credibility) criteria.[6]

[6] The 'Guidelines on Conducting and Using Evaluability Assessments in CGIAR' (2022) and Management Engagement and Response guidelines (under development) aim to facilitate attention to such key elements as MEL–related resourcing and infrastructures, I.e. performance results management systems.

[7] An additional source for learning about the CGIAR Evaluation Policy & Framework https://youtu.be/DeUn7T3UW6s (video).

2 What Frames Evaluation of Quality of Science in CGIAR?

2.1 The Accountability and Measurement Parameters

The 2030 Research and Innovation Strategy emphasizes the critical role of science and innovation in providing new evidence, insights and solutions that feed into strategic alliances for change. The supporting Performance and Results Management Framework 2022-2030 (PRMF) lays out CGIAR's intent to measure its contributions from research to impact along **three main pathways**: (1) targeted capacity development, (2) policy advice and (3) science-based innovations and technologies (CGIAR 2030 Strategy (page 20); Table 1). The PRMF, operationalized in part through the TRA (2022), will steer and inform decision-making around research initiatives stage gating in the project cycle; i.e., monitoring and learning processes to help initiative teams dynamically manage their programs during delivery, through the evidence-based validation and/or adaptation of theories of change; to aid management oversight of portfolio implementation; and to support investment decisions.

Thus, an important role for evaluation is understanding whether and how the above-mentioned innovation types and other outputs exhibit high-quality science and contribute towards pathways to development outcomes.

Table 1: *Types of Innovations*

Innovations: Packages of complementary contributions needed to develop and take to scale products, services and solutions.

Capacity development	Policies	Technologies
The know-how and capacities of individuals, firms, organizations, and networks to design, test, validate and use innovations.	**Public policy, legislation, public and private delivery and business strategies that create an enabling environment in which innovations can move to scale, or which in themselves represent innovations that can lead to impact.**	**The varieties, machines, management practices, products and tools - including big data and information tools - whose use can lead to benefits, gains or efficiencies and whose deployment at scale can lead to impact.**

Source: PRMF 2022-2030, Page 5

2.2 Evaluation Framework and Policy

In response to the growing pace of change to support the CGIAR's evolving needs and demands, the CGIAR System Board (23rd Session) and System Council (15th Meeting) approved the new fit-for-purpose CGIAR Evaluation Framework and revised Evaluation Policy in 2022. The Framework and Policy define and set out 15 standards and principles, as well as the overall approach to process and performance evaluations in CGIAR.

Process evaluations are evaluations of the organizational functioning, instruments, mechanisms, and management practices of institutional and procedural issues across CGIAR and assessments of experience with CGIAR frameworks, policies, criteria and procedures.

Performance evaluations provide rigorous and impartial assessments of organizational effectiveness and operating models by assessing progress toward the achievement of outcomes or processes by comparing performance data with the stated objective and reporting back on a predetermined schedule, to inform decision-making about how to best use or invest financial or technical resources, resolve challenges and support ongoing progress.

Aligned to the OECD/DAC framework, the seven evaluation criteria provide structure to the substantive focus of evaluation questions (Text Box 1, (Annex 3, CGIAR Evaluation Policy 2022).

Text Box 1: *Seven CGIAR evaluation criteria* (See Annex 2 for more information)

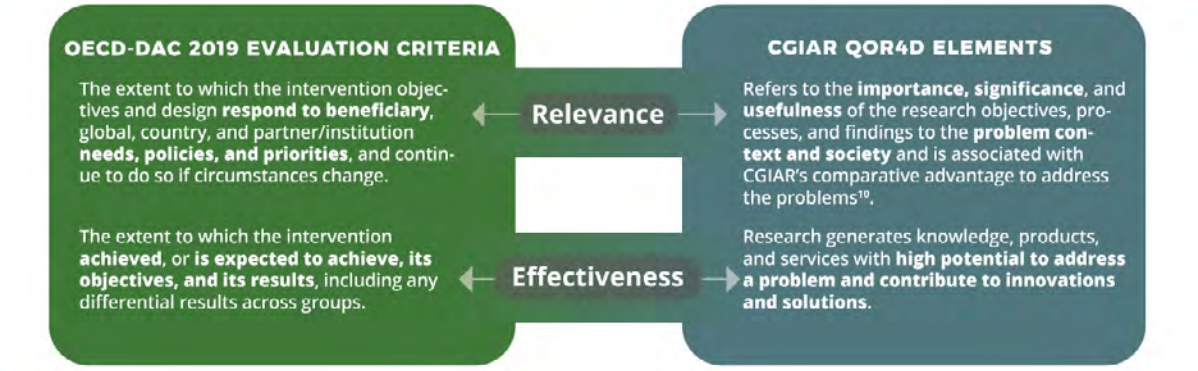
- Relevance
- Coherence
- Effectiveness
- Efficiency
- **Quality of Science**
- Sustainability
- Impact

2.3 Quality of Research for Development Frame of Reference

The Quality of Research for Development (Qo4RD) frame of reference aims to bring coherence and enhance the overall quality of R4D within the CGIAR portfolio, recording system-wide agreement on the nature and assessment of research quality. It encompasses relevance to user groups, scientific credibility, legitimacy and the likelihood of achieving development outcomes, distilled into four key elements: Relevance, Scientific Credibility, Legitimacy and Effectiveness.

Relevance and **Effectiveness** are defined in the QoR4D frame slightly differently from the OECD/DAC evaluation criteria. In CGIAR these two criteria use the same terms and speak directly to QoR4D elements and OECD/DAC criteria (Figure 1). CGIAR evaluation criteria are responsive to the definitions per QoR4D and aligned with OECD/DAC for ex-post evaluations, for example, a process and performance evaluation[8] of effectiveness must look at both fidelity to plan (per OECD/DAC) and readiness for use (per QoR4D). Relevance criterion for OECD/DAC pertains to the responsiveness to the development challenge, whereas in QoR4D relevance element also incorporates an association with CGIAR's comparative advantage. Table 2 presents the differences in the definitions of relevance and effectiveness criteria between the two frameworks (see also Figure 1).

Table 2: *Considerations around 'Relevance' and 'Effectiveness* [9]



The **Quality of Science (QoS) evaluation criterion** assures evaluative coverage of the other two QoR4D elements (**legitimacy and scientific credibility**), which are not explicit OECD/DAC criteria but vital in the QoR4D frame (Figure 1). As one of seven criteria, the **QoS evaluation criterion** reflects the identity of CGIAR as a global research-for-development partnership. The single QoS evaluation criterion rests on using mixed methods to judge the degree and extent of legitimacy and scientific credibility.

Mapping of QoR4D elements and ISDC review criteria to the evaluation standards and principles in Annex 5 illustrates alignment between Evaluation and QoR4D frameworks, for the specific purpose of conducting performance and process evaluations in CGIAR.

[8] Definitions can be found in section 2.2

[9] Consistent with the Evaluation Policy, other evaluation criteria (Efficiency, Coherence, Sustainability, Impact) follow OECD-DAC definitions (see Annex 2).

[10] Independent Science for Development Council. 2022. Identifying and Using CGIAR's Comparative Advantage. Rome: CGIAR Independent Advisory and Evaluation Service. https://iaes.cgiar.org/sites/default/files/pdf/ISDC-Technical-Note-Identifying-and-Using-CGIAR-Comparative-Advantage.pdf

3 Evaluating Quality of Science: Dimensions, Questions and Timing

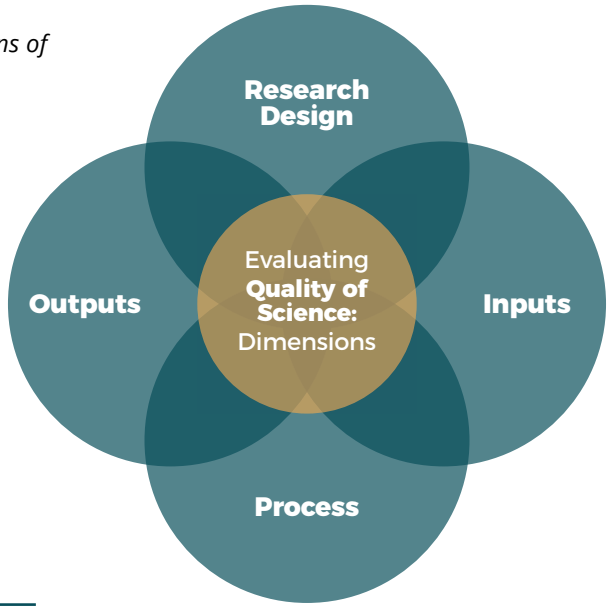
The context and timing of evaluation and the characteristics of the objective of the evaluation (evaluand) will dictate the overall evaluation approach and selection of methods. Consistent with the **key evaluation types** in the CGIAR Evaluation Policy, focused evaluations of Quality of Science (QoS) are most likely to be conducted as part of a larger process and/or performance evaluation (Sections 2.2 and 2.3), e.g. the 2020 CRP Reviews which used effectiveness and QoS[11] evaluation criteria.

3.1 Four Dimensions to Evaluating Quality of Science and Mapping to Other Evaluation Criteria

Toward comprehensive evaluation of the Quality of Science, this guideline recommends analysis of four interlinked dimensions – Research Design, Inputs, Processes, and Outputs (Figure 3).[12]

- **Research Design:** The appropriateness of the research design as implemented is judged in terms of commonly accepted standards in a designated subject-matter field. Assessment of the technical appropriateness of the research agenda and strategy, and overall relevance and coherence.
- **Inputs:** The necessary inputs of research are assessed for their adequacy in relation to outputs. Examples of inputs include research staff, team compositions, availability of adequate research infra structure and funding.
- **Processes:** Management and coordination are driven by incentives for achieving and maintaining the high scientific credibility of outputs. Further, the evaluation explores the effectiveness of building and leveraging partnerships, i.e., whether based on mutual understanding, trust, and commitment, with a clear recognition of various perspectives, needs, roles and contributions. Fairness and the ethical aspects of actual research portfolio implementation are also assessed. For example, to what extent did research and processes consider the implementation of ethics guidelines and management decision processes and if they were representative.
- **Outputs:** Outputs will vary considerably; however, they are most often tangible products or services. For example, new seeds or germplasm, or technical outputs such as policy documents, journal articles, technical briefs and new soil management tech niques. Other technical outputs include software outputs, guidelines, decision support tools and training materials, policy briefs and other policy-change-oriented actions.

Figure 3: *Four interlinked dimensions of evaluating Quality of Science*



[11] Terms of Reference (and Addendum) for 2020 CRP Reviews

[12] CGIAR adapts the concept of 'Research Rigor' from the International Development Research Centre's RQ+ Assessment Instrument (IDRC - International Development Research Centre 2022) for CGIAR needs, framing it in a way that can be considered also in lack of specific outputs.

Table 3 maps the four QoS dimensions to the seven evaluation criteria: it demonstrates how the four dimensions 'fit' to seven evaluation criteria, which frame how evaluative judgments are made to answer core evaluative questions in project and performance evaluations of the CGIAR portfolio. This means that the dimensions suggest evaluation questions that can be deployed in association with multiple criteria, so there are multiple pathways for evaluative exploration of science. The suggested questions by the evaluation criteria are further elaborated in Step 2 and Annex 3.

Table 3: *Sample mapping of seven evaluation criteria to four Quality of Science dimensions*

EVALUATION CRITERIA								
Quality of Science			Relevance	Coherence	Efficiency	Effectiveness	Sustainability	Impact
DIMENSIONS	Design	X	X	X				X
	Input	X	X	X	X			
	Process	X			X	X	X	X
	Outputs	X	X				X	X

Across all QoS dimensions, attention to evaluation standards and principles is required (Annex 5). For example:

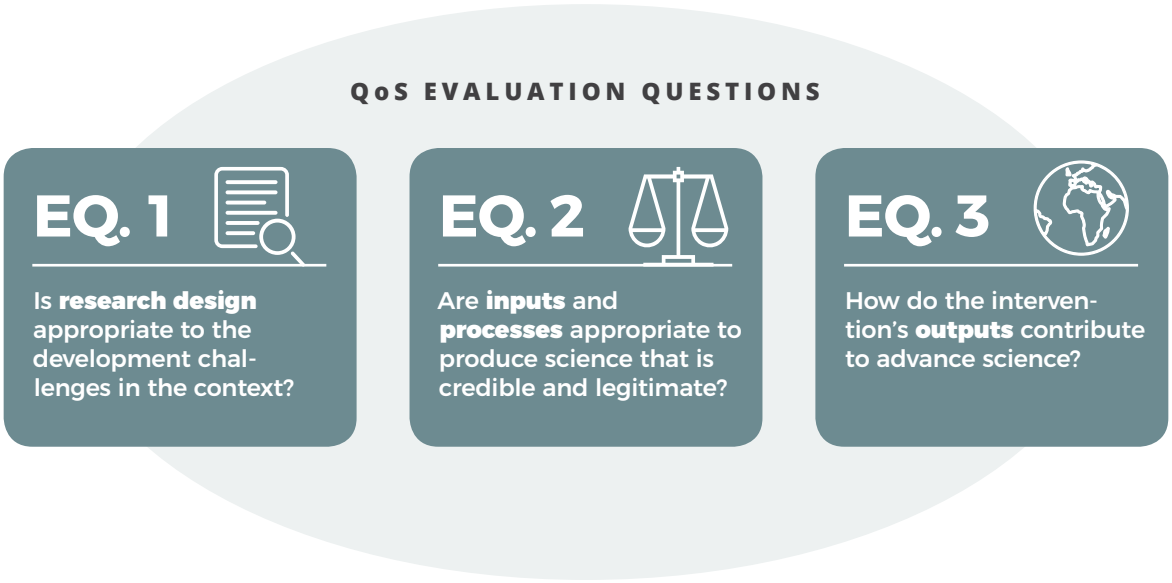
- The consideration of the principle of *Gender, Diversity, and Inclusion* (GDI) would be suitable in design, inputs and processes (2022, EF). Evaluation design and conduct, the commissioning of teams and the reporting strive to fully address GDI parameters. Evaluations will consider who is engaged in the work and who benefits from it. The evaluation would consider the composition of a research team in terms of gender, nationality, age, and discipline diversity, as well as how involved women, researchers from LIMCs, and local communities were in the design and delivery of the research.
- The principles of *Legitimacy* and participation, and *Transparency* address the increasingly critical role of partnerships to facilitate good research design, process and outputs. Evaluations interrogate perceptions and practices, by inviting relevant informants and using consultative processes to design evaluation questions and select methods. They also ensure the process of delivering science including, where appropriate and feasible, representatives of end and intermediate users of outputs.

Evaluation inception and final evaluation reports would cover how the standards and principles are addressed, and potential and actual limitation and mitigation considerations.

3.2 Key Quality of Science Evaluation Questions

The following three key evaluation questions (EQ) are recommended, especially when using a designated QoS evaluation criterion (Figure 4 and Table 3).

Figure 4: *Three evaluation questions to operationalize the Quality of Science evaluation criterion*



EQ.1 captures ***design***, how appropriate and responsive it is to addressing development challenges. Assessment and judgment are made in terms of commonly accepted standards in a designated subject-matter field, technical appropriateness of the research agenda and strategy. This dimension overlaps with the relevance and coherence evaluation criteria towards impact and it touches on the credibility element (see Table 2). The evaluative design should describe how the QoS and relevance criteria will be interwoven. This question provides the evaluation team a space to look deeply at the contextualization of the scientific endeavor.

EQ.2 considers the quality of ***inputs and processes***. It is meaningful only if measured against clear industry benchmarks, and/or CGIAR quality standards or targets (e.g., gender diversity among scientific staff, requirements for infrastructure and materials). The credibility results from the appropriateness of ***inputs***; and the ***process*** confers legitimacy. EQ.2 mostly overlaps with the effectiveness evaluation criterion, touching on the legitimacy element.

EQ.3 assesses the contribution of ***output(s)*** to a research field, domain, discipline or transdisciplinary grouping of these. Answering EQ.3 requires an understanding of the scope of the scientific endeavor given the interdisciplinary and transdisciplinary approaches often used to deliver CGIAR mission. Evaluation teams, and their designs, need to approach EQ.3 with an understanding that system transformation requires transdisciplinary ways of working considering that many different disciplines may be involved in an intervention that leads to an output(s). This may overlap with the efficiency and coherence evaluation criteria.

3.3 Timing of Evaluating Quality of Science in CGIAR

Selected outputs and most societal outcomes of research have been shown to typically take five, ten, or even twenty years to be fully realized (Langfeldt, 2015). This is specifically the case for technology-related research, i.e., plant breeding. However, increasingly it appears that systems-type research is geared toward assessing and achieving impacts in real time. The timing of evaluating QoS in the context of process and performance evaluation would depend on the kind of intervention and on the dimensions to be evaluated, as well as the overall project cycle. Different methods and indicators might be calculated at different points during or beyond the duration of an intervention. For example, it is often useful to assess inputs and processes before the next phase of an intervention, to support adaptive management and course correction. Therefore, a comprehensive QoS evaluation, incorporating all four dimensions is best at the end of a project cycle of an AR4D-type intervention.

Figure 5 provides an example of CGIAR intervention with three-year funding cycles that are designed to deliver outcomes over three consecutive phases (as is the case with the CGIAR pooled funding initiatives).

Figure 5: *Typical intervention cycle for CGIAR**

Phases	Year 0	Year 1	Year 2	Year 3	Year 4
Intervention Design					
Implementation					
Monitoring					
Process/performance Evaluation					

**Publishing of related publications, and associated citation impact realization windows, applied to 44 CGIAR Initiatives and Programs (2022).*

A three-pronged approach is recommended for evaluation's timeliness and robustness with the following considerations:

- **Year 2:** Evaluability Assessment or mid-term evaluation would balance qualitative and quantitative methods, to assess input and process dimensions. Where resources allow, an interim or mid-term evaluation would identify practices that could be improved, allowing for the adaptive management of inputs and practices in ongoing processes. Most quantitative output indicators can only be calculated after a longer timeframe (year 3-8). However, an indicator like the H-index could be used already at this stage as a measure of the productivity of scientists involved in the intervention.
- **Year 3-5:** End-line (process/performance) evaluation would also prioritize qualitative methods to assess input and process dimensions and quantitative indicators, e.g., a restricted set of bibliometric indicators.
- **Year 8:** Comprehensive and targeted mixed methods evaluation of a portfolio of projects, to capture realized mid-term societal outcomes (e.g., uptake of publications in policy-related documents, international public goods), or accrued citation impact of transformative research articles.

Figure 6 provides suggestions regarding the process and performance evaluation of QoS dimensions by evaluation criteria, depending on the project cycle and overall duration of R4D projects. However, changes to the suggested timelines are possible if a specific intervention allows for an earlier evaluation of selected dimensions, i.e., design at the mid-term. Notably, the boxes in blue point to the evaluation of QoS. Light blue includes the aspects that can be evaluated before the end of the intervention, while dark blue for the aspects that need to be evaluated after the end of the intervention.

Therefore, a comprehensive QoS evaluation, incorporating all four dimensions is best at the end of a project cycle of an AR4D-type intervention in CGIAR.

Figure 6: *Timeliness to evaluate each criterion*

Evaluation criteria/ Project cycle	Year 0	Year 1	Year 2	Year 3	Year 4
Relevance					
Coherence					
Effectiveness					
Efficiency					
Sustainability					
Impact					
Quality of Science					

3.4 Quality Assurance

To ensure the robustness and credibility of an evaluation, embedded within the CGIAR Evaluation Framework and aligned to the standard evaluation process, **a multilayered quality assurance (QA) system** should be followed throughout.

Triangulation is a method used to increase the credibility and validity of evaluative findings and is essential in evaluation design, implementation and QA. Different types of triangulation are appropriate at different steps of evaluating quality of science.

(1) investigator (evaluation team members) triangulation, defined as the use of different experts in data gathering and the analysis process.

Text Box 2: *Embedding quality assurance into the evaluation process*

For robustness and credibility of an evaluation, a diversified team of multi-disciplinary experts is recommended. Each team should be composed of a team leader (preferably, a professional evaluator) and subject-matter experts (SMEs). The latter are thought leaders in domains relevant to the CGIAR 2030 Research and Innovation Strategy. This team, complemented by rigorous peer review and other quality assurance procedures, ensures the credibility of the evaluation process and results (Section 4).

In addition to a diversified team of SMEs in the evaluation team, external peer reviewers that are experts in the topic to be evaluated but are external to the evaluation team, should also be included. They will be involved in at least two steps of the evaluation process. Complementing review by IAES/Evaluation, evaluation peer-reviewers check the choice of methodology for quality and technical soundness and review the inception and draft evaluation reports. They also review the module or component or other core intermediary deliverables. Issues raised during QA steps need to be mitigated to a degree possible by the SMEs and evaluation team leader with clear documentation. (Step 1)

(2) theory triangulation, which implies the use of multiple approaches during the evaluation.

(3) methodological triangulation, namely the use of more than one method to gather data, where possible.

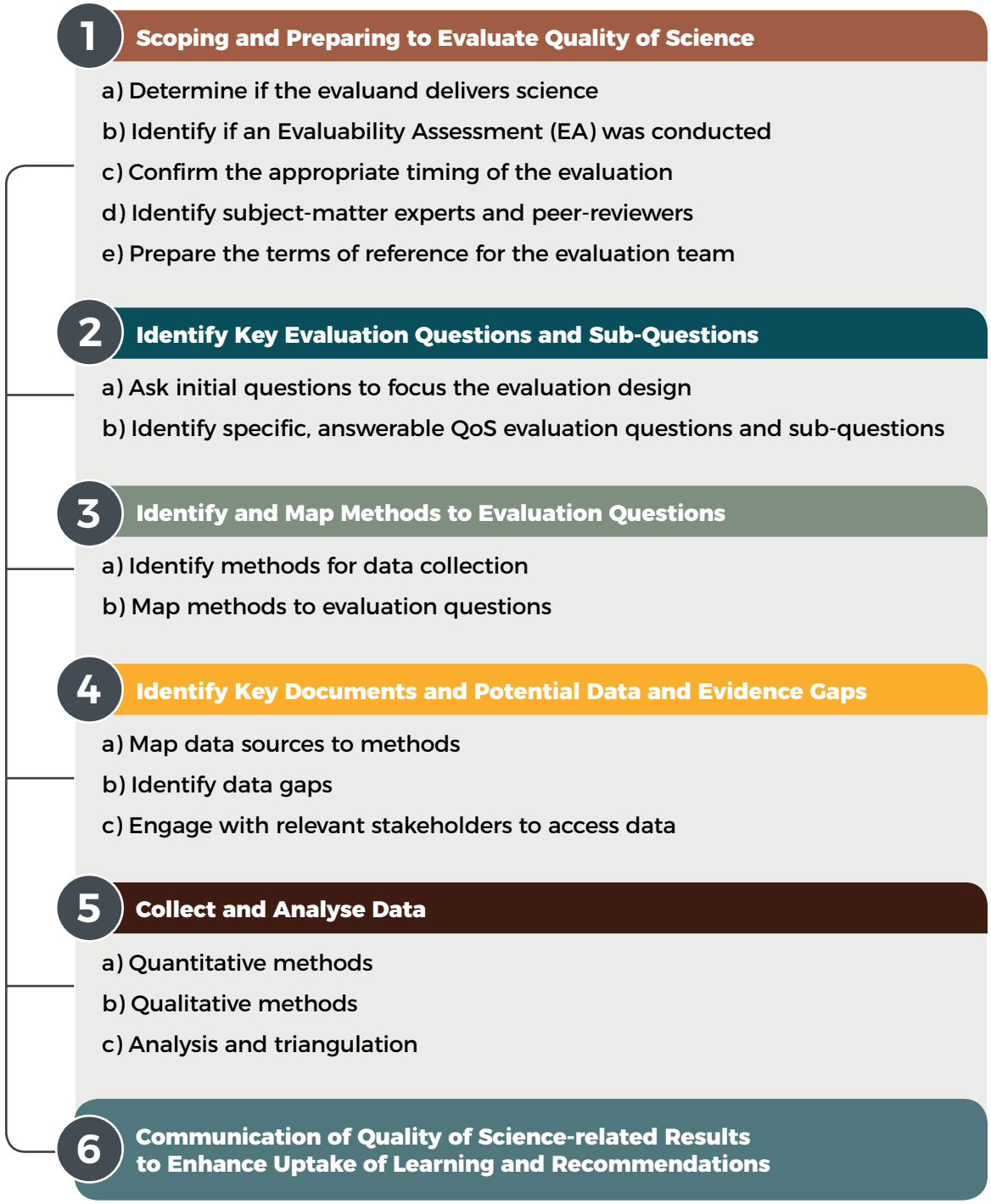
(4) data triangulation, namely the use of different sources of data to reinforce the result.

In Section 4, a six-key-step process is described to facilitate evaluating the quality of science, either as part of a broader evaluation or very targeted to the QoS evaluation criterion as such. Steps 3 and 4 cover a menu of methods and Step 5 focuses the role of data triangulation.

4 Evaluating Quality of Science: Key Steps

For any given evaluation, an Evaluability Assessment, or other pre-scoping or planning stage would recommend a choice among the seven CGIAR criteria that are most appropriate for the specific evaluation purpose. The use of a designated Quality of Science (QoS) evaluation criterion may be deployed (preferably alongside relevance and effectiveness evaluation criteria at a minimum), or science quality can be uncovered obliquely through the use of combinations of the other six evaluation criteria (see Table 3). Which route to take is determined by the evaluation timing and purpose. In either case, a six-step approach to evaluating QoS is suggested. Each step (Figure 7) presents a decision point to ensure that the choice of relevant dimensions (design, inputs, process, and outputs - Section 3.1) and methods remain applicable. For evaluations that are guided by multiple criteria, these steps would be implemented in tandem with all the evaluation criteria in use.

Figure 7: *Six key steps to evaluate Quality of Science*



Step 1: Scoping and Preparing to Evaluate Quality of Science

This first step determines whether the evaluand delivers scientific research. Certain interventions evaluated, e.g., platforms, are not mandated per se to deliver (generate) science. Thus, evaluation is driven by an evaluand's objectives; and the core guiding question is "Was the objective of the evaluand to deliver science?[13]" If yes, QoS is an important determinant of its potential effectiveness (Annex 2). So, Step One is necessary to understand whether the QoS criterion should be applied in relation to the intervention being evaluated. If yes, it entails subsequent preparatory sub-steps.

1a. Determine if the evaluand delivers science

- Ask the question, "Was the objective of the evaluand to deliver science outputs and outcomes for development?" If no, do not include the QOS criterion and pursue another pathway to the evaluation design. If yes, then …
- … identify at what stage the evaluand is in the project cycle, i.e., output delivery, to focus on dimensions: design, input, process, and outputs (Section 3) to control the scope, design evaluation questions, and select appropriate methods.

Example: A clear distinction in the **CGIAR Research Portfolio** could be made between the programs that deliver the science (CGIAR Research Program Evaluative Reviews 2020) versus those that coordinate the delivery of science (for example, the CGIAR Platforms on Big Data in Agriculture and Excellence in Breeding, whose objectives focused on support and coordination around science outputs and processes. Specifically, Big Data in Agriculture aimed to harness the capabilities of big data to accelerate and enhance the impact of international agricultural research but did not directly deliver science while Excellence in Breeding, with a strong focus on data, is focused on driving change rather than the delivery of science.

1b. Identify if an Evaluability Assessment (EA) was conducted. An EA identifies if an evaluand is ready to be evaluated by assessing whether an EA provided and/or captured the following information:

- The (clarified) theory of change associated with each intervention and the original objectives.
- The needs, policies and priorities of users, including global, regional, and country partners and institutions.
- The importance, significance and usefulness of the science produced in a specific context.

If an EA has not been conducted, these areas need to be addressed prior to the evaluation taking place (as a separate process) or addressed in the evaluation design.

1c. Confirm the appropriate timing of the evaluation. For specific guidance, see Section 3. For example, it might be useful for evaluation users to learn about the adequacy of inputs and processes before the end of the intervention or the beginning of the next phase.

1d. Identify subject-matter experts and peer-reviewers. In CGIAR and similar AR4D contexts, qualified experts who are knowledgeable about research for development in CGIAR and broader AR4D contexts should be included.

In addition:

- Peer reviewers: Peer reviewers in an Evaluation Reference Group (i.e., ERG for IAES) who are evaluation experts to review the evaluation design and relevant outputs.
- Subject-matter experts (SMEs): Scientists with expertise that is specifically relevant to the science being evaluated.

1e. Prepare the terms of reference for the evaluation team members. The TORs for evaluation team leader and subject matter experts should explicitly require familiarity with these guidelines on applying the QoR4D Framework to process and performance evaluations.

[13] In case of CGIAR, aligned to the CGIAR's research strategy 2022-2030 "… to deliver science and innovation that advance transformation of food, land, and water systems in a climate crisis… Science-based innovation — co-development of sets of knowledge products, technologies, services, and other solutions along a scaling pathway. CGIAR will work with partners on innovations that include genetics, agrifood management practices, social sciences and institutional solutions, biophysical sciences and solutions, databases, and tools. Activities will include participatory design, testing, and piloting, working closely with private sector partners and regulatory bodies, advancing the enabling environment and providing global architecture for collaborative international agricultural research".

Step 2: Identify Key Evaluation Questions and Sub-Questions

There are two sets of questions that need to be prepared: (1) questions that inform the evaluation design and (2) those that address the evaluative intent (i.e., evaluation questions).

2a. Ask initial questions to focus the evaluation design. The first set of questions broadly focuses on the evaluation design by clarifying what should be evaluated. For this, use the clarified Theory of Change (ToC), and ask the commissioner's representative and program manager to elaborate on it by inquiring:

- How does an output contribute to advancing a research field, domain or discipline or transdisciplinary grouping of these?
- Is the science delivered relevant to the identified development challenges?
- How do innovations (of three types) map to processes and outputs? (Table 1)

2b. Identify QoS evaluation questions and sub-questions that are specific and can be answered by an evaluative exercise (Table 4). Informed by the previous steps, develop QoS-related key evaluation questions and sub-questions, using Annex 5 and Annex 6[14].

To further guide the development of sub-questions and assessment parameters by QoS dimensions, Table 7 presents a menu of themes by QoS dimensions, with suggested indicators and assessment criteria.

Table 4: *Evaluation criteria with sample evaluation questions*

Evaluation Criteria	Sample evaluation question	QoS dimension(s) covered (design, Input, Process, Outputs)
Quality of Science	EQ 1: Is research design appropriate to the development challenges in the context? EQ 2: Are inputs and processes appropriate to produce science that is credible and legitimate? EQ 3: How do the intervention's outputs contribute to advancing science? (i.e., per the full conception of relevance under QR4D)	ALL
Relevance	- Is there evidence of (continuing) demand for the program from intended beneficiaries? - Is the program consistent with the PRMF and Results Framework, and the agreed CGIAR reform agenda?	Design, Outputs
Coherence	Is the design of the intervention coherent with other interventions in the research portfolio/country/sector?	Design, Inputs
Efficiency	- Was the funding adequate and timely? - Are facilities and services adequate and properly utilized? - Was the composition of research teams adequately diverse (inclusive in terms of gender, age/young researchers, and nationality)?	Inputs, Process
Effectiveness	- Did the intervention achieve its objectives and results? Are deliverables positioned for uptake? - Were roles and responsibilities clearly defined and implemented as planned, along ToCs spheres of control and influence? - Any activities that should be modified, discontinued, or added to the current portfolio to enhance the program's likely effectiveness?	Process, Outputs
Sustainability	- Does implementation of the program theories of change and the assumptions underlying these theories include sustainability aspects? - Is the contribution generated by the intervention scalable and likely to be continued? - Have trade-offs between different longer-term outcomes been taken into account in program design and implementation, for instance regarding environmental sustainability?	Outputs
Impact	What was the impact of the studies produced/interventions? Did it have a transformative effect?	Outputs

[14]The data collection matrix covers dimensions of design, input, process and outputs by specific evaluation criteria.

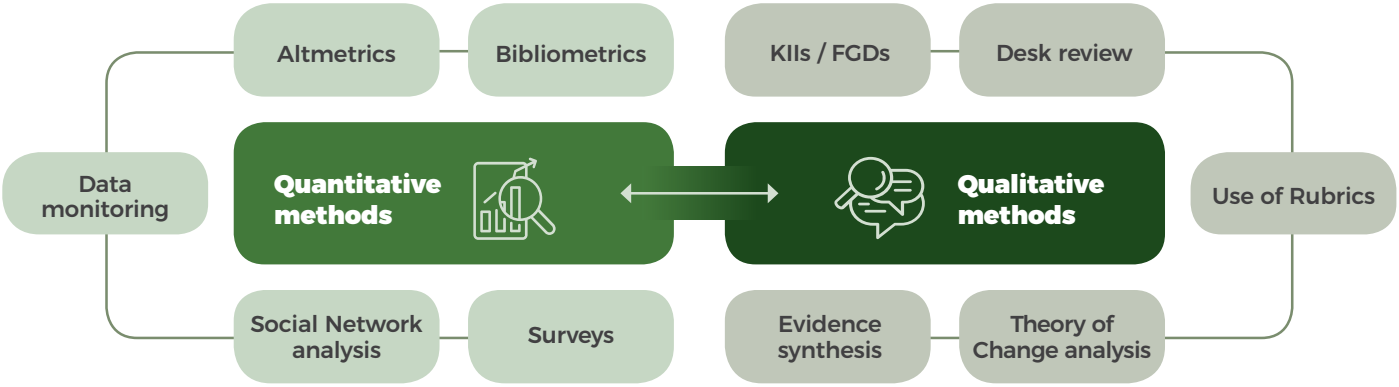
Step 3: Identify and Map Methods to Evaluation Questions

The Evaluation Framework, consistent with the evaluation industry standards, advises the use of mixed methods in performance and process evaluations, which also includes evaluating Quality of Science. Likewise, primary and secondary data sources need to be considered.

3a. Identify methods for data collection

Methods need to be selected based on the evaluation questions, data availability, preference of the key users (i.e., what kinds of data primary users find credible), and timing of the evaluation. For example, questions that focus on process, capacity building, or communication can draw on quantitative (e.g., numbers trained, numbers of methods or tools, scores in tests of trainee comprehension) and/or qualitative (e.g., quality of training, usefulness to planned activities, relevance to the target audience) data. Figure 8 provides a glimpse on the menu if methods to chose consider.

Figure 8: *Evaluation methods for data collection and analysis*



3b. Map methods to Evaluation Questions

Evaluation team members, especially SMEs, need to engage in mapping key evaluation questions and sub-questions to the evaluation methods for data collection and analysis (Figure 8, Table 5, Text Box 3, and Annex 3). The evaluation design matrix, including the QoS criterion (Annex 6), has been developed to identify and group sources of information to quantitative and qualitative methods. The evaluation team leader engages the team through the development of the evaluation design matrix- a key element of the evaluation inception report[16].

Table 5: *Sample evaluation questions and sub-questions by QoS dimensions, data sources and methods*

QoS Dimensions	Evaluation Question	Sample Sub-questions	Data Sources, Methods
Design	Is the research design appropriate and clearly articulated?	- Are research questions and methodology fit-for-purpose and aligned to the research problem? - How interconnected is the research design to SDGs? - Is there evidence of how and what partners were involved in the co-design? - Is the link between the MEL(IA) plan and ToC in the research initiative design?	Initiative proposals and reports; ISDC Initiative review reports; Theory of Change (original and revisions); ISDC ex-ante proposal reviews *Primary*: Interviews
Inputs	To what extent were necessary inputs adequate and sufficient to deliver planned outputs and outcomes?	- Was the composition of research delivery teams adequately diverse? - Were research physical infrastructures (e.g., labs, experimental plots, etc.) adequate? - Did capacity strengthening of the research team and partners address needs vis-a-vis the planned work, including non-scientific aspects?	Initiative reports; Social Network Analysis; Budget reviews against plans; needs assessments and training records *Primary*: Expert field/lab visits; Interviews
Processes	To what extent did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations?	- What were the levels of trust, commitment, and engagement from different partners? - Were there policies followed on mentoring and training junior research staff? - Was performance and monitoring data used for adaptive management?	Initiative reports; Meeting records; Internal policies; *Primary*: Interviews, FGDs

**Additional detail in Annex 3*

[16]The Inception Report (IR) guidance for CGIAR is under finalization: see IR example from the evaluation of CGIAR Big Data in Agriculture Platform and related blog on IAES approach to Inception Reports

Table 5: *Sample evaluation questions and sub-questions by QoS dimensions, data sources and methods (cont'd)*

QoS Dimensions	Evaluation Question	Sample Sub-questions	Data Sources, Methods
Outputs	What is the quality of research outputs, such as improved varieties, knowledge tools, and publications, of high quality? How do the intervention's outputs contribute to advancing science?	- What is the contribution of outputs to science-based innovations, targeted capacity development, and advice on policy? - Were research findings and related outputs clearly communicated/disseminated? - How GDI or environmental concerns are reflected in the outputs? - Can these products have broader applicability and potential for impact at scale?	Bibliometrics; Altmetrics; Initiative reports; Theory of Change; Expert Analysis and Assessment; Download statistics; Social media trends *Primary*: Interviews, FGDs

Methods for consideration in evaluating quality of science is broken down by type in Box 3, with additional detail following on strengths and limitations in Step 5.

Text Box 3: *Overview of quantitative and qualitative methods suggested to evaluate Quality of Science*

Quantitative methods could include **surveys, bibliometric analysis, Altmetrics, social network analysis and monitoring data**. Quantitative analysis mitigates subjective SME judgments. The process of doing science is difficult to measure through quantitative methods, despite the increasing number of bibliometric indicators, and it highly relies on qualitative methods. (See additional detail in Step 5)

Qualitative methods could include **expert reviews, interviews, focus group discussions, theory of change analyses, rubrics and evidence synthesis**. The use of qualitative methods is key to assessing whether a process was ethical and inclusive, with the integration of learning, and the extent to which QoS lies in the content that is produced and displayed in the outputs. Some qualitative methods may be more prone to bias. (See additional detail in Step 6)

Step 4: Identify Key Documents and Potential Data and Evidence Gaps

Determination of what primary and secondary data collection and analysis should be made at this step and depending on the available data sources and preferred methods, any gaps in data sources and feasibility of methods should be determined. If necessary, the evaluation team needs to engage with relevant stakeholders to access data and evidence. At this step, the checklist in Text Box 4 will complement information gathered during scoping, if not during the Evaluability Assessment. Finally, an output could be the creation of new institutions (2021 PRMF/MELIA Glossary; IAES).

Text Box 4: *Checklist for document and information collection*

√ Collect the names and contact information for prominent stakeholders and SMEs (for KII/FGD)

√ Review availability of QoS and project-related information if not covered earlier in EA

√ Depending on documentation availability and quality, determine whether to use external data sources and outsource some elements of the QoS evaluation, for example, bibliometric analysis (see Annex 6)

As an example, to facilitate **bibliometric and/or Altmetric analysis**, Figure 9 presents outputs that are recommended by the CG Core. The CG Core aims to describe all types of information products that are published by the different CGIAR centres. An example of an output-related indicator on the Data Assets: presented data asset types are recommended by the metadata schema, aligned to the industry standards. While the list contains common outposts, it can also be used to generate ideas for likely data sources with the evaluand.

Figure 9: *List of data assets (outputs) in the CG Core grouped into two high-level categories*

Publications

1. Audio
2. Book Publications
3. Book Chapter
4. Brochure
5. Conference Paper
6. Conference Proceedings
7. Directive document
8. Financial Report
9. Geospatial raster file
10. Geospatial vector file
11. Image
12. Infographic
13. Journal (full)
14. Journal Article
15. Knowledge Org. System
16. Legal Document
17. Magazine Article
18. Manual
19. Map
20. News Item
21. Newsletter
22. Newspaper
23. Other
24. Personal Communication
25. Post
26. Poster
27. Presentation
28. Press Item
29. Research Report
30. Software
31. Source Code
32. Standard
33. Technical Report
34. Thesis
35. Video
36. Video Recording
37. Website
38. Working Paper

Datasets

39. Data file
40. Database
41. Dataset As A Collection Of Data

Step 5: Collect and Analyze Data

With the information in hand on evidence and gaps, and the needs known for primary or secondary data collection, the next step entails targeted consideration of data collection and analysis methods. Before collecting and analyzing data however, consider the strengths and limitations of different data collection methods, outlined here.

5a. Use of Quantitative Methods

Table 6 sets out the respective strengths and limitations of various quantitative methods to consider toward assessing QoS.

Table 6: *Menu of suggested **quantitative** methods to evaluate QoS dimensions, with strengths and limitations*

Methods[17]	General Definition	Strengths	Limitations
Bibliometric analysis[18]	Traditionally used to evaluate outputs and their effectiveness is a statistical analysis of books, articles, and other publications, specifically those with scientific content.	Includes some of the most widely accepted indicators of science impact, e.g., impact within a research field (ex. Impact factor, citation index); it provides a good indication of QoS since published papers have already passed a high-quality threshold as they have been peer-reviewed; its value is recognized by funders. Now offers indicators for a broad range of dimensions that allow also for the evaluation of processes and inputs. New bibliometric indicators allow measuring cross-disciplinarity, gender equity, preprinting as an open science practice, or the prevalence of complex multi-national collaborations.	Limiting to peer-review articles measured through bibliometric indicators, would not consider the well-conducted science that does not produce significant results and is then difficult to publish. Not all science, innovation, and research products are included in bibliographic databases; bibliographic databases coming from different sources might not be harmonized; there might be exaggerated attention to a specific author; long periods (it might take decades for results from investments in agricultural research to become visible); it provides little information on policy outreach, contextual relevance, sustainability, innovation and scaling of the contribution; relying uniquely on bibliometrics might miss the rounded picture of the context; bibliometric measures may be skewed depending on the research domain; I.e, comparison across domains is inappropriate).
Altmetrics	'Alternative metrics' (Altmetrics) are used to monitor the reach and impact of scholarship and research through online interactions.	Altmetrics are qualitative data that complement traditional, citation-based metrics. Quicker to accumulate than citation-based metrics; they can capture more diverse impacts than citation-based metrics; they apply to more than journal articles and books.	Altmetrics cannot be used individually as they do not tell the whole story; it has not been widely recognized in the scientific community; there's potential for gaming of Altmetrics that could bias results; it is a relatively new tool, and more research is needed to best benefit from it.
Social Network Analysis	Process of investigating social structures using network analysis and graph theory.	Very practical and visually attractive; much information can be found when interpreting network graphs; it shows also gaps in connections.	The research question needs to be clear to know exactly what needs to be mapped. Interpretation of the network can be complex. It might be necessary to rely on external experts for the methods and interpret the data internally considering the CGIAR context.
Surveys	Data collection from a pre-defined group of respondents.	It is relatively easy to create and does not necessarily need external experts; there are many low-cost tools used to design surveys; that can reach many respondents.	Response rate might be limited, and respondents might be biased; needs to be shared through the right channels to limit biases; survey fatigues could lead to response bias; questions could be interpreted differently by respondents if not clearly stated.
Performance Monitoring data	The Performance & Results Management System (PRMS)[19] encompasses planning, monitoring, and reporting within CGIAR.	PRMS elements - Results Dashboard and others https://www.cgiar.org/dashboards/ compile related information into evaluation products.	Quality of data and information in the dashboards relies on quality assurance mechanisms, which would vary by the dashboard. The varying data quality has been assessed during the 2020 CRP Reviews.

[17] See additional detail on selected methods in Annex 3

[18] https://cas.cgiar.org/evaluation/publications/bibliometric-analysis-evaluate-quality-science-context-one-cgiar

[19] This list of indicators was informed by the RQ+ Assessment Instrument (IDRC, 2022) and includes indicators previously used for 2020 CRP reviews.

5b. Use of Qualitative Methods in social science collect and work with non-numerical data and seek to interpret meaning from these data that help understand social life through the study of targeted populations or places including through words and pictures, as opposed to quantitative methods that focus on numbers. For illustration, Table 7 presents qualitative data themes by QoS dimensions, with suggested indicators and assessment criteria.

Table 8 presents qualitative methods to evaluate QoS and sets out their strengths and limitations. A menu in Table 8 is followed by description of considerations around the analysis.

Table 7: *Qualitative data themes, indicators per Quality of Science dimension with assessment criteria*

Dimension	Theme	Indicator	Assessment Description
Design	Research topic & plan	Global/regional challenge	Appropriate, realistic
	Design	Coherence, clarity	Appropriate
	Methodology	Integrity, fitness	Rigor, clarity
Inputs	Skill base	Discipline*	Appropriate
	Composition of teams	Diversity, gender, discipline*	Appropriate, inclusive, multi- & trans-disciplinarity
	Infrastructures	Laboratories, fields*	Adequate
	Funding	Donor commitment*	Adequate
	Capacity building	Useful for planned activities	Appropriate, adequate
Processes	Partnerships	Inclusiveness, recognition*	Equal team member, involvement in co-design and delivery
	Gender	Awareness, responsiveness*	Gender integrated into design & implementation
	Roles and responsibilities	Defined roles & responsibilities*	Clarity
	Performance evaluation	Incentives*	Rewards for quality
	Negative consequences	Consequences, risks	Risk assessment and mitigation strategy
Outputs	Communication	Methods & tools*	Relevance for the target audience
	Enabling environment	Awareness, understanding	Appropriate positioning and targeting
	Networking	Multi-stakeholder engagement*	Adequate and inclusive
	Policy linkages	Policy makers engagement*	Appropriate and targeted
	Scaling readiness	Multi-stakeholder engagement	Contribution to development outcomes
	Generation of international public good (IPG)	Positioning for uptake and impact*	Broadness of applicability

**Indicators used in the 2020 CRP reviews.*

Table 8: *Menu of suggested **qualitative** methods to evaluate QoS, with strengths and limitations*

Methods	General Definition	Strengths	Limitations
Interviews/ Focus-group discussions	Consultations with main stakeholders either individually (interviews) or by gathering people from similar backgrounds or experiences together to discuss a specific topic of interest (FGD).	Useful for collecting information on experiences, understanding met and unmet needs, and providing ideas for improvement.	The subjectivity of the process. Risk of not including important stakeholders due to power asymmetries. Risk of self-reporting biases/inaccuracies and groupthink.
Expert review, Desk-review	Evaluation of available documentation, literature, and reports.	*Expert (SME)* document review adds credibility and rigor. Decreased time pressure on evaluand after documents are furnished; investments in automation and mining procedures allow for high throughput analysis; materials may have already been peer reviewed and or quality assured by third parties, supporting credibility.	Expert biases. Laborious (typically manual) process that requires many person-hours and relevant subject matter expertise.
Evidence synthesis	Bringing together information from a range of sources.	More useful to decision/policy makers to receive synthetic information; encourages the observation of trends and patterns over time and space.	Dependent on the variety of quality evidence that can be found.
Theory of Change	Comprehensive description and illustration of how a desired change is expected to happen in a particular context.	Helps identify whether a project is delivering on its original objectives; can help adjust the projects to best meet the needs of the final beneficiaries; might be a good tool to assess relevance and reach.	Some assumptions might be wrong, hence, relying uniquely on ToC might lead to a lack of attention to challenging findings; its value might not be always recognized by funders.

Additional detail and considerations on selected qualitative methods follows next.

Interviews and FDGs: Semi-structured interviews are likely the most useful method for QoS reviews. These are interviews that have specific, focused questions and also provide room for open-ended or more exploratory questions. These interviews can be done individually with key informants or take place in form of focus group discussions, which only focus on 1-2 main themes. Focus group discussions (FGDs) are useful when there is a key question that is likely better answered through structured discussion. People, as opposed to documents, are the 'data source' and should be selected using transparent criteria[21] (e.g., knowledge of the specific sector).

Some key actors to engage with could be:

- CGIAR Management and Science Leads (Action Areas, Impact Platforms).
- Intervention teams: Director, Program Head, MEL Lead, and others as applicable.
- Key staff including project leaders, managers.
- Research managers, from the network of research centers involved in a project.
- Scientists in the National Agricultural Research Systems (NARS).
- Donors as per additional information from the Funder Analysis Dashboard.
- Partners including academia, NGOs, and the private sector.
- Early career researchers, defined as researchers with recently obtained a PhD.

[21] Qualitative research provides specific sampling protocols that should be adhered to.

Desktop reviews, or document reviews, are used to identify key search terms, patterns, and themes that address evaluation questions around the design's outputs, processes, inputs, and design/rigor.

In CGIAR, key documents include but are not limited to:

- **Initiative reports.** The 2022-2030 PRMF mandates each CGIAR Initiative to develop annual work and budget plans, track progress and provide an annual report against the stated objectives and results achieved.
- **Theory of change.** Initiatives plan and report their annual progress against a ToC that incorporates results and indicators across the spheres of control, influence, and interest of the Initiative, and is adjusted annually in a reflection process.
- **Outcome and impact reports (e.g., results stories).** Short reports describing the contribution of research projects to development outcomes and impact. A CGIAR example is the Outcome and Impact Case Reports which are useful to understand effectiveness in terms of outcomes and achievements along impact pathways.
- **Impact assessment studies.** CGIAR interventions and partners implement impact assessment studies to test the assumptions in the ToC to contribute to their improvement and increased impact.

5c. Analyse and Triangulate Data

Different analysis and triangulation techniques would accompany selected methods.

Qualitative data: Qualitative data drawn from documents or interviews can be organized by hand, such as using Word or Excel, or organized with computer software. Some of the more common software packages include Atlas.ti, MAXQDA and NVivo. While most qualitative data analysis is iterative through the data collection process, some software packages can assist with analyzing data drawn from interviews and documents. These include Cynefin Sensemaker, Sprockler, and Narrafirma. These data analysis packages require specific training.

Rubrics. Rubrics set out criteria and standards for different levels of performance and describe and value what performance would look like at each level. In qualitative analysis, the use of rubrics allows mitigation for subjectivity.[22] Questions in Table 4 and Annex 7 are formulated in a way that a simple light scoring system can be applied, where red indicates a serious problem (=No), yellow a minor problem that can be solved (=partly), and green that the QoS-related dimension is performing well for that specific indicator/question (=Yes).

Table 9: *Criteria to assess the quality of selected peer-reviewed publications*

Assessment Criterion	Assessment Approach
Do the results (knowledge presented in the paper) represent broadly applicable knowledge (international public goods) relevant to the intervention's objectives?	Rating scale: 0=results not relevant to agriculture and climate change 1=no broader applicability (local relevance only) 2= potentially broader applicability, but not spelled out 3=broader applicability is presented 4=significant international applicability
Quality (and appropriateness) of publication venue	Observation of low-quality or inappropriate venue relative to subject and quality of paper
Co-authorship	Observation of the extent of co-authorship, with whom, and whether it is appropriate
The overall quality of publication (including additional criteria at evaluator/SME's discretion)	Brief overall assessment (around 100-150 words)

Mixed methods-Assessment of the quality of peer-reviewed publications for *relevance* and *credibility*: In the 2020 CRP reviews (see Annexes for RTB, CCAFS, WLE, and Livestock), in-depth expert reviews of selected outputs, including peer-reviewed and other technical publications and physical products (germplasm, digital innovations, and services) added credibility and rigor to address the QoS evaluation criterion, and guide recommendations on future orientation. Criteria combined quantitative and qualitative aspects and used a unified rating scale across the twelve reviews. Focused specifically on the outputs, Table 10 provides specific criteria to use for assessing the quality of peer-reviewed publications.

Data triangulation supports high-quality science-specific conclusions and evidence-based recommendations and enhances and evaluation's credibility. Data triangulation can happen throughout the data collection and analysis process, as well as at the end. Triangulation facilitates validation of data through cross verification from more than two sources. It tests the consistency of findings obtained through different instruments and increases the chance to control, or at least assess some of the threats or multiple causes influencing our results.

[22] An example of use of rubrics from a program in CIFOR & IUCN https://www.cifor.org/wp-content/uploads/dfid/KNOWFOR%20-%20Rubrics%20and%20Guidance%20Notes.pdf from the following report International Forestry Knowledge Programme (publishing.service.gov.uk)

Step 6: Communication of Quality of Science-related Results to Enhance Uptake of Learning and Recommendations

Evaluation of science quality supports robust decision making processes in research for development settings. Effective communication of the evaluation of science quality processes and QoS-related findings and conclusions is key to credibility, learning, and the uptake of evaluation recommendations and lessons. When the core evaluation users are aware of the potential learning from the evaluation, and their roles both in the evaluative process and subsequent uptake of this learning, the recommendations are more likely to be acted upon. For QoS and indeed all types of process and performance evaluations, there are typically two kinds of recommendations[23]:

- **Formal recommendations:** are numbered in an evaluation report's 'recommendations' section. Formal recommendations, including sub-recommendations, must receive a written Management response according to CGIAR Evaluation Policy (2022). A well-documented MR is a learning document that may contribute to helping CGIAR avoid and mitigate strategic, policy or systemic problems arising in future programming.

- **Informal recommendations:** An evaluation team may decide to make a 'suggestion' or observe 'a lesson learned' instead of a formal recommendation. There are many reasons for this including when the recommendation falls outside of the scope of their TOR, need to prioritize more substantive recommendations or the recommendation is not likely to be feasible or actionable. For such informal recommendations, while not required, management may choose to respond.

In particular, when the QoS criterion is applied, it is important to balance having stand-alone recommendations by QoS dimensions and embed the other recommendations that may have come out of assessing efficiency, sustainability, and other evaluation criteria.

Irrespective of the type of recommendations, it is highly advisable that evidence synthesized by QoS-related dimensions is presented to science managers, researchers, and other relevant stakeholders during the validation phase, and detailed in the final evaluation report. Subsequently, effective communication will increase the understanding of performance and process evaluations of QoS and build stakeholder confidence to motivate positive responses to recommendations and lessons learned.

One evaluation may result in multiple communication products aimed at different user groups. Identification of the specific information needs of key groups will determine the appropriate type of communication products and ways to manage evaluative knowledge (KM). These may include presentations, short videos or podcasts, blogs, briefs, conference interventions, and other means to reach audiences with timely and appropriate content, to facilitate the delivery of messages and use of information.[24]

[23] See CGIAR Evaluation Guidelines: Management Engagement and Response.

[24] See CGIAR Evaluation Guidelines: Final Evaluation Report

5 Bibliography

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. SAGE Open, 9(1), 215824401982957. https://doi.org/10.1177/2158244019829575

Bryman, A. (2004). Triangulation and measurement. Retrieved from Department of Social Sciences, Loughborough University, Loughborough, Leicestershire

Carneiro, B., Resce, G., Läderach, P., Schapendonk, F., & Pacillo, G. (2022). What is the importance of climate research? An innovative web-based approach to assess the influence and reach of climate research programs. Environmental Science & Policy, 133, 115-126. https://doi.org/10.1016/j.envsci.2022.03.018

Carter, N. (2014, September). The use of triangulation in qualitative research. In Oncology nursing forum (Vol. 41, No. 5, p. 545). 10.1188/14.ONF.545-547

CAS Secretariat (CGIAR Advisory Services Shared Secretariat). (2020). CGIAR Research Program 2020 Reviews: Policies, Institutions, and Markets. Rome. CAS Secretariat Evaluation Function. https://cas.cgiar.org/ and List of Annexes

CAS Secretariat (CGIAR Advisory Services Shared Secretariat). (2020). CGIAR Research Program 2020 Reviews: Roots, Tubers and Bananas (RTB). Rome. CAS Secretariat Evaluation Function. https://cas.cgiar.org/

CAS Secretariat (CGIAR Advisory Services Shared Secretariat). (2021). Synthesis of Learning from a Decade of CGIAR Research Programs. Rome. CAS Secretariat Evaluation Function. https://cas.cgiar.org/

CGIAR Independent Advisory and Evaluation Service (2022). CGIAR Evaluation Framework. Rome: CGIAR Independent Advisory and Evaluation Service.

CGIAR System Management Office. (2020). CGIAR Performance and Results Management Framework 2022-2030. Montpellier. CGIAR. https://hdl.handle.net/10568/113793

CGIAR System Organization. (2021). CGIAR 2030 Research and Innovation Strategy: Transforming food, land, and water systems in a climate crisis. Montpellier. CGIAR System Organization. https://cgspace.cgiar.org/handle/10568/110918

CGIAR System Organization (2021). 2022-24 Investment Prospectus: pooling funds for research and innovation to transform food, land, and water systems. Montpellier. CGIAR System Organization. https://www.cgiar.org/research/investment-prospectus/, https://storage.googleapis.com/cgiarorg/2021/06/Document-SC13_02_Endorsed-2022-24-Investment-Prospectus.pdf and Companion Document.

De Col, V., Jani, S., Rünzel, M., Tobon, H., Almanzar, M., See, D. S., & Bonaiuti, E. (2021). Case Study on the Monitoring-Quality Assurance Processor-API: A Tool to Support CGIAR Quality Assurance Process for Peer-reviewed Publications. Beirut, Lebanon: International Center for Agricultural Research in the Dry Areas (ICARDA). https://repo.mel.cgiar.org/handle/20.500.11766/66480

De Col, V. (2022). Alone we can do so little; together we can do so much. CAS Secretariat. https://cas.cgiar.org/evaluation/news/alone-we-can-do-so-little-together-we-can-do-so-much

Hcéres (2022). Evaluated entities ares. https://www.hceres.fr/en/evaluated-entities-area

Heale, R., & Forbes, D. (2013). Understanding triangulation in research. Evidence-based nursing, 16(4), 98-98. http://dx.doi.org/10.1136/eb-2013-101494

5 Bibliography cont'd

IDRC (International Development Research Centre). (2022). The International Development Research Centre's Research Quality Plus (RQ+) Assessment Instrument. www.idrc.ca/RQplus

IEA (Independent Evaluation Arrangement) (2015). CGIAR Standards for independent external evaluation. https://iaes.cgiar.org/sites/default/files/pdf/Standards.pdf

IEA (Independent Evaluation Arrangement). (2015) IEA Workshop on Evaluating Quality of Science. Report. Rome 10-11 December 2015 https://iaes.cgiar.org//sites/default/files/pdf/Report_QoSWorkshop-final-1.pdf

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 2005–2019. PLOS One, 15(4), e0231735. https://doi.org/10.1371/JOURNAL.PONE.0231735

Langfeldt, L., & Scordato, L. (2015). Assessing the Broader Impacts of Research: A Review of Methods and Practices. Oslo: Nordic Institute for Studies in Innovation, Research and Education. https://nifu.brage.unit.no/nifu-xmlui/handle/11250/282742

Neylon, C., Willmers, M., & King, T. (2014). Rethinking impact: Applying Altmetrics to southern African research. Paper/Scholarly Communication in Africa Programme. Canada. IDRC. http://hdl.handle.net/10625/53461

OECD/DAC. (2019). Better criteria for better evaluation: Revised evaluation criteria definitions and principles for use. The Organisation for Economic Co-operation and Development (OECD) Development Assistance Committee (DAC). https://www.oecd.org/dac/evaluation/revised-evaluation-criteria-dec-2019.pdf

Science-Metrix & CGIAR Advisory Services Secretariat Evaluation Function (2022). Bibliometric Analysis to Evaluate Quality of Science in the Context of One CGIAR. Technical Note. Rome. https://cas.cgiar.org/evaluation/publications/bibliometric-analysis-evaluate-quality-science-context-one-cgiar

Science-Metrix. (2018). Review of the Human Frontier Science Program Final Report. Quebec. Science-Metrix inc. https://www.hfsp.org/node/12547#book/

Serrat, O. (2017). Social network analysis. In Knowledge solutions (pp. 39-43). Springer, Singapore. Social Network Analysis | SpringerLink

Wilson, V. (2014). Research methods: Triangulation. Evidence-based library and information practice, 9(1), 74-75. https://doi.org/10.18438/B8WW3X

Annex 1: Stakeholders Key to Developing the Guidelines

The guidelines development followed a consultative, inclusive, and iterative approach that included strategic and operational discussions with internal and external CGIAR stakeholders, including, among others, CGIAR governance and management, the CGIAR MELCOP, and an external peer review. Annex 1 provides a list of experts engaged in performed activities. The core grounding and sources of knowledge came from a workshop on evaluating quality of science (2015), EvalForward[25] discussion (EN, FR, ES) on evaluating science, technology and innovation in a development context (2022), and expert engagement at the European Evaluation Society Conference (June 2022); and from selected recommendations on the use of bibliometrics in mixed methods evaluations from the technical note[26].

Stakeholder group	Validation workshop, June 23 2022	EvalForward[27] CoP discussion (EN, FR, ES)
Funders	- Raphael Nawrotzki, M&E officer for the Fund International Agricultural Research (FIA) at the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), Germany	
CGIAR Independent Science for Development Council (ISDC)	- Holger Meinke, ISDC Chair; also Adjunct Research Professor for Global Food Sustainability at the University of Tasmania, Australia - Andrew Ash, ISDC member. Director and Principal at AJ Ash and Associates. Adjunct Professor at School of Agriculture and Food Science, The University of Queensland, Australia - Amy Beaudreault, ISDC Secretariat Lead, IAES, Italy - Pierre Boulanger, ISDC Secretariat Advisor, IAES, Italy	
Research and other partners, universities		- Claudio Proietti, Monitoring and Evaluation Advisor at CIRAD, France - Nobert Tchouaffe, Researcher at the Pan-African Institute for Development, Cameroon - Rachid Serraj, Associate Director of Strategy at Mohammed VI Polytechnic University, Morocco - Richard Tinsley, Professor Emeritus at Colorado State University, USA - Valeria Pesce, Partnership Facilitator at Global Forum on Agricultural Research and Innovation (GFAR), Italy
	- Etienne Vignola Gagné, Analyst at Science-Metrix / Elsevier, co-author of the Technical Note, Canada	
Evaluation Reference Group (ERG) to CGIAR Independent Advisory and Evaluation Service	- Guy Poppy, Director and Professor at the University of Southampton, UK	
	- Zenda Ofir, Scientist & full-time inter national evaluator – Written feedback, Switzerland/South Africa	- Sonal D Zaveri, Founder and Coordinator GENSA, Community of Evaluators South Asia, India - Ola Ogunyinka, Monitoring, Evaluation and Impact Specialist at Natural Resources Institute (NRI), University of Greenwich, UK
CGIAR	- Alessandra Furtado, Interim head of project coordination unit; Head of Project Management at the International Potato Center (CIP), Mozambique	- Valentina De Col, Agricultural Information System Officer at the International Center for Agricultural Research in the Dry Areas (ICARDA), Germany – Graham Thiele, former director of CGIAR research program on Roots, Tubers and Bananas, Peru
	- Bia Carneiro, Social Research & Media Specialist for CGIAR FOCUS Climate Security, Portugal	

[25] EvalForward is a Community of Practice on Evaluation for Food Security, Agriculture and Rural Development. It brings together officers and professionals to exchange experience and to strengthen capacities for evaluation at country level. EvalForward intends to contribute to the evaluation of progress towards Sustainable Development Goal 2: End hunger, achieve food security and improved nutrition and promote sustainable agriculture.

[26] Science-Metrix integrated with Elsevier's Research Analytics and Data Services (RADS) team in 2018.

[27] EvalForward is a CoP on Evaluation for Food Security, Agriculture and Rural Development.

Annex 1: Stakeholders Key to Developing the Guidelines cont'd

Stakeholder group	Validation workshop, June 23 2022	EvalForward[27] discussion (EN, FR, ES)
Independent and private	- Beverly Parsons, Executive Director at InSites; president of American Evaluation Association (2013-2015), USA - Sara Vaca, Data Visualization Consultant, CGIAR and UNICEF, France	- Keith Child, Evaluation Consultant to IAES, previously MEL expert for CGIAR and Water, Land and Ecosystems (WLE) CRP, Canada - Lennart Raetzell, Manager at Syspons GmBH, Germany
Independent experts, 2020 CRP Evaluative Reviews	- Jillian Lenne - Independent Consultant and editor. Previously SME for Quality of Science for 2020 CRP Reviews (RTB and GLDC), UK	
	- Donna Podems – Evaluator for 2020 WHEAT CRP Review, South Africa - John Morton – Professor of Development Anthropology at NRI; SME for the 2020 CCAFS CRP Review, UK	- Paolo Sarfatti, Evaluation Senior Strategic and Technical Advisor, Italy - Paul Engel, SME for 2020 PIM CRP Review; Team Leader at Knowledge, Perspectives and Innovations, Netherlands
The Food and Agriculture Organization of the United Nations (FAO)	- Rachel Sauvinet Bedouin, Senior Evaluation Officer; previously Head of Independent Evaluation Arrangement (IEA) of CGIAR - written feedback	- Nanae Yabuki, Evaluation Officer - Ibtissem Jouini, Regional Evaluation Specialist for Near East and North Africa (RNE); previously Evaluation consultant at IAES - Serdar Bayryyev, Senior Evaluation Officer
CGIAR Independent Advisory and Evaluation Service (IAES)	- Svetlana Negroustoueva, Evaluation Function Lead	
	- Allison Grove-Smith, IAES Director - Gaia Gullotta, Data analyst consultant - Inese Berzina, Administrative coordinator	

Key Informant Interviews		Topic
CGIAR	- Bia Carneiro, Social Research & Media Specialist, formely CCAFS CRP	Social Network Analysis (SNA)
	- Valentina De Col, Agricultural Information System Officer, ICARDA	
Curtin University, Australia	- Cameron Neylon, Professor of Research Communication, Curtin University, Center for Culture and Technology	Altmetrics
University of Cape Town, South Africa	- Michelle Willmers, Publishing and Implementation Manager of the Digital Open Textbooks for Development project	
United Nations Population Fund (UNPFA)	- Lamin Massaquoi, Data Expert and East Africa Team Leader	Use of RQ+ framework, International Development Research Centre (IDRC)
	- Kais Al-Abhar, Monitoring and Evaluation Analyst	

[27] EvalForward is a CoP on Evaluation for Food Security, Agriculture and Rural Development.

Annex 2: CGIAR Evaluation Criteria[28]

Apart from Quality of Science evaluation criterion, extended guidance on other criteria is available under the OECD DAC Network on Development Evaluation (EvalNet) https://www.oecd.org/development/evaluation/.

Quality of Science: The QoS evaluative criterion pertains to scientific credibility and legitimacy. The definition of the criterion derives from the QoR4D frame of reference, which records CGIAR's System-wide agreement on the nature and assessment of research quality. The QoR4D describes research quality according to four key elements: relevance, scientific credibility, legitimacy, and effectiveness.[29] Relevance and Effectiveness are treated as separated evaluation criteria.

Relevance: The extent to which the intervention's objectives and design respond to the needs, policies, and priorities of users/clients and global, regional, and country partners/institutions and continue to do so if circumstances change. Consistent with the QoR4D framework, attention is given to the importance, significance, and usefulness of the work implemented in the problem context, associated with CGIAR's capacity to address the problems.

Effectiveness: The extent to which the intervention achieved, and/or is expected to achieve, its objectives, and its results, including any differential results across subgroups of users/clients. Consistent with the QoR4D framework and in the CGIAR context, this criterion considers the extent to which research is positioned for use and has generated knowledge, products, and services with high potential to address a problem and contribute to innovations, outcomes, and impacts. Effectiveness, therefore, implies that research has been designed, implemented, and positioned for use within a dynamic theory of change, with appropriate leadership, capacity development, diversity of research skills, and support to the enabling environment to translate knowledge into use and to help generate desired outcomes.

Coherence: The compatibility of the intervention with other interventions in a country or a sector or within CGIAR; its overall fit. Internal coherence addresses the synergies and interlinkages between the intervention and other interventions carried out within CGIAR, and the consistency of the intervention with the relevant international norms and standards to which CGIAR adheres. External coherence considers the consistency of the intervention with other actors' interventions in the same context—that is, its complementarity, harmonization, and coordination with others, its value-added, and its avoidance of duplication of effort.

Efficiency: The extent to which the intervention delivers, or is likely to deliver, results in an economical and timely way—that is, the overall use of resources. "Economical" refers to the conversion of inputs (funds, expertise, natural resources, time, etc.) into outputs, outcomes, and impacts in the most cost-effective way possible compared with feasible alternatives in the context. "Timely" delivery is within the intended timeframe, or a timeframe reasonably adjusted to the demands of the evolving context. This criterion may include assessing operational efficiency (how well the intervention was managed).

Sustainability: The extent to which the net benefits of the intervention continue or are likely to continue. This criterion focuses on continuation of benefits, not on external funding, and highlights the multidimensional nature of sustainability.

Impact: The extent to which the intervention has generated or is expected to contribute to generating significant positive or negative, intended or unintended higher-level effects. Impact addresses the ultimate significance and potentially transformative effects of the intervention.

[28] https://iaes.cgiar.org/evaluation/publications/cgiar-evaluation-policy

[29] A co-designed guideline on evaluating the Quality of Science in CGIAR details the approach and methods for operationalizing the QoS evaluation criterion of this Policy.

Annex 3: Sample Evaluation Questions and Sub-Questions by Methods and Data Sources

Evaluation Question	Sample Sub-questions	Methods & data sources
Design		
EQ 1. Is research design appropriate to the development challenges in the context?	1. Is there a documented link between a stated objective and Impact Areas? 2. Is there program-level or other evidence of changed methodology if research objectives changed? Learning: are prior research outputs/findings clearly described and integrated? 3. Are research questions and methodology fit-for-purpose and aligned to the research problem? 4. How innovative is the research and science? Was comparative advantage considered? 5. How interconnected is the research design to SDGs including within each SDG? 6. How aligned is research design to shared, multi-funder, and partners priorities? 7. How and what partners were involved in the co-design of the delivered science? 8. Is the link between the MEL(IA) plan and indicators in the ToC clearly defined in the research initiative design?	Initiative proposals and reports; ISDC Initiative review reports; Interviews; Theory of Change (original and revisions); ISDC ex-ante proposal reviews
Input		
EQ 2a. To what extent were necessary inputs adequate and sufficient to deliver planned outputs and outcomes? (relates to EQ.2 in Figure 4)	1. Was the composition of research delivery teams adequately diverse (inclusive in terms of gender, age/young researchers, and nationality)? 2. Was there an appropriate range of disciplines and skills given the topic of the research? 3. Were research physical infrastructures (e.g., labs, experimental plots, etc.) adequate? 4. Was research funding sufficient and timely received? 5. Did capacity strengthening of the research team and partners address needs vis-a-vis the planned work, including non-scientific aspects?	Initiative reports; Bibliometrics; Interviews; Social Network Analysis; Expert field/lab visits; Budget reviews against plans; needs assessments and training records
Process		
EQ 2b. To what extent did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations? (relates to EQ.2 in Figure 4)	1. What was the level of trust, understanding, and commitment with partners (of different types)[30]? 2. To which extern were partners embedded in the research team and their operations? 3.Were there policies in place for research ethics; were they well implemented? 4. Were **roles and responsibilities** clearly defined and implemented as planned, along ToCs spheres of control and influence? 5. Were there policies in place for internal peer-review mechanisms, to enhance learning? 6. Were there policies in place for **mentoring and training** junior research staff? 7. Were **risk assessment and mitigation strategies** put in place?	Interviews; FGDs; Initiative reports; Meeting records; Internal policy analysis; Risk matrix analysis

[30]Alignment to a 'Partnership Framework'

Annex 3: Sample Evaluation Questions and Sub-Questions by Methods and Data Sources cont'd

Evaluation Question	Sample Sub-questions	Methods & data sources
Outputs		
EQ 3. How do the intervention's outputs contribute to advance science? In what ways are the research outputs of high quality: a. Physical products: germplasm, digital innovations & services b Research and technical publications, training materials, toolkits, decision support mechanisms, and policy advice	1. Quality and quantum of scientific and technical publications and other outputs?	Bibliometrics; Altmetrics; Initiative reports; interviews; Theory of change; FGD; Expert analysis of scientific publications (desk review quality)
	2. How many publications were produced? What was the impact factor of journals? What was the share of highly cited publications? Who were the most productive authors?	
	3. Were publications cited through different channels (rather than the most traditional ones), such as blogposts, Twitter, etc.?	
	4. Were research findings and related outputs clearly communicated/ disseminated?	
	5. Was there a request from partners and/or other stakeholders to present the research and its outputs? What was the reach of publications in the focal countries and to NARS?	
	6. What is the contribution of outputs to science-based innovations, targeted capacity development, and advice on policy?	
	7. Pathways and documented contribution of outputs to SDG?	
	8. How GDI or environmental concerns or localization efforts/tailoring to particular contexts, for example, are reflected in the outputs? Co-authors from the global south (prevalence).	
	9. Can these products have broader applicability and potential for impact at scale?	Initiative reports; Interviews; FGDs; Expert Analysis
	10. To which extent are these products relevant to the target audience?	

Annex 4: Data Parameters and Analysis

Document reviews: Box 1 below shows sample documentation key in document review towards evaluating QoS and also using other evaluation criteria (see Evaluability Assessment guidelines).

Box 1: Examples of sample documentation and external sources:

1. **Proposal or strategy documents**
2. **Theory of change**
3. **Results Framework** or other document with articulated Inputs, activities and outputs, desired outcomes and impacts
4. **Project lists** (with related documentation)
5. **Contact lists** for internal and external stakeholders and key informants
6. Previous independent or other evaluations, studies, and impact assessments
7. **Impact Assessments**
8. **Key databases** with potentially relevant information
9. **Peer reviewed publications**
10. **Policy briefs**
11. **Working papers**
12. **Pre-prints**
13. **Technical reports**

Altmetrics track a range of sources to collate conversations about research happening online daily. Altmetrics are metrics and qualitative data complementary to traditional, citation-based metrics. They can include (but are not limited to) peer reviews on Faculty of 1000, citations on Wikipedia and in public policy documents, discussions on research blogs, mainstream media coverage, bookmarks on reference managers like Mendeley, and mentions on social networks such as Twitter. Sourced from the Web, Altmetrics can tell a lot about how often journal articles and other scholarly outputs like datasets are discussed and used worldwide. It is useful to monitor and report on the attention that a work is getting through channels that are different from the most common ones. Altmetrics contains the potential for a comprehensive reconceptualisation of what qualifies as impact, what should be rewarded in institutional reward and incentive structures, and how to track and promote engagement with civil society partnerships (Neylon, 2014). Evaluative review of the Policies, Institutions, and Markets (PIM) CGIAR research program (2020) describes use of Altmetrics attention score[31] in evaluating QoS (see Annex). Altmetric Attention Score is obtained for free by the PRMS team and does not require initiatives or centers to pay an annual subscription unless they wish to use advanced services provided by Altmetric for their own use.

Social Network Analysis (SNA) is a graphic way of depicting the number and strength of connections between people, including researchers, institutions, government partners, etc. Social network analysis seeks to understand networks and their participants and has two main focuses: the actors and the relationships between them in a specific social context (Serrat 2017). An example of using SNA in the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) as a case study (Carneiro (2022)). The authors effectively repurpose publicly available data from digital sources such as social media and websites by employing text mining and SNA to assess the influence and reach of the program among stakeholder at various levels. Furthermore, the following blog on RTB and GLDC CRPs describes the potential of Network Analysis to complement other methods and metrics used to monitor and assess the QoS. The example co-authored by the CGIAR centers ICARDA and CIP with the University of Florida, shows how integrating different dimensions like geographical focus, gender and duration of the collaborations and bibliometrics into the network analysis, brings value to analyzing the QoS inputs and processes, under the QoR4D. Suggested softwares for SNA include: VOSviewer, a software tool for constructing and visualizing bibliometric networks (open-source and free); Gephi, a visualization and exploration software for different kind of graphs and networks (open-source and free). This kind of analysis requires cooperation between an expert in the field and an internal evaluator who could interpret the results in light of the CGIAR context.

Bibliometrics includes a powerful set of tools to assess the scientific performance of various entities—countries, regions, institutions, or researchers—by characterizing dimensions of their scientific outputs (i.e., mostly peer-reviewed scientific publications), such as the size of their production, their collaboration patterns, their scientific impact, and the extent to which they recombine different fields of knowledge through, among other things,

[31] It is a score for a specific research output that indicates the amount of attention it has received. The score is derived from an automated algorithm and represents a weighted count of the amount of attention received for research output.

partnerships with experts from a diversified set of fields. Bibliometrics is used to evaluate scientific funding, policies, and activities—particularly to assess the outcomes of those interventions on research excellence—and is being implemented and used for this purpose by a wide range of governmental and nongovernmental bodies internationally (Jappe, 2020). Bibliometric indicators can also be used to assess other dimensions, such as processes and inputs- see Annex 6 for evaluation questions related to input and processes that can be at least partly addressed using bibliometrics. All bibliometric indicators have some weaknesses when considered individually. For the quantitative evaluation, an extended use of bibliometrics is crucial and it can be fueled and improved by qualitative inputs. Consequently, it is important to use various lines of evidence to triangulate the results.

The list of high-priority indicators reported in Table 6 includes indicators of: equal gender participation; shares of publications that are academic-private co-publications; cross-disciplinary integration of the social sciences and humanities (SSH) within publications; normalized citation impact; cross-disciplinarity; and South-South, South-North co-publication, thematic alignment with SDGs. Except for some normalized citation impact indicators and cross-disciplinarity indicators, the indicator formulae are a simple division, expressed as a proportion or percentage of an overall publication set. The indicators can be computed by assembling an overall publication set (for example, all publications from a 2022–24 Initiative or, say, all publications from an Action Area). The number of publications in this set is the denominator in the indicator formulae. The numerator is determined by counting the number of publications within the overall set that fulfill a criterion—for example, publications that include at least a female co-author or a Southern co-author or that have received at least one journalistic mention as tracked in Altmetrics databases. Inclusion in the numerator count can also be based on multiple criteria—for example, the publication has a women author as either first, last, or corresponding author and women authors make up 50% or more of authorships within the publication. The prioritization of indicators does not indicate the level of authority of a single indicator against other indicators. It is always recommended to use a panel of complementary indicators to capture different aspects even of a single phenomenon. Annex 7 of the Technical Note includes a wider list of indicators to consider for the evaluation, based on their feasibility and on the evaluation's needs.

Web of Science: paid-access platform that provides access to multiple databases that provide reference and citation data from academic journals, conference proceedings and other documents in various academic disciplines. Useful to calculate bibliometric indicators and Altmetrics. For Bibliometric analysis it is suggested to use Bibliometrix, an R-tool for science mapping analysis and Biblioshiny, a Shiny app for bibliometrix (both free).

Performance monitoring data: CGIAR Results Dashboard: The Dashboard provides access to data in bulk and access to publications and Outcome Impact Case Reports (OICR) with detailed information. This tool was originally created for management and accountability purposes and is expected to evolve as an M&E tool. Information within has gone through quality assurance. The use of the dashboard through MEL and MARLO combined into CLARISA varies from one CRP to another. CLARISA: CLARISA (CGIAR Level Agricultural Results Interoperable System Architecture) Is a web service that helps to transform raw data on CGIAR research and activities into meaningful information that can shape how we work and reveal what our impacts are on development – on reducing poverty, improving food and nutrition security for health, and improving natural resources and ecosystem services. In programming terms, CLARISA is a REST-API, which means it is a type of web service that enables computer systems to work together over the Internet. CLARISA enables systems like MARLO, MEL and others to communicate with each other, finds common ground in their data, and produces standardized, aggregated information in the language needed for System-level reports. It works by using control lists of standardized key terms, such as those commonly used by the CGIAR Strategy and Results Framework and the the SDGs.

Annex 5: Mapping of QoR4D Proposal Review Criteria and Elements against Evaluation Framework Standards and Evaluation Criteria

Eval Framework standards (CGIAR, 2022) *Ex post performance/process evaluation*	**Quality of Research for Development** **QoR4D in Practice for One CGIAR (2021)** *Ex ante proposal review criteria*	**QoR4D Frame of Reference (2020)** *QoR4D elements*	**Eval Policy criteria (CGIAR, 2022)** *Ex post performance/ process evaluation*
1. *Relevance, use, and utility:* All evaluations are applicable to the question(s) at hand and designed in a respon sive and timely manner for use in decision-making, accountability, and learning processes.	1. Clearly defined research problem that addresses Impact Areas, is a high priority in the targeted geographies, is well aligned to shared, multi-funder priorities, and is well informed by previous research findings.	Relevance **Effectiveness**	Relevance Coherence Effectiveness
2. *Independence and lack of bias:* Evaluations instil confidence among all users that the evaluation is as objective as possible with the highest ethical standards and codes of conduct; impartial, with a system in place against conflicts of interest; and unbiased operationally and analytically.	8. Ethics, including equitable partnerships, information disclosure, biases, and potential conflicts of interest are considered; proposal defines how formal research ethics approvals will be sought/granted.	**Legitimacy** Credibility	Quality of Science
3. *Transparency:* Processes (including methods) and results are transparently disclosed, traceable, and accessible to the public.	14. Justified and transparent costing explicitly linked to expected Research for Development results. 15. Anticipated research outputs (knowledge, technical, or institutional advances, specific technologies or products, policy analyses) are described and knowledge/ gaps they will fill are evident with a demonstrated focus on quality, forward-looking, and impact relevance and how they will be disseminated. Protocols for open-data and open-access compliance are evident in plan (including budget).	**Legitimacy** Effectiveness **Credibility** Effectiveness	Quality of Science Effectiveness Efficiency
4. *Legitimacy and participation:* Evaluations include relevant informants and use consultative processes to prepare terms of reference and the evaluation matrix. Such processes ensure the quality of the process, including, where appropriate and feasible, representatives of end and intermediate users of evaluation outputs.	2. Evidence that the Initiative is demand driven through code-sign with key stakeholders and partners (Investment Advisory Groups, governments, private sector, funders) and research collaborators within and outside CGIAR. 11. Capacity statements indicate why the proponents are the ideal implementers for the work. The value proposition is stated and CGIAR capacity and appropriateness to lead the work is justified. This includes the skills, diversity and multi-/ trans-disciplinarity of the research team and collaborators.	**Relevance** Effectiveness	Relevance Coherence Effectiveness Quality of Science Sustainability
5. *Responsiveness* to gender, diversity, and inclusion (GDI): Evaluation design and conduct, the commissioning of teams, and the reporting strive to fully address GDI parameters. Evaluations will consider who is engaged in the work and who benefits from it.	9. Research design and proposed implementation demonstrates gender and social inclusion that can be tracked in outcomes	**Legitimacy** Effectiveness	Relevance Effectiveness Quality of Science Impact
6. *Ethics and equity:* Evaluations consider questions of ethics in research and outcomes and integrate ethical and equity considerations in the evaluation design and implementation.	8. Ethics, including equitable partnerships, information disclosure, biases, and potential conflicts of interest are considered; proposal defines how formal research ethics approvals will be sought/granted	**Legitimacy** Credibility	Relevance Effectiveness Quality of Science

Annex 5: Mapping of QoR4D Proposal Review Criteria and Elements against Evaluation Framework Standards and Evaluation Criteria cont'd

Eval Framework standards (CGIAR, 2022) *Ex post performance/process evaluation*	**Quality of Research for Development** **QoR4D in Practice for One CGIAR (2021)** *Ex ante proposal review criteria*	**QoR4D Frame of Reference (2020)** *QoR4D elements*	**Eval Policy criteria (CGIAR, 2022)** *Ex post performance/ process evaluation*
7. *Evaluability*: Evaluability refers to the extent to which an intervention can be evaluated in a reliable and credible fashion; the concept is central to a culture of results. A strong focus on evaluability at the design stage facilitates overall measurability, monitoring, and subsequent evaluation.	15. Anticipated research outputs (knowledge, technical, or institutional advances, specific technologies or products, policy analyses) are described and knowledge/ gaps they will fill are evident with a demonstrated focus on quality, forward-looking, and impact relevance and how they will be disseminated. Protocols for open-data and open-access compliance are evident in plan (including budget).	**Credibility** Effectiveness	Effectiveness Efficiency Quality of Science
	16. Monitoring and evaluation (M&E) plan for the Initiative is clearly defined, with flexibility to adapt. M&E plan supports effective management and learning, including baseline data collection, and evaluative and review processes corresponding to stage-gates and course-correction decisions. M&E occurs during the life of Initiative and is used proactively to reflect on and adapt the Theory of Change, where appropriate.	**Credibility** Effectiveness Legitimacy	
8. *Credibility and robustness:* Methods employed are credible and replicable. The quality of an evaluation depends on the professional and methodological competency of the evaluators and the use of reliable, triangulated data.	5. Research methodology and methods (and supporting activities) are fit-for-purpose, feasible, are state-of-the-art, and rigorous in data collection and analysis, and limitations clearly stated	**Credibility** Relevance Effectiveness	Quality of Science
9. *Measurability:* Sound methods underpin measurability and replicability. To the extent possible, evaluations measure, using quantitative and/or qualitative methods, the performance of CGIAR. Measurability provides comparability between time frames, groups, or alternative theories.	3. Research questions, objectives, outputs, and outcomes are aligned to the research problem, are measurable with well-defined milestones and stages amenable for assessment and corrective action through the project lifecycle.	**Relevance** Effectiveness	Coherence Effectiveness
10. *Mutual accountability:* In CGIAR, expectations for evaluation are matched with adequate investments in requisite financial and human resources. The capacity and systems for data collection and real-time information underpin mutual accountability.	13. Project management mechanisms and (if applicable) additional scientific oversight and governance measures effectively and efficiently support the Initiative objective	**Legitimacy** Credibility	Quality of Science Efficiency
	14. Justified and transparent costing explicitly linked to expected Research for Development results	**Legitimacy** Efficiency	
11. *Efficiency:* Evaluation avoids unnecessary duplications, costs, or redundancy to other evaluative assessment.	6. Analysis of trade-offs and synergies across the CGIAR Impact Areas; ex-ante assessment of project benefits provides logical rationale for scaling of impacts	**Effectiveness** Credibility	Coherence Effectiveness Efficiency Impact
	14. Justified and transparent costing explicitly linked to expected Research for Development results	**Legitimacy** Efficiency	

Annex 5: Mapping of QoR4D Proposal Review Criteria and Elements against Evaluation Framework Standards and Evaluation Criteria cont'd

Eval Framework standards (CGIAR, 2022) *Ex post performance/process evaluation*	**Quality of Research for Development** **QoR4D in Practice for One CGIAR (2021)** *Ex ante proposal review criteria*	**QoR4D Frame of Reference (2020)** *QoR4D elements*	**Eval Policy criteria (CGIAR, 2022)** *Ex post performance/ process evaluation*
12.*Comparative advantage:* Evaluation gives due consideration to exploring the comparative advantage of CGIAR in contributing to the achievement of qualityresearch-for-development results.	6. Analysis of trade-offs and synergies across the CGIAR Impact Areas; ex-ante assessment of project benefits provides logical rationale for scaling of impacts 7. Evidence that the Initiative will likely lead to impact at scale through integrated systems approaches that drive innovation in research and partnerships, including linking to and leveraging of other Initiatives within and outside CGIAR	**Effectiveness** Credibility Relevance	Coherence Effectiveness Impact
13.*Fairness*, confidentiality, and no harm: The evaluators and commissioning office(s) are responsible for ensuring and protecting the confidentiality and anonymity of information, as required. In line with a do-no-harm approach, evaluators attend to actions, omissions, and unconscious choices throughout evaluation design and implementation.	8. Ethics, including equitable partnerships, information disclosure, biases, and potential conflicts of interest are considered; proposal defines how formal research ethics approvals will be sought/granted	Credibility	Relevance Effectiveness
14.*System framing and complexity awareness:* Evaluations consider the contextual realities in terms of boundaries, interrelationships, dynamics, and perspectives that delineate the systems that CGIAR aspires to improve incrementally or to transform. Evaluation attends to nonlinearities, emergence, uncertainties, turbulence, and adaptive capacity, in line with complexity awareness.	10.A risk framework that details main project risks and mitigation actions, including intended and unintended consequences of technologies/innovations for natural resources, GHG emissions, and social and economic aspects	**Credibility** Legitimacy Relevance	Coherence Effectiveness Efficiency
15.*Capacity building:* Learning and evaluation-related capacity building will be embedded into evaluation practice to promote coherent monitoring, evaluation, and learning (MEL).	12.Capacity building within project teams, partners, and stakeholders evident in project activities. This can include development of early career researchers and partner staff, support/empowerment for under-represented stakeholders, building partner networks	**Credibility** Legitimacy	Effectiveness Quality of Science Sustainability
16.*AR4D Items:* Use of theories of change and theory-based approaches: Theories of change (ToC) describe the pathways to impact-which can be complex, intersecting, and often nonlinear-drawing on insights from the social sciences, including economic and international relations theory. When theory-driven interventions are evaluated, the evaluations assess the relevance of the ToC against the development problem, including the assumptions and risks it describes, and use of the ToC towards measuring and explaining results and conditions for achieving outcomes and ultimate impact.	4. Theory of Change with intended outputs, outcomes, and impacts at scale clearly described. Assumptions are documented, causal linkages are clear, especially the role of partners in driving impact, and all indicators including stagegate indicators made explicit. 17. Well-defined plan for Initiative-level evaluation and impact assessment based on expected end-of-Initiative outcomes and impact. Links between the impact assessment plan and indicators in the Theory of Change are clear.	**Effectiveness** Relevance **Effectiveness** Relevance	All

Annex 6: Evaluation Design Matrix, Quality of Science Criterion

Dimension	Evaluation sub-question	Elements to be assessed	Assessment criteria	Type of Method	Data sources, methods, analysis	Evaluation policy
QoS_EQ1: Is research design appropriate to the development challenges in the context?						
Design	1.1. Does the research behind the intervention objective align to shared, multi-funder priorities?	Research relevance	Relevance and coherence of the research agenda.	Qualitative	Theory of Change Interventions' reports; Interviews; Rubrics	Relevance
	1.2. Has the comparative advantage been systematically assessed and documented?	Research relevance and coherence		Mixed	Trade-off analysis	Relevance Coherence
	1.3. Is the link between the impact assessment plan and indicators in the ToC clearly defined in the research design?	Research design	Rigor of the experimental research	Qualitative	Theory of Change Interventions' reports; Interviews; Rubrics	Relevance Quality of Science
	1.4. How interconnected is the research design to SDGs including within each SDG?	Research relevance	Alignment with SDGs	Qualitative/ Quantitative	Theory of Change Interventions' reports; Bibliometrics; Rubrics	Relevance
	1.5. Is the research design of the intervention appropriate and clearly articulated?	Research design	Rigor of the experimental research	Qualitative	Interventions' reports; Interviews	Relevance Quality of Science
	1.6. Are research methodology and methods fit-for-purpose for an intervention?	Research design	Rigor of the experimental research	Qualitative	Interventions' reports; Interviews	Relevance Quality of Science
	1.7. Does a defined objective of an intervention address CGIAR Impact Areas?	Research relevance	Relevance and coherence of the research agenda	Qualitative	Theory of Change Interventions' reports; Interviews	Relevance
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate?						
Inputs	2.1. To what extent were necessary inputs adequate and sufficient to deliver planned outputs and outcomes?	Composition of research teams	Adequacy of skills and scientific disciplines; level of multi- & trans- disciplinarityintegration; Inclusiveness in relation to diversity of age, gender, and nationality	Quantitative Qualitative	Interventions' reports; Bibliometrics; Interviews; Social Network Analysis	Quality of Science
		Attractiveness of research team	Attractiveness of team members	Qualitative	Interviews; team members profile; Interventions' reports	Quality of Science
	2.2. Are previous research outputs/ findings clearly described and learning integrated?	Reputation of research unit	Scientific reputation of the unit; recognition gained through the success in competitive calls for projects	Qualitative	Interviews; Interventions' reports	Quality of Science
		Funding	Adequacy and predictability; commitment of donors	Quantitative Qualitative	Interviews; Interventions' reports	Quality of Science

Annex 6: Evaluation Design Matrix, Quality of Science Criterion cont'd

Dimension	Evaluation sub-question	Elements to be assessed	Assessment criteria	Type of Method	Data sources, methods, analysis	Evaluation policy
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate? cont'd						
Inputs	2.1. To what extent were necessary inputs adequate and sufficient to deliver planned outputs and outcomes? 2.2. Are previous research outputs/ findings clearly described and learning integrated?	Research infrastructures	Adequacy of laboratories and fields	Qualitative	Interventions' reports; Interviews	Efficiency QoS
		Capacity building	Appropriate and adequate, useful to planned activities	Qualitative	Interviews; reports	Efficiency QoS
		Comparative advantage	Best knowledge available, identifying the relative costs of the key deliverables among the identified organizations, including CGIAR.	Qualitative	Reports	Coherence Efficiency Effectiveness
		Research relevance	Research design is appropriate and builds on filling evident gaps; appropriate and comprehensive literature review	Qualitative	Theory of Change Interventions' reports; Interviews	Relevance Effectiveness
		Comparative advantage	Best knowledge available, identifying the relative costs of the key deliverables among the identified organizations, including CGIAR.	Qualitative	Reports	Coherence Efficiency Effectiveness
Processes	2.2. To what extent did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations?	Stakeholder's involvement	Appropriate stakeholders involved at the right stage	Qualitative	Interventions' reports; Interviews; FGD SNA	Quality of Science Coherence
		Partnerships	Multistakeholder approach; Mutual trust, understanding, and commitment; Clear recognition of partners' perspectives, needs, roles, and contributions, comparative advantage, including resilience[32]	Qualitative	Interventions' reports; Interviews; FGD SNA, Rubrics	Quality of Science Coherence
		Research ethics	Policies in place for research ethics and their implementation	Qualitative	Interventions' reports; Interviews; FGDSNA	Quality of Science

[32]Independent Science for Development Council. 2022. Identifying and Using CGIAR's Comparative Advantage. Rome: CGIAR Independent Advisory and Evaluation Service. https://iaes.cgiar.org/sites/default/files/pdf/ISDC-Technical-Note-Identifying-and-Using-CGIAR-Comparative-Advantage.pdf

Annex 6: Evaluation Design Matrix, Quality of Science Criterion cont'd

Dimension	Evaluation sub-question	Elements to be assessed	Assessment criteria	Type of Method	Data sources, methods, analysis	Evaluation policy
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate? cont'd						
Processes	2.2. To what extent did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations?	Engagement with local knowledge	Local communities, stakeholders or populations were effectively engaged and have been considered in the research process	Qualitative	Interventions' reports; Interviews; FGD	Quality of Science
		Roles and responsibility	Clearly defined roles and responsibilities	Qualitative Quantitative	Document review, Survey, Interviews; FGD	Quality of Science
		Internal review mechanisms	Policies in place for internal review mechanisms and their implementation	Qualitative	Interventions' reports; Interviews; FGD	Quality of Science
		Mentoring and training of junior staff	Policies in place for mentoring and training of junior staff and their implementation	Qualitative	Interventions' reports; Interviews; FGD	Quality of Science
		Gender	Gender, diversity and inclusion in implementation	Qualitative Quantitative	Interviews, FGD GDI dashboard	Quality of Science
		Performance evaluation	Quality work is rewarded	Qualitative	Interviews, FGD Survey	Effectiveness QoS
		Risk management	Risk assessment and mitigation strategies are put in place	Qualitative	Internal audit reports, Interviews	Efficiency QoS
		Protocols for open-data and open-access compliance (including budget)	Accessibility of data and information	Qualitative Quantitative	Bibliometrics; Interviews	
QoS_EQ3: How do the intervention's outputs contribute to advance science?						
Outputs	3.6. In what ways are the research outputs, such as improved varieties, knowledge tools, and publications, of high quality?	Quality and quantum of scientific and technical publications	Number of publications; H index of most productive authors; Impact factor of journals; share of highly cited publications (HCP); Citation distribution index (CDI); Average of relative citation (ARC); multi-disciplinarity integration; Altmetrics scores.	Quantitative	Altmetrics ; Bibliometrics ; Interventions' reports	Quality of Science

Annex 6: Evaluation Design Matrix, Quality of Science Criterion cont'd

Dimension	Evaluation sub-question	Elements to be assessed	Assessment criteria	Type of Method	Data sources, methods, analysis	Evaluation policy
QoS_EQ3: How do the intervention's outputs contribute to advance science? cont'd						
	3.7. Were outputs reflected in new policies and/or contributed to the society where change is sought?	Policy linkages	Policies citing the research products; network with governments; impact studies produced; policy makers engagement	Quantitative	Reports (including impact assessments if available); Interviews; Social Network Analysis; Theory of Change	Quality of Science Relevance Effectiveness
Outputs	3.8. Were physical products, e.g., improved varieties and digital innovation, of high quality and relevant to next stage users?	Development of physical products, e.g., improved varieties and digital innovations	Broader applicability; Adaptability of the physical product to the context; Scaling readiness; Relevance for target audience.	Qualitative Quantitative	Interventions' reports; Interviews; FGD; Theory of Change	Effectiveness Sustainability
	3.9. Were research findings clearly communicated?	Communication of research findings	Relevance to target audiences	Qualitative	Interventions' reports; Interviews	Relevance

Annex 7: Use of Rubrics for Qualitative Assessment of Quality of Science

Dimension	Evaluation question	Elements to be assessed	Assessment criteria	No	Partly	Yes
QoS_EQ1: Is research design appropriate to the development challenges in the context?						
Design	1.1. Does the research behind the intervention objective align to shared, multi-funder priorities?	Research relevance	Relevance and coherence of the research agenda.	The research does not align to shared, multi-funder priorities	The research aligns only partly to shared, multi-funder priorities	The research fully aligns to shared, multi-funder priorities
	1.2 Has the comparative advatage been systematically assessed and documented?	Research relevance and coherence		The comparative advantage was not assessed and documented	There is partial, not systematic evidence on comparative advantage	The intervention has a clearly documented systematic assessment of exiting comparative advantage of the intervention in the context
	1.3. Is the link between the impact assessment plan and indicators in the ToC clearly defined in the research design?	Research design	Rigor of the experimental research	In the research design it is not clearly defined the relation of the study with the ToC	ToC is mentioned but its link with the impact assessment plan is not clearly defined in the research design	The link between the impact assessment plan and indicators in the ToC is clearly defined in the research design
	1.4. How interconnected is the research design to SDGs including within each SDG?	Research relevance	Alignment with SDGs	The research design does not align with SDGs	The research design aligns with at least one SGD	The research design clearly aligns with more SDGs and is relevant also to other SDGs indirectly
	1.5. Is the research design of the intervention appropriate and clearly articulated?	Research design	Rigor of the experimental research	There was not clearly articulated research design	Research design was articulated but left some gaps	Research design was appropriate and clearly articulated
	1.6. Are research methodology and methods fit-for-purpose for an intervention?	Research design	Rigor of the experimental research	The research design did not adhere to methodological standards and are not fit for purpose	Adherence to methodological standards was partly achieved	Research methodology and methods are fit-for-purpose
	1.7. Does a defined objective of an intervention address CGIAR Impact Areas?	Research design	Rigor of the experimental research	Research questions are not clearly stated and/or are not aligned to the research problem	Research questions are only partly aligned to the research problem	Research questions are clearly stated and address properly the research problem
	3.4. Does a defined research problem address Impact Areas?	Research relevance	Relevance and coherence of the research agenda	The research problem does not directly relate to any Impact Areas	The research problem refers to Impact Areas but the link is not clearly defined	The defined research problem clearly addresses Impact Areas

Annex 7: Use of Rubrics for Qualitative Assessment of Quality of Science cont'd

Dimension	Evaluation question	Elements to be assessed	Assessment criteria	No	Partly	Yes
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate?						
Inputs	2.1.To what extent were required inputs adequate and sufficient to deliver planned outputs and outcomes?	Composition of research teams	Adequacy of skills and scientific disciplines; level of multi- & trans-disciplinarity integration	Low level of integration. There's lack of diversity in terms of skills and there is low disciplinarity integration.	Adequate level of integration in terms of skills and scientific disciplines but there is space for further improvement	High level of integration in terms of skills and scientific disciplines
		Composition of research teams	Inclusiveness in relation to diversity of age, gender, and nationality	Low level of integration. There's lack of diversity within the teams in relation to gender, age and/or nationality	Adequate level of integration but there is space for further improvements	High level of integration and diversity within the research teams
		Attractiveness of research team	Attractiveness of team members	Team members are not attractive for their caliber	Team members have good profiles but lack some experience to be fully attractive	High caliber of teams members, high attractiveness
		Attractiveness of research unit	Scientific reputation of the unit; recognition gained through the success in competitive calls for projects	The unit has not a record of winning competitive calls for projects	The unit has a record of winning some competitive calls for projects (between 1 and 5)	The unit is known for having won several competitive calls for projects (more than 5)
		Funding	Adequacy and predictability; commitment of donors	Insufficient fundings. Poor commitment of donors.	Good level of funding but not predictable. Donors are partly committed	Fundings are adequate and predictable. Donors are highly committed
		Research infrastructures	Adequacy of laboratories and fields	Laboratories and fields and not adequate for research purposes	Laboratories and fields are partly adequate. Some improvements are needed	Laboratories and fields are adequate to conduct research
		Capacity building	Appropriate and adequate, useful to planned activities	There was a lack for capacity building activities throughout the project	There were some capacity building activities but the focus on them was limited	There were numerous and good capacity building activities throughout the project.
	2.2. Are previous research outputs/ findings clearly described and integrated?	Research relevance	Research design is appropriate and builds on filling evident gaps; appropriate and comprehensive literature review	Literature/documental review, if at all evident, was insufficient and largely outdated	Literature/document review was appropriate but not fully exhaustive or not fully integrated	Literature/ document review was appropriate and exhaustive

Annex 7: Use of Rubrics for Qualitative Assessment of Quality of Science cont'd

Dimension	Evaluation question	Elements to be assessed	Assessment criteria	No	Partly	Yes
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate? cont'd						
Processes	2.2. To what extend did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations?	Stakeholder's involvement	Appropriate stakeholders involved at the right stage	Qualitative	Interventions' reports; Interviews; FGD SNA	Quality of Science
		Partnerships	Mutual trust, understanding, and commitment; Clear recognition of partners' perspectives, needs, roles, and contributions; Multistakeholder approach	Relationship with partners was not clear. It was difficult to find alignments of needs with partners. Partners' contribution was weak	Good commitment of some partners and with differences in engagement throughout the duration of the intervention	The relationship with partners has been good throughout the whole duration of the project. Partners' perspectives, needs, roles and contributions were always clear
		Research ethics	Policies in place for research ethics and their implementation	There are not policies in place for research ethics and/or policies in place were not respected	Policies for research ethics were not exhaustive and/or only partly applied	Policies for research ethics were exhaustive and widely applied
		Engagement with local knowledge	Local communities, stakeholders or populations were effectively engaged and have been considered in the research process	Engagement with appropriate contexts has been neglected during the research process.	Contexts and engagement have been considered during the research process	Engagement with local communities, populations or stakeholders happened in an appropriate and credible manner
		Roles and responsibility	Clearly defined roles and responsibilities	Roles and responsibilities were not clearly defined	Roles and responsibilities were partly clear. During the research some aspects have been identified as unclear	Roles and responsibilities were clearly defined throughout all duration of the research
		Internal review mechanisms	Policies in place for internal review mechanisms and their implementation	Lack of policies for internal review mechanisms	Policies are in place but not exhaustive/not fully applied	Policies are in place and fully applied throughout the entire research process
		Mentoring and training of junior staff	Policies in place for mentoring and training of junior staff and their implementation	There we no policies or programs to mentor and train junior staff	Mentoring and training for junior staff was in place but in need for improvement	Good policies in place for mentoring and training of junior staff. Policies were widely implemented

Annex 7: Use of Rubrics for Qualitative Assessment of Quality of Science cont'd

Dimension	Evaluation question	Elements to be assessed	Assessment criteria	No	Partly	Yes
QoS_EQ2: Are inputs and processes appropriate to produce science that is credible and legitimate? cont'd						
Processes	2.2. To what extend did the management process ensure the Quality of Science, including scientific credibility, and legitimacy, of the research and operations?	Gender	Gender, diversity and inclusion in implementation	There was no consideration of gender balance and roles in the research team.	Limited gender consideration was shown in the composition and roles of the research team.	Emphasis was given to gender balance and appropriate roles in the research team
		Performance evaluation	Quality work is rewarded	There is not a system to evaluate and reward performance	There is a system in place to evaluate performance, but quality work has not always been rewarded	Performance was evaluated during the research process and quality work has been rewarded
		Risk management	Risk assessment and mitigation strategies are put in place	Absence of risk assessment and mitigation strategies	Risk assessment and mitigation strategies were identified but not always put in place	Risk assessment was well done and mitigation strategies were put in place when necessary
		Protocols for open-data and open-access compliance (including budget)	Accessibility of data and information	Lack of data and information	Data and information only partly available	Data and information fully available. Presence of protocols for open-data and open-access compliance
EQ3: How do the intervention's outputs contribute to advance science?						
Outputs	3.1. Were physical products, e.g., improved varieties and digital innovations, of high quality and relevant to next stage users?	Development of physical products, e.g., improved varieties and digital innovations	Broader applicability; Adaptability of the physical product to the context; Scaling readiness; Relevance for target audience.	Products are not applicable broadly and are not ready for scaling	Products can be partly applied. Product are not easy to be scaled	Applicability of the product. The product is ready for scaling. It is relevant for target audience
	3.2. Were research findings clearly communicated?	Communication of research findings	Relevance to target audiences	Research findings are not well communicated	Research findings are communicated but not all means are used to reach the target audiences	Research findings are well communicated, and target audiences easily reached

Annex 8: Bibliometrics: Glossary and Priority Indicators

ID: Both an alphabetical reference to the QoR4D dimension of relevance and a unique numeral.

Indicator title: Name of the indicator.

Implementation: Implementation modality (by whom and when):
- CGIAR +: Could be implemented in house by PPU, the MEL community, or CAS-engaged analysts on recommendation from Science-Metrix in the future.
- Extern: Would have to be implemented by an external provider in the future.
- Pilot: Indicator still in design; may be implemented by PPU, the MEL community, or CAS-engaged analysts or external providers, but in all cases requires some R&D, with no guarantee of success.

Time: Number of years after a project concludes during which publications produced through that project can be assessed (considering that relevant publications are still released in the two years immediately following the last formal year of a project).

Limits: A typology of generic limitations includes the following:
- Un-normalized: Indicator is not currently or can never be normalized to control for field biases and yearly trends.
- Cleaning: Requires substantial efforts to harmonize metadata.
- Unknown optimum: Current knowledge does not fully allow for determining a best practice in the dimension measured by this indicator; high scores on the measurement may have adverse effects on research practices.
- Imperfect proxy: Indicator captures only a narrow component of a broader phenomenon of interest.
- May capture tokenism: Quantitative indicators of equity among groups typically do not capture fully realized equity, but only outward manifestations of equity. This limitation overlaps with the imperfect proxy limitation.
- Complex categorical definition: Assigning an output to a category may rely on judgment or necessarily imperfect guidelines.
- Metadata errors: There are recognized shortcomings to the metadata typically used to compute this indicator, either because publication authors themselves make mistakes, or because coding and parsing in bibliographic databases are imperfect
- Discrepancies between plans and achievements: Project proposals and project realization may differ greatly.

ID: Both an alphabetical reference to the QoR4D dimension of relevance and a unique numeral.
Indicator title: Name of the indicator.

Table 10: *Sample Data Collection Matrix for high priority bibliometric indicators*[33]

ID	Title	Implementation	Time (in years)	Limits
L23	Share of publications with women's participation in authorship	CGIAR+	+3	Does not capture balance or equity; may capture tokenism; paying software (NamSor); margin of error (especially for Asian names)
L24	Share of publications achieving gender balance in key authorship	CGIAR+	+3	Paying software (NamSor); margin of error (especially for Asian people)
L26	Share of North-South/South-South co-publications	CGIAR+	+3	Un-normalized; cleaning; unknown optimum; imperfect proxy; does not capture balance or equity
L27	Southern authors' participation as first, corresponding, or last author	CGIAR+	+3	Error rate in affiliation data; imperfect proxy (South-North equity);
L31	Chord diagram visualization of international co-publications	Pilot	+3	Metadata errors (affiliation data); limited knowledge base (novel indicator); imperfect proxy (equity in multinational integration)
R31	Thematic alignment with SDG-relevant topic	CGIAR+	+3	Imperfect proxy (knowledge transfer for development); limited knowledge base; metadata errors
R34	Share of academic-private co-publications	CGIAR+ or Extern	+3	Difficult normalization; extensive cleaning; complex categorical definition; imperfect proxy (technology transfer)
R38	Share of highly cited publications (HCP)	Extern	+5	Imperfect proxy (publication quality and intellectual achievement); 30 publications or more required; computable 2 years or more after publication year
R39	Citation distribution index (CDI)	Extern	+5	Imperfect proxy (publication quality and intellectual achievement); 30 publications or more required; computable 2 years or more after publication year
R41	Average of relative citations (ARC)	CGIAR+	+5	Imperfect proxy (publication quality and intellectual achievement); sensitive to outliers; 30 publications or more required; computable 2 years or more after publication year
R42	Index of interdisciplinary integration	Extern	+3	Imperfect proxy (intellectual disciplinary integration); bias toward novel and radical interdisciplinarity; abstract index most meaningful as part of comparisons
R43	Share of highly interdisciplinary publications	Extern	+3	Imperfect proxy (intellectual disciplinary integration); bias toward novel and radical interdisciplinarity

[33] Bibliometric Analysis to Evaluate Quality of Science in the Context of One CGIAR | CAS | CGIAR Advisory Services

Table 10: *Sample Data Collection Matrix for high priority bibliometric indicators (cont'd)*

ID	Title	Implementation	Time (in years)	Limits
R44	Index of multidisciplinary integration	Extern	+3	Imperfect proxy (collaborative disciplinary integration); bias toward novel and radical disciplinary diversity
R45	Share of highly multidisciplinary publications	Extern	+3	Imperfect proxy (collaborative disciplinary integration); bias toward novel and radical disciplinary diversity
R46	Chord diagram visualization of interdisciplinarity (notably to capture social sciences and humanities integration)	Extern	+3	Imperfect proxy (interdisciplinary integration); bias toward novel and radical interdisciplinarity

CGIAR | Independent Advisory and Evaluation Service

Contact

CGIAR Independent Advisory and Evaluation Service (IAES)
Via di San Domenico 1, 00153 Rome, Italy
Email: IAES@cgiar.org
URL: https://iaes.cgiar.org/evaluation

Follow the Independent Advisory and Evaluation Service on social media: