CGIAR | Advisory Services

Bibliometric Analysis to Evaluate Quality of Science in the Context of One CGIAR

Technical Note

March 2022

Citation: 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.

Cover image: Stock MS Office

Design and layout: Luca Pierotti and Macaroni Bros

Bibliometric Analysis to Evaluate Quality of Science in the Context of One CGIAR

Technical Note

Science-Metrix, Elsevier

March 2022

Acknowledgments

Science-Metrix was commissioned by Evaluaton Fucntion of the CGIAR Advisory Services (CAS) in 2021 to develop this technical note. Science-Metrix expresses gratitude to all who contributed to the consultative process, especially the CAS team (Jillian Lenne, Svetlana Negroustoueva, Max Runzel, Paolo Sarfatti, Stefania Sellitti, and Allison Smith) and CGIAR colleagues (Enrico Bonaiuti, Valentina De Col). We would also like to thank focus group participants Paul Engel, Margaret Gill, and Ravi Kumar; interview participants Brian Belcher, Ben Bennet, Bas Bouman, Michael Friedmann, Robert McLean, Raphael Nawrotzki, Victor Sadras, and Rachid Serraj; and survey participants Rex Dunham, Paul Engel, Victor Sadras, Laurian Unnevehr, and Rattan Yadav. The contents of the technical note benefited from review by the CAS evaluation reference group members Guy Poppy and Zenda Ofir, and inputs from a validation meeting with Helen Altshul, Brian Belcher, Enrico Bonaiuti, Valentina De Col, Karen Garrett, Jillian Lenne, Svetlana Negroustoueva, Guy Poppy, Paolo Sarfatti, Stefania Sellitti, Allison Smith, and Graham Thiel.

Contents

Figures

Tables

Boxes

Acronyms

A4NH	CGIAR Research Program on Agriculture for Nutrition and Health
AR4D	Agricultural Research for Development
API	Application Programming Interface
CAS	CGIAR Advisory Services
CAS Secretariat	CGIAR Advisory Services Shared Secretariat
CCAFS	CGIAR Research Program on Climate Change, Agriculture and Food Security
CLARISA	CGIAR Level Agricultural Results Interoperable System Architecture
CRP	CGIAR Research Program
DCM	Data Collection Matrix
DID	Difference-in-Differences
DOI	Digital Object Identifier
FAIR	Findability, Accessibility, Interoperability, and Reusability
Fish	CGIAR Research Program on Fish
FTA	CGIAR Research Program on Forests, Trees and Agroforestry
FTE	Full-time Equivalent
GLDC	CGIAR Research Program on Grain Legumes and Dryland Cereals
ICARDA	International Center for Agricultural Research in the Dry Areas
IDO	Intermediate Development Outcome
IEA	CGIAR Independent Evaluation Arrangement
ISDC	Independent Science for Development Council
KI	Key Informant
KII	Key Informant Interview
LIVESTOCK	CGIAR Research Program on Livestock
Maize	CGIAR Research Program on Maize
MARLO	Managing Agricultural Research for Learning and Outcomes
MEL	Monitoring, Evaluation, and Learning
MIS	Management Information System
M-QAP	Monitoring, Evaluation, and Learning Quality Assurance Processor
NARS	National Agricultural Research System

OA	Open Access
OD	Open Data
OICR	Outcome Impact Case Report
PCU	Project Coordination Unit
PIM	CGIAR Research Program on Policy, Institutions, and Markets
PRMF	Performance and Results Management Framework
PPU	Performance Portfolio Unit
PRMS	Performance and Results Management System
QA	Quality Assurance
QoR	Quality of Research
QoR4D	Quality of Research for Development
QoS	Quality of Science
Rice	CGIAR Research Program on Rice
RTB	CGIAR Research Program on Roots, Tubers and Bananas
SLO	System-Level Outcome
SME	subject matter expert
SPIA	Standing Panel on Impact Assessment
SRF	Strategy and Results Framework
TN	Technical Note
ToC	Theory of Change
WHEAT	CGIAR Research Program on Wheat
WLE	CGIAR Research Program on Water, Land and Ecosystems
WoS	Web of Science

Executive Summary

One CGIAR is a reformulation and integration of CGIAR's capabilities, knowledge, assets, people, and global presence for a new era of interconnected and partnership-driven research towards achieving the Sustainable Development Goals (SDGs). Science-Metrix developed this technical note to provide recommendations on the use of bibliometrics in mixed-methods evaluations and potential monitoring of Quality of Science (QoS) in One CGIAR research. As such, this technical note combines bibliometric expertise with input from CGIAR stakeholders and lessons from the independent evaluative reviews of the CRPs.

The recommendations for the use of bibliometrics in the evaluation of QoS in One CGIAR research are as follows:

R1. Expand and systemize metadata curation throughout CGIAR.

R2. Require funding acknowledgment based on a formal template which includes mention of funding sources, CGIAR Center, Initiative and Action Area.

R3. Expand and pilot test the use of qualitative indicators beyond those used in the 2020 CRP reviews.

R4. Expand the mixed-methods approach used in CRP reviews as part of QoS evaluations.

R5. Ensure the PPU, the MEL community professionals, and/or CAS/Evaluation-engaged analysts are responsible for formulating and computing bibliometrics used as part of evaluation of QoS.

R6. Conduct bibliometric components of evaluating QoS at a minimum three years after the completion of a group of projects or a portfolio (non-pooled or initiative projects).

R7. (For comparative analyses using MEL metadata where equivalent metadata on non-CGIAR publications is of much lower quality) repeat same analysis using both datasets and report findings next to one another in a multidimensional panel of indicators.

R8. Conduct a 2025 interim evaluation for 2022–24 Initiatives using QoS criteria based on qualitative assessments and a restricted set of bibliometric indicators and conduct a comprehensive and targeted evaluation of QoS in the 2022–24 Initiatives in 2030.

R9. Provide normalized versions of indicators of citation impact, and possibly of other indicators, to SMEs or external reviewers evaluating QoS at CGIAR.

R10. Increase the use of reference levels and comparative strategies in evaluation of QoS to support more robust interpretation of bibliometric findings.

R11. Develop thematic queries to delineate a foundational AR4D publication set that will enable comparative assessments of CGIAR research against external institutions.

R12. Only include publications that have been written in South-North co-authorship when comparing to Northern institutions.

R13. Simultaneously deploy a panel of bibliometrics that together cover dimensions of credibility, effectiveness, legitimacy, and relevance for One CGIAR evaluations of QoS.

R14. Conduct a bibliometric pilot study to ease the implementation of the recommendations contained in this report pertaining to bibliometrics expertise and production.

R15. Explore setting quantitative targets to guide bibliometric assessments.

R16. Assess the RQ+ set of contextual factors for One CGIAR projects and then cluster projects along contextual profiles when evaluating QoS.

R17. Monitor upcoming developments on the use of databases such as Google Scholar and altmetrics databases to capture societal outcomes linked to a larger set of documents than just journal-based publications.

The development of this technical note and its recommendations followed a consultative and co-design process. Recommendations on the use of indicators for evaluating QoS were to be drawn from the use of bibliometric indicators in the evaluation of QoS in the 2020 CRP evaluative reviews, the QoR4D framework, consultations with the CAS Secretariat Evaluation Function, and a focus group, interviews, and surveys with stakeholders and subject matter experts (referred to herein as "key informants"). Through these recommendations, Science-Metrix shares some of the best practices in bibliometrics and information sciences, as well as some of the best practices of private bibliometric service providers, in order to provide a high-level framework for designing and conducting or provisioning bibliometric analyses by CAS-engaged analysts; the Performance Portfolio Unit (PPU); and the monitoring, evaluation, and learning (MEL) community in One CGIAR. In view of potential challenges such as changing data sources, unexpected metadata errors, or other obstacles that are identified and defined only when implementation of a project is launched, Science-Metrix advises analysts to undertake frequent reassessments of the overall framework by triangulating with the empirical experience collected during implementation.

1 Background and Context

CGIAR is a global research partnership for a food-secure future. The CGIAR reform of 2020 restructured CGIAR's partnerships, knowledge, and operations to create One CGIAR, the CGIAR System whose mission is to end hunger by 2030 through science to transform food, land, and water systems in a climate crisis. One CGIAR integrates CGIAR's capabilities, knowledge, assets, people, and global presence for a new era of interconnected and partnership-driven research aimed at achieving the Sustainable Development Goals (SDGs).

In 2011 CGIAR research shifted from being driven largely by its global Research Centers to being centered on 12 cross-Center CGIAR Research Programs (CRPs) and four platforms. Now, under One CGIAR, CGIAR research aims to become further integrated. One CGIAR will allocate pooled funding across approximately 30 portfolio initiatives spanning three Action Areas, partnerships between funders and Research Centers, and five SDG-focused Impact Areas. The three Action Areas and the five SDG-focused Impact Areas are presented as follows (Figures 1 and 2):

Figure 1. One CGIAR Action Areas.

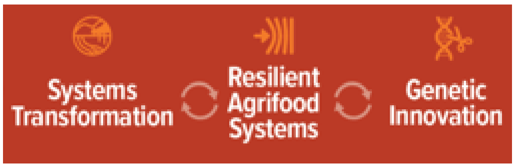
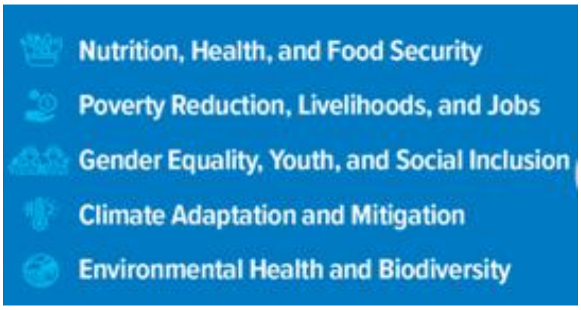


Figure 2. One CGIAR SDG-focused Impact Areas.


Nutrition, Health, and Food Security
Poverty Reduction, Livelihoods, and Jobs
Gender Equality, Youth, and Social Inclusion
Climate Adaptation and Mitigation
Environmental Health and Biodiversity

This technical note is designed to inform the use of bibliometric analysis to evaluate Quality of Science (QoS) in the context of One CGIAR by combining bibliometric expertise with input from CGIAR stakeholders and lessons from the independent evaluative reviews of the CRPs. In 2017, the CGIAR Independent Science and Partnership Council (ISPC)[1] introduced the Quality of Research for Development (QoR4D) framework for system-wide agreement on the nature and assessment of scientific quality. Subsequently, the Independent Science for Development Council (ISDC) supported CGIAR by carrying this work forward. The ISDC published a technical note and companion brief in 2020 as well as an additional companion brief on operationalizing the framework to assess One CGIAR research initiative proposals in 2021. With the QoR4D framework and other ways of assessing agricultural research for development (AR4D) in mind, Science-Metrix, integrated with Elsevier's Research Analytics and Data Services (RADS) team since 2018, developed this technical note to provide recommendations on the use of bibliometrics in mixed-methods evaluations and potential monitoring of QoS in One CGIAR research. This technical note will contribute to the development of guidelines on how to evaluate QoS as part of the One CGIAR evaluation framework to inform future programmatic or project evaluations and to meet the needs of the CGIAR's Performance and Results Management System (PRMS). The target audience of the technical note is the CGIAR System Council and System Board.

[1] The Independent Science and Partnership Council (ISPC), which was reconstituted in 2019 as the ISDC, developed the Quality of Research for Development (Qo4RD) framework in 2017; it was updated in 2021.

1.1 Scope of the Study

The CGIAR Advisory Services (CAS) comprise the Independent Science for Development Council (ISDC), the Standing Panel on Impact Assessment (SPIA), and an independent evaluation function. CAS provides external, impartial, and expert advice related to strategy and positioning, program evaluation, and impact assessment. The CAS Secretariat facilitates and supports these independent advisory services by delivering operational support to ISDC and SPIA and executing the System's multiyear evaluation workplan.

This technical note will support CAS and help CGIAR build its capacity for bibliometric analysis, critical analysis, and interpretation of bibliometric findings, including by expanding its in-house bibliometric production. In this note, Science-Metrix shares (1) some of the best practices in bibliometrics and information sciences broadly speaking; (2) some of the best practices of private bibliometric service providers; and, most crucially, (3) more than 20 years of experience operating a complex bibliometric production and analysis pipeline.

The ultimate goals of this process are to

- broaden the panel of indicators available to CGIAR, covering a higher number of dimensions from the QoS/QoR4D frameworks and interests from various stakeholder groups;
- enable CGIAR to increase the robustness of its bibliometric analyses by implementing best practices when designing analytical groups, analytical periods, and comparative strategies;
- support internal capacity building for CGIAR's bibliometric pipeline; and
- on a best effort-basis (given this falls out of its main area of expertise), support the definition of One CGIAR's PRMS guidelines, including by contributing to the assessment of qualitative indicators.

It must be noted that Science-Metrix cannot conduct a systematic review of bibliometric evaluation frameworks already deployed in various contexts, institutions, or initiatives, including those frameworks used in national research systems assessments (Hicks, 2012; Zacharewicz et al., 2019). Furthermore, such an exercise would not have been useful for the current study, given that the bibliometric practices used in many formal Northern evaluations focus on publication count, journal impact, and citation impact only, with most considerations of relevance and legitimacy left to peer review or expert judgment methods (Zacharewicz et al., 2019). Such bibliometric frameworks are too narrow to fully account for the expectations put on research and innovation in an AR4D context (Noyons & Ràfols, 2018; Tijssen & Kraemer-Mbula, 2018).

Science-Metrix is best positioned to transfer knowledge on analytical strategies and indicators for which it has accumulated direct experience. Given the operational challenges faced in the development of any bibliometrics pipeline, it appears there is a strong rationale for balancing both feasibility and authority in selecting indicators. The added value of consulting with Science-Metrix lies precisely in this concern for feasibility, which would remain secondary in any framework elaboration that would rely solely on the published scientific literature. Therefore, one pragmatic goal in the conduct of this study has been to try to avoid the production of an authoritative but abstract framework based solely on the literature and that does not account for the specific AR4D context within which CGIAR operates.

The current study provides a high-level framework for designing and conducting or provisioning bibliometric analyses by CAS-engaged analysts; the Performance Portfolio Unit (PPU); and the monitoring, evaluation, and learning (MEL) community in One CGIAR. Science-Metrix cannot assess in advance all potential obstacles or challenges encountered in the implementation of bibliometric analyses. In view of potential challenges such as changing data sources, unexpected metadata errors, or other obstacles that are identified and defined only when implementation of a project is launched, Science-

Metrix advises analysts to undertake frequent reassessments of the overall framework by triangulating with the empirical experience collected during implementation.

Science-Metrix itself, in the conduct of its commercial bibliometric activities, can seldom rely on generic indicator and assessment frameworks. Instead, it must invest substantial time and resources in building tailored evaluation designs for most of its clients and/or projects. Bibliometric indicators can seldom be rolled out in a generic manner to all experimental, institutional, and geographical contexts. On the contrary, robust delineation of publication sets in given thematic areas, for given comparators, or in given time periods often requires extensive redesign, even when indicator formulas and specifications themselves tend to remain constant from one project to another. Based on its own accumulated experience, Science-Metrix advises CGIAR to set time aside to adjust and customize the overall framework presented here before launching more specific bibliometric assessments of specific Action Areas or initiatives, or especially in comparative exercises. These reassessments should aim to evaluate the composition of the panel of indicators most relevant to the task at hand in view of constraints in data availability or costs.

1.2 Approach to the Development of the Technical Note and Recommendations

The development of this technical note followed modified terms of reference written by the CAS and based on the proposal by Science-Metrix. From October 2021 through January 2022, a consultative and co-design process was implemented to develop this technical note. Recommendations on the use of indicators for evaluating QoS were to be drawn from the use of bibliometric indicators in the evaluation of QoS in the 2020 CRP evaluative reviews, the QoR4D framework, consultations with the CAS Secretariat Evaluation Function, and a focus group, interviews, and surveys[2] with stakeholders and subject matter experts (referred to herein as "key informants"). The resulting draft document was then reviewed by Guy Poppy and Zenda Ofir of the CAS Evaluation Reference Group; as well as Enrico Bonaiuti (ICARDA Research Team Leader – MEL); and CAS consultants Jillian Lenne, Paolo Sarfatti, and Stefania Sellitti.

2 Frame of Reference

2.1 QoR4D Criteria as a Framework Guiding Dimensions of Evaluation

Consistent with the 2012 CGIAR evaluation policy and QoR4D framework, QoS refers to ways in which research is designed, conducted, documented, and managed. The 2020 revised QoR4D consists of four key elements—relevance, scientific credibility, legitimacy, and effectiveness—defined by the ISDC as included below (Table 1, Figure 3). The 2022 CGIAR Evaluation policy (CAS Secretariat, 2022) links six CGIAR evaluation criteria and QoR4D elements: Evaluation criteria reflect the characteristics of research for development in the CGIAR context, consistent with the QoR4D framework elements.

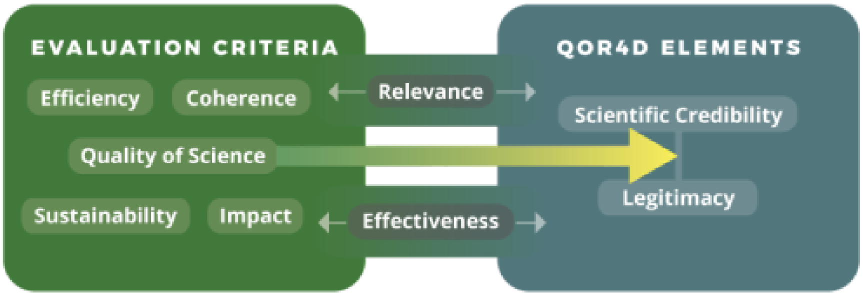
[2] See Annex 3 for the Survey Instrument and annexes 5 and 6 for a list of all participants to interviews, focus group discussion and a final validation meeting.

Table 1. Four key elements of Quality of Research for Development (QoR4D), 2021

1. Relevance Relevance refers to the importance, significance, and usefulness of the research objectives, processes, and findings to the problem context and to society and is associated with CGIAR's comparative advantage to address the problems.	**3. Legitimacy** Legitimacy means that the research process is fair and ethical and perceived as such. This feature encompasses the ethical and fair representation of all involved and consideration of the interests and perspectives of intended users. It suggests transparency, sound management of potential conflicts of interest, recognition of the responsibilities that go with public funding, genuine involvement of partners in co-design, and recognition of partners' contributions. Partnerships are built on trust and mutual commitment to delivery of agreed-upon outcomes.
2. Scientific Credibility Scientific credibility requires that research findings be robust and that sources of knowledge be dependable and sound. It includes a clear demonstration that data used are accurate, that the methods used to procure the data are fit for purpose, and that findings are clearly presented and logically interpreted. It recognizes the importance of good scientific practice, such as peer review.	**4. Effectiveness** Effectiveness means that research generates knowledge, products, and services with high potential to address a problem and to contribute to innovations and solutions. It implies that research is 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 to use and to help generate desired outcomes. To achieve target outcomes, research requires a clear path to impact in one or more of the five impact areas, regardless of where it sits along the spectrum from fundamental to applied research.

In line with the revised 2022 Evaluation Policy for CGIAR, and with a view toward developing a comprehensive approach to evaluating QoS, the list of recommended indicators for consideration in this technical note spans all four QoR4D elements.

Figure 3. Mapping of CGIAR evaluation criteria to QOR4D, CGIAR Evaluation Policy 2022



Note: The criterion of "coherence" was not part of 2012 CGIAR evaluation policy.

2.2 Methodological Approach for Evaluating QoS during the 2020 CRP Reviews

In cross-referencing the QoR4D framework's four key elements with the CGIAR evaluation criteria based on OECD-DAC evaluation criteria, which was adopted for the previous round of reviews, the element of legitimacy was covered in QoS, and is dispersed through the sustainability and effectiveness evaluation criteria (Table 2, Figure 4).

For the twelve 2020 CRP reviews commissioned by CAS, the QoS evaluation criterion was operationalized through the two elements of scientific credibility and legitimacy, and three dimensions as follows:

- **Inputs:** The 2020 CRP reviews assessed the extent to which CRPs benefited from sufficient high-quality inputs to deliver planned outputs and outcomes.
- **Management processes:** The reviews assessed how well CRP management processes ensured the QoS, including relevance to next-stage users, scientific credibility, and legitimacy, of the research and operations.
- **Outputs:** Evaluation of scientific credibility addressed research outputs, such as published results and improved varieties, as well as leadership, staff disciplinary expertise, processes, and incentives for achieving and maintaining the high scientific credibility of those outputs. The assessment of scientific credibility also included research teams' track records in terms of state-of-the-art research literature, methods, and novelty. The shift to One CGIAR will continue to address the QoS evaluation criteria using QoR4D elements of scientific credibility and legitimacy, as expressed through relevance and effectiveness.

  A lean approach to the CRP evaluative reviews restricted the time to comprehensively evaluate peer-reviewed publications.

Figure 4. Evaluation Criteria and QoR4D elements adopted in 2020 CRP reviews, 2012 CGIAR Evaluation Policy

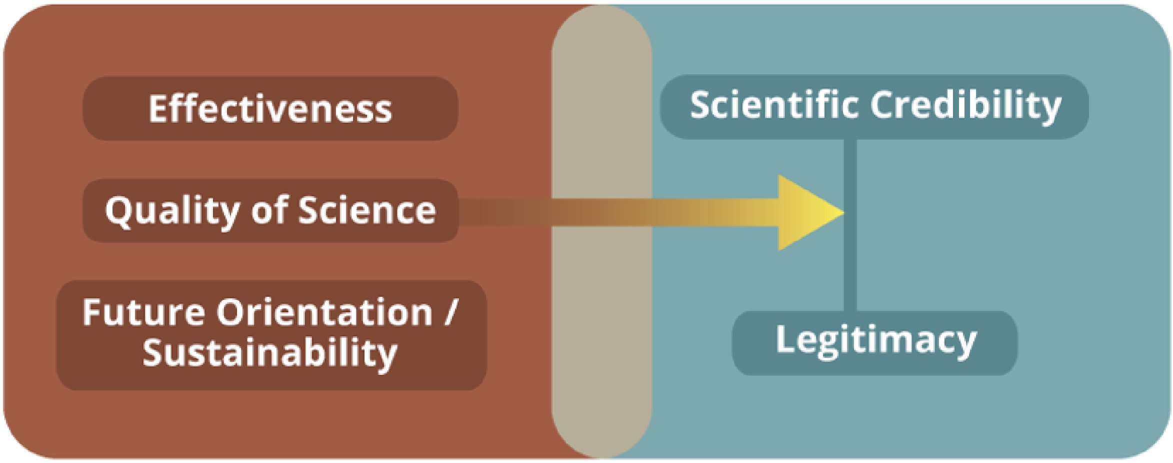


Table 2. Data Collection Matrix for QoS elements in the 2020 CRP Reviews

QoS dimensions	Evaluation question	Elements to be assessed	Assessment criteria	Data sources	Credibility	Legitimacy
INPUTS	To what extent does the CRP benefit from sufficient high-quality inputs, necessary to deliver planned outputs and outcomes?	Composition of research teams	Adequacy of skills and scientific disciplines; Adequacy in relation to diversity of age, gender, and nationality	CRP; **Bibliometrics** (H indexes); Interviews	✓	
		Funding	Adequacy and predictability	CRP reports; Interviews	✓	
		Research infrastructures	Adequacy	CRP reports; Interviews	✓	
PROCESSES	To what extent do the CRP management processes ensure the quality of science, including relevance to next-stage users, scientific credibility, and legitimacy, of the research and operations?	Partnerships	Mutual trust, understanding, and commitment; Clear recognition of partners' perspectives, needs, roles, and contributions; Multistakeholder approach	CRP reports; Interviews, FGD	✓	✓
		Research ethics	Policies in place for research ethics and their implementation	CRP reports; Interviews, FGD		✓
		Internal review mechanisms	Policies in place for internal review mechanisms and their implementation	CRP reports; Interviews, FGD	✓	✓
		Mentoring and training of junior staff	Policies in place for mentoring and training of junior staff and their implementation	CRP reports; Interviews, FGD		✓
OUTPUTS	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; Quality assessment of scientific publications; Quality assessment of technical publications; Altmetrics scores; Quality and relevance to next-stage users	**Altmetrics**; **Bibliometrics**; CRP reports; **Publications**	✓	✓
		Development of physical products, i.e. improved varieties and digital innovations	Broader applicability; Potential for impact at scale	CRP reports; Interviews	✓	✓
		Communication of research findings	Relevance to target audiences	CRP reports; Interviews	✓	✓

3 Alignment with Broader Objectives and Context within CGIAR

This section describes how different audiences and clients within CGIAR may have different requirements and uses for bibliometric analyses in the monitoring and evaluation (M&E) of CGIAR research in One CGIAR, and how the CGIAR context could inform changes to bibliometric analyses to facilitate and contribute to the QoS evaluation of One CGIAR.

3.1 Governing Bodies

Part of the work of the CGIAR System Council (SC), as supported by the Strategic Impact, Monitoring, and Evaluation Committee (SIMEC), is to (1) request independent evaluations; (2) oversee the strategic direction and efficiency of the System Organization; and (3) monitor the efficiency, effectiveness, and impact of CGIAR Research. According to a SIMEC-member key informant, bibliometric indicators are not currently used at the System level for decision-making but would be useful in System-wide monitoring and evaluation. QoS is an important component of the entire CGIAR System and its functioning and being able to demonstrate CGIAR's research capabilities in terms of QoS is of intrinsic importance for the System. However, as another key informant who was part of the development of QoR4D noted, SC and SIMEC lack a common frame to evaluate activities across the board,[3] and there is a challenge in applying the same frame. System-level bibliometric analyses will need to consider how to weigh the relative importance of scientific credibility and legitimacy criteria within the evaluation of CGIAR research. Looking forward toward One CGIAR, with increasing pooled funding of initiatives, finding ways throughout the entire System to evaluate and monitor science impacts and quality will be important. In that sense, a formal collaboration and ongoing forum could ideally be used to do real-time bibliometric monitoring of progress (as mentioned by two key informants), although the feasibility of this would depend on several other factors.

3.2 CGIAR Research Centers

CGIAR research has been delivered through its CGIAR Research Centers, which are nonprofit legal entities housing more than 8,000 scientists, researchers, technicians, and staff. From 2012 to 2021, most research portfolios were CRP driven. Under the portfolio and structure prior to 2022, Centers participated in one or more of the 12 CRPs or the four platforms (Excellence in Breeding, Big Data in Agriculture, Gender, and Genebanks). As one key informant noted, the move from Center-driven research to CRPs was transformative for research progress. In 2022, under One CGIAR, CGIAR is adopting an integrated, global operational structure. Enhanced use of bibliometric analyses would allow for a better understanding of how the shift to One CGIAR impacts research in terms of the five SDG-focused CGIAR Impact Areas: (1) Nutrition, health, and food security; (2) Poverty reduction, livelihoods, and jobs; (3) Gender equality, youth, and social inclusion; (4) Climate adaptation and mitigation; and (5) Environmental health and biodiversity.

[3] This technical note, and supporting guidance it will inform towards operationalizing the new CGIAR evaluation framework and review evaluation policy, are all designed to remedy this lack of a common frame.

3.3 Funders and ISDC

The System Council, comprising CGIAR's largest funders, unanimously endorsed the One CGIAR recommendations set out by the System Reference Group (SRG) in November 2019. Those recommendations included the achievement of at least 50% pooled funding by 2022 and at least 70% by 2024. The CGIAR 2030 Research and Innovation Strategy (the CGIAR Strategy) is operationalized through research-for-development programming, supported by a broad range of funders and investors, including investments in a prospectus of Initiative projects. New CGIAR Initiatives to help radically realign food, land, and water systems were announced in 2021. The CGIAR Performance and Results Management Framework 2022–2030 (PRMF) will measure the results of these efforts.

Funders, as key stakeholders, have been involved in developing and prioritizing new Initiatives. The Independent Science for Development Council (ISDC) coordinated independent review of the Initiatives and operationalized the QoR4D framework (Beaudreault & Meinke, 2021). As such, an important use for bibliometric indicators will be for Initiative-by-Initiative monitoring and evaluation in order to understand each prioritized Initiative and the associated quality and effectiveness of AR4D outcomes.

3.4 Stratification and Breakdown of CGIAR Research

One type of stratification of research is geographic attribution, which is normally done bibliometrically using author affiliation. However, the geographic attribution of CGIAR research, as represented by authors' publication affiliations, has a level of complexity and fluidity that surpasses those of other contexts. To overcome this problem, CGIAR outputs, including but not limited to those elements to be assessed with bibliometric data sources, should be documented with some key metadata as it is done by the MEL professionals and others with related job responsibilities. With good-quality publication-level metadata in place, it is possible to conduct bibliometric assessments for all types of stratification. More details and recommendations are made in the following chapter.

4 Overview of CGIAR Data Architecture and CAS Evaluation Strategies

4.1 Research Co-Creation and Joint Quality Assurance in the CGIAR

The 2021 report *Synthesis of Learning from a Decade of CGIAR Research Programs*, based on existing evaluative evidence, indicates that scientific quality assurance (QA) processes (normally the responsibility of Centers) were applied by CRPs—but applied inconsistently (CAS Secretariat, 2021). While some host Centers had strong science quality processes, these processes were not enforced consistently across CRPs. Additionally, in the current CGIAR research context, Centers' commitment to open access and open data (OA-OD) requires harmonized implementation across CGIAR research to comply with donor policies. However, as noted by one key informant, analysis of OA-OD research practices comes with limitations, including the equity issues that persist for researchers from low- and middle-income countries (or the Global South) that are prioritized within the context of QoR4D.

The focus group, interviews, and surveys conducted with key informants addressed other limitations as well. An important potential use for bibliometric analyses that has not been part of the CGIAR Strategy and Results Framework (SRF) 2016–30 and PRMF 2022–30 is to monitor progress and to evaluate process and performance in order to inform improvement. As one key informant noted, the 2020 CRP reviews were conducted too late for any formative learnings to be implemented in subsequent research plans, and bibliometric analyses could play an important role in monitoring and learning for researchers if implemented early enough to be put into practice in a timely way.

The Monitoring, Evaluation, and Learning Quality Assurance Processor (M-QAP) tool developed by the MEL team at the International Center for Agricultural Research in the Dry Areas (ICARDA) has already demonstrated an improved quality assurance process for tracking research output across CRPs in CGIAR, as noted in its case study (De Col et al., 2021). This tool supports CGIAR quality assurance processes for peer-reviewed publications in an “automatic, reproducible, reliable, and rapid” way. It assesses publication metadata in the Web of Science (WoS) Core Collection and other related metadata linked through the integration of different application programming interfaces (APIs), replacing the previous process of manually checking all publications provided by CRPs and platforms. Through this tool, proxies for examining open science and open access requirements for CGIAR research can be implemented and used in bibliometric analyses.

4.2 Monitoring of Publications and Other Outputs in CGIAR

The 2020 CRP reviews relied on CGIAR entities (CRPs and platforms) to submit peer-reviewed publication information (DOIs) as part of annual reporting, with the responsibility for QA processes falling to each Center. Key informants differed in their reflection on the completeness of provided data. While some 2020 CRP reviewers noted that the data was quite complete during their review (e.g., A4NH, RTB), others responded that there were limitations, including some missing data, inconsistency between institutions, and lack of quality assurance (e.g., WHEAT, FISH, GLDC). These limitations pertained largely to the first cohort of reviews, and many were addressed in subsequent reviews. The 2021 Synthesis Report indicates that during the 2020 CRP reviews, publication-level QA performed by Centers varied from one CRP to the next (CAS Secretariat, 2021a and 2021b).

With the new One CGIAR research portfolio and the framing of research into initiatives instead of CRPs, it is not clear yet how differently publications will be monitored under the One CGIAR structure. As part of the CGIAR Level Agricultural Results Interoperable System Architecture (CLARISA), which transforms raw data on CGIAR research and activities into interoperable information that can be used for evaluation purposes, the M-QAP tool has been deployed to begin semi-automated validation processes for publications. It uses DOIs to query through WoS, with manual validation required for unindexed DOIs. This process may help address some of the limitations identified by key informants.

4.3 Bibliometric Database Access and Curated Data Acquisition Practices in CGIAR

Although not implemented for the 2020 CRP reviews, the M-QAP system pilot has demonstrated improved reporting of CGIAR results that can be used for M&E (ICARDA, 2021). The current M-QAP setup accesses data and metadata from WoS, Scopus, Unpaywall, and Crossref to assess peer-reviewed publications submitted by CGIAR entities as part of annual reporting.

Importantly, data acquisition through M-QAP is followed by exhaustive and in-depth curation of publication records by the MEL community. This curation step allows users to correct metadata on authors’ affiliation and country of origin, identify the gender of authors, and identify the Centers and programs of provenance for publications. Other categorizations, such as crop category and partners integrated in the work, are also applied.

It is important to note that bibliographic databases or commercial providers cannot provide metadata on publication alignment with Action Areas or Initiatives, on crop categories, or on partners. These metadata fields must be coded by CGIAR.

It is also recommended to proactively encourage redundancy in metadata curation for any future retrospective evaluations. This approach will enable CGIAR to efficiently conduct bibliometric assessments at multiple levels of stratification and to produce bibliometric findings by specific operational breakdowns within CGIAR (Portfolio, CGIAR Centers, Action Areas, CRPs, or Initiatives); by donor or funder; and by country or other geographic groupings.

Recommendation 1 (R1): Science-Metrix recommends that data curation currently performed around the M-QAP system pilot system be expanded and systematized throughout CGIAR. Publication records provided by different CGIAR centers should be reviewed to ensure consistency. Consistency in metadata curation in turn supports improved precision and recall values for bibliometric assessments, whether these assessments are conducted in house or subcontracted to commercial providers. In addition to standard publication metadata (DOI, year, title, journal, authors, and so forth), curation efforts should continue to characterize CGIAR publications by:

- Action Area
- Initiative
- CGIAR Center
- Funding source
- Crop category
- Partner participation
- Any linkage of the publication or the underlying research to an Outcome and Impact Case Report (OICR)

Again, it is impossible to conduct bibliometric assessments for certain stratifications if the associated categorization has not been implemented, or has been only partially implemented, at the level of individual publication metadata.

R2: CGIAR should provide requirements and a formal template to be used in the funding acknowledgment of CGIAR publications. The funding acknowledgment template should mandate mention of funding sources, CGIAR Center, Initiative and Action Area.

It is sometimes suggested that correct use of acknowledgment templates be required for a given CGIAR publication to be part of an evaluation (presumably at the Center or researcher level). In terms of bibliometrics program evaluations (which are likely to provide input into strategic decision-making rather than Center- or researcher-level decision-making), this strategy to increase compliance presents drawbacks, since it might lead to fewer observations. Affiliation and funding acknowledgment fields from commercial bibliographic databases are typically used to complement and complete manual curation and in turn increase the number of available observations for the statistical analysis. In fact, it might be worthwhile to investigate these fields to validate that CGIAR researchers themselves do not forget publications in compiling their lists of project outputs, a common situation in Science-Metrix's experience.

4.4 Brief Case Studies of QoS Evaluation Approaches Deployed by CAS in the 2020 CRP Reviews

Quality of Science in One CGIAR may look different within different research initiatives, as it did in the CRPs. The following illustrative case studies from four of the 2020 CRP reviews provide a brief overview of the evaluation approaches deployed by CAS, to answer the evaluative review question "To what extent does the CRP deliver QoS, based on its work from 2017 through 2019?" They highlight variations in approaches, priorities, and limitations for evaluating QoS in each CRP. All CAS-commissioned CRP 2020 reviews, from which the following four case studies are drawn, adopted bibliometrics and altmetrics as part of a mixed-methods approach, using them as proxies to measure and evaluate QoS in addition to other data sources. All 2020 CRP reviews also faced time and budget limitations that led to the adoption of a lean approach in review guidelines, which used only two core evaluation criteria (see Figure 3) (CAS Secretariat, 2012).

4.4.1 CGIAR Research Program on Wheat (WHEAT)

The WHEAT CRP was reviewed in the first of four cohorts of 2020 CRP reviews, together with A4NH and GLDC (CAS Secretariat, 2020f, a, b).

Approach. The WHEAT review team used bibliometric indicators complemented with surveys and interviews. Data sources included documents and records provided by CAS and WHEAT; surveys and interviews were conducted with managers, scientists, and other stakeholders. A subsample of publications was read and analyzed in detail to complement the quantitative analysis.

Priorities. The WHEAT review team prioritized network development, which may improve delivery of outputs relative to investment, with an emphasis on the implications of diversity and resilience in these networks. For bibliometric output indicators, journal rank for publications (divided into quartiles) was prioritized. A WHEAT CRP reviewer, who was a key informant for the current study, noted that working with journal impact factor quartiles was useful for taking disciplinary norms into account. They also noted the value of reading publications for understanding QoS, even though it required a diversity of experts to create inter-rater reliability for assessing each publication.

Limitations. According to the WHEAT CRP reviewer, databases were inconsistent between Centers and over time for a given Center, and databases included duplications. The key informant revealed that there had been concerns about the quality of publication data, which led the team to create their own system for evaluating WHEAT bibliometric indicators.[4] Additionally, network maps of coauthorship were of limited use without comparators. It was noted that it would be of value to have bibliometric indicators of gender rather than just relying on qualitative data; that comparison of CGIAR performance with other similar institutes such as Cornell would be useful; and that altmetrics have questionable value. The key informant also noted the importance of the use of responsible metrics; that is, reviewers need to understand the assumptions underlying each indicator and how they should be interpreted, and indicators must be complemented with other complementary data sources.

4.4.2 CGIAR Research Program on Livestock (LIVESTOCK)

Approach. Bibliometric data were used to assess individual articles, journal quality and access, h-indices of researchers, and altmetrics for LIVESTOCK (CAS Secretariat, 2020c). A sample of publications was selected for individual assessment according to the following criteria: methodological rigor, novelty/originality, international public good (IPG) value, quality of publication (impact factor), coauthorship, and overall quality. Thirty-five physical products and 10 communication products of various types (blogs, posters, newsletters, and flyers) were also assessed. The review team conducted 48 individual interviews and two focus group discussions (FGDs)—one with the five members of the Independent Steering Committee (ISC) and one with a group of eight junior researchers—bringing the

[4] This experience also spurred the CAS evaluation team of staff and consultants to devise and implement much deeper QA and pre-analysis of the source data.

total number of people interviewed to 61. Two Outcome and Impact Case Reports (OICRs) were selected for deep dives.

Priorities. In the 2020 CRP review of LIVESTOCK, emphasis was placed on inputs, such as staff and partner diversity and funding. Bibliometric priorities included the quality of research outputs. Other output priorities included physical products. Notably, 2020 LIVESTOCK CRP reviewers, who were key informants for the current study, highlighted the value of the OICRs in balancing qualitative and quantitative indicators and enabling longitudinal analysis of research, as many research endeavors trace back decades.

Limitations. The key informant to this technical note, who was a 2020 LIVESTOCK CRP reviewer, indicated that the quality of the data, including bibliometric data, was not a limitation in assessing QoS. However, the indicators used to assess QoS outputs inhibited assessment of CGIAR authors' contributions to the research, particularly those of early-career researchers, who may have been low in the order of authorship on papers. The key informant noted the distortion in comparing the same metrics across differing disciplinary orientations, a limitation that needs to be addressed in terms of book and book chapter contributions to scientific output. This limitation was mitigated somewhat by analyzing how much research was published using journal impact factor quartiles, which takes disciplinary field or subfield into account. They also noted that an emphasis on open access and open data, which may be more accessible for publishing researchers and research entities in the Global North, might pose some equity issues. Additionally, there is a need to evaluate research published in peer-reviewed journals above and beyond the impact factor and to examine whether placing a paper in a more applied journal is more aligned with research objectives that include knowledge uptake by intended audiences. It was noted that further emphasis on those QoS indicators that are assessed through OICRs could strengthen research programs.[5]

4.4.3 CGIAR Research Program on Policy, Institutions, and Markets (PIM)

Approach. Unlike LIVESTOCK and WHEAT, which are agri-food systems programs, PIM is a crosscutting global integrating program that implements diverse AR4D projects that aim to contribute to policy, institutional, and market reform in line with global development objectives (CAS Secretariat, 2020d). For PIM the review team used a mixed-methods approach. Bibliometric and altmetrics data were provided by the CAS Secretariat and triangulated with PIM's governance and management documentation and a sample of OICRs. Findings are triangulated through semi-structured interviews: 23 individual interviews were taken; 11 additional people were reached through three group interviews.

Priorities. Given the crosscutting nature and breadth of the PIM CRP, the 2020 CRP review prioritized bibliometric indicators for evaluating QoS. Priorities for measuring QoS included the use of publication output, H-indexed authors, Altmetric attention scores, country collaboration networks, and an added emphasis on lists of top 25 publications for each bibliometric indicator. Additional priorities included examining PIM's output in terms of policies and innovations. Within the context of PIM priorities, these international public goods, including policy studies, analytics, and ensuing policy recommendations or strategic suggestions were directed toward decision-makers. This led to engagement of PIM's next-stage users (government ministries and agencies, nongovernmental and private sector organizations, and multilateral agencies) in partnerships and was standard practice in PIM. The diversity of PIM's research team in terms of social science disciplines and gender was also an identified priority, given the emphasis on integration of different disciplinary and institutional perspectives. CRP reviewers found the Marlo database particularly informative, as it allowed them to organize PIM research outputs by category, although the data had to be checked for completeness. Among qualitative indicators, reviewers were interested in the way in which researchers set out their publication priorities, as it gave a good indication of how the program guides performance. Furthermore, the indicators capturing international collaboration through the location of researchers were considered relevant. In contrast, the subject matter expert (SME) viewed the assessment of scientific quality through a qualitative review of a sample of publications as an activity that took a lot of time for little value, and this time could have been used better in evaluating non-peer reviewed materials.

Limitations. The relevance of research processes and outputs to the context and to society in line with declared development objectives was not part of the 2020 CRP reviews. Reviewers noted that leaving out the relevance criterion from the integrated frame of reference proposed by the ISPC posed severe limitations on the assessment of the quality of scientific research and policy incidence for development. All future evaluations of CRPs should assess relevance, either according to the QoR4D definition or the OECD-DAC definition. Additionally, reviewers found databases incomplete, and it was difficult to obtain a clear and clean view of the capacity input (for total FTEs mapped to PIM by flagship program, see Table 2 of the PIM review report) into research during the review. This, in turn, made it difficult to assess productivity and efficiency in terms of QoS. They noted that it needed to be made clear to SME reviewers that data needed to be provided in the context of responsible use of such metrics, with information on how data was transformed prior to use or how SMEs should be transforming data for evaluation of QoS. Additionally, the PIM CRP reviewer indicated that citations were not well suited for QoS assessment in the context of PIM, given the priority of uptake into policy and policymakers' tendency not to cite research. They also noted the importance of using quantitative and qualitative indictors of QoS in self-monitoring exercises, coupled with regular self-assessments and reporting by program staff, so that any lessons on research practices that need to be learned can be learned and addressed early.

4.4.4 CGIAR Research Program on Roots, Tubers and Bananas (RTB)

Approach[6]: Bibliometric data were used to assess scientific credibility (journal impact factors and quartiles, number of citations, altmetrics, and h-index); legitimacy (acknowledgment of coauthors); and relevance (international public goods [IPG] rating) of 371 publications from the RTB CRP (CAS Secretariat, 2020e). Fifty-seven percent of publications were in good to excellent journals with impact factors (IFs) greater than 3, while more than 93% of publications were multiauthor and multi-institute, with a high author collaboration index of 5.28, demonstrating excellent legitimacy. An analysis of 271 publications from the 12 most productive countries found that 89% were multi-country publications, again supporting high levels of international collaboration. Twenty-seven publications were assessed in depth, with the overall quality being good to high. Eleven technical publications, including working papers, project reports, and manuals, were analyzed for quality and relevance to next-stage users as well as for their potential for capacity development. Ten communication products, including mobile phone apps, web portals, web-based tools, newsletters, leaflets/brochures, design, and databases, were analyzed for relevance to target audiences. Sixty-three physical outputs were rated for potential IPGs. The review team conducted 40 individual interviews, two focus group discussions, and two OICRS for deep dives.

Priorities: In the RTB review, emphasis was placed on comprehensive bibliometric analysis and disciplinary skill-based diversity, especially the integration of gender and biophysical research and

[6] The same subject matter expert conducted the reviews of the GLDC and RTB CRPs, applying lessons learned from GLDC review, which was in the first cohort, to the RTB review.

partnerships to complement internal skills through high-quality external partnerships with appropriate NARSs, NGOs, and private sector partners. The RTB key informant highlighted the value of the cross-cutting activities and mechanisms for building cross-crop and cross-center collaboration to increase efficiencies and effectiveness.

Limitations: Because the review was desk based, it was not possible to carry out an assessment of infrastructure and technical outputs or to interview final beneficiaries. However, because most of the CRP's outputs were disseminated by national authorities and other stakeholders, feedback from those interviewees was used to judge the value of the CRP to the final beneficiaries.

4.5 Rationale for a Quantitative, Bibliometric Contribution to a Mixed-Methods Approach

Bibliometric methods constitute a powerful set of tools to assess the scientific performance of various entities—countries, regions, institutions, or researchers—by characterizing various 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, 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 data is used not only to better manage scientific policies, but increasingly also to measure progress toward the achievement of various social and economic objectives, though with more limitations than for excellence metrics (Noyons & Ràfols, 2018; Tahamtan & Bornmann, 2020; Technopolis Group & Science-Metrix, 2020; Vignola-Gagné et al., 2021; Wilsdon et al., 2017).

In the context of CGIAR research and the evaluation of QoS, performing comparative and/or time-series analyses on program data allows analysts to, for example, situate program research in comparison with other research entities that conduct AR4D and demonstrate the success of a funding program's selection process in choosing the best research proposals; determine whether or not the selected individuals are outstanding scientists (internationally or more locally); track important changes—e.g., in terms of production size, scientific impact and/or collaboration intensity—resulting from the funding; and establish whether or not the selected individuals reinforce the scientific strength of their host organizations or departments. These questions are highly relevant to the evaluation of the majority of funding programs as they generally put a strong emphasis on promoting scientific excellence, and CGIAR is no different in this regard. For instance, the bibliometric analyses recommended in this technical note provide valuable suggestions for addressing the points detailed above.

Nevertheless, all bibliometric indicators have some weaknesses when considered individually. Consequently, Science-Metrix promotes the use of various lines of evidence to triangulate the results. As noted by a key informant with expertise in applying the QoR4D framework, the purpose of the research—which in this case, is for development—helps define the weight of different dimensions in research assessment and should take into consideration the context(s) in which research is conducted and the way it is managed. This allows evaluators to deliver robust conclusions upon which to build constructive recommendations toward improving the performance of scientific programs. Indeed, when all the indicators point in the same direction, the results are regarded as being more reliable than those based on a single indicator.

4.6 Qualitative Contribution to a Mixed-Methods Approach

In addition to bibliometrics, comprehensive evaluation of QoS requires balancing consideration of qualitative criteria, including design and methodology, with criteria related to inputs, execution, and likelihood of use. The 2020 CRP reviews used a mixed-methods approach to assess qualitative and quantitative aspects of the QoS (Table 3). Emphasis was placed on the *credibility* and *legitimacy* elements for qualitative criteria across *inputs, processes, and outputs*. During development of this technical note, subsequent analysis, discussion, and feedback from KIIs, who in general had insufficient understanding of the value of qualitative indicators, pointed to the need and potential to expand the qualitative criteria to include the elements of *relevance* and *effectiveness*, distinct evaluation criteria, aligned to QoR4D elements.

Table 3 provides an expanded list of qualitative criteria and indicators (including those from the 2020 CRP reviews) with descriptions and QoR4D elements. This was informed by the International Development Research Centre's RQ+ Assessment Instrument (IDRC, 2022).

Table 3. Qualitative criteria, indicators, descriptions, and QoR4D elements

QoS Category	Criterion	Indicator	Description	QoR4D element
Rigor	Research topic & plan	Global/regional problem	Appropriate, realistic	Relevance
	Design	Coherence, clarity	Appropriate	Credibility, effectiveness
	Methodology	Integrity, fitness	Rigor, clarity	Effectiveness
Inputs	Skill base	Discipline[1]	Appropriate	Credibility
	Composition of teams	Diversity, gender, discipline[1]	Appropriate, inclusive, multi- & trans-disciplinarity	Credibility, legitimacy
	Support structures	Laboratories, fields[1]	Adequate	Credibility, effectiveness
	Funding	Donor commitment[1]	Adequate	Credibility, effectiveness
	Capacity building	Useful to planned activities	Appropriate, adequate	Relevance, legitimacy, effectiveness
Processes	Partnerships	Inclusiveness, recognition[1]	Equal team member, involvement in co-design and delivery	Legitimacy
	Gender	Awareness, responsiveness[1]	Gender integrated in design & implementation	Legitimacy
	Roles and responsibilities	Clarity[1]	Defined roles & responsibilities	Legitimacy
	Performance evaluation	Incentives[1]	Rewards for quality	Legitimacy

QoS Category	Criterion	Indicator	Description	QoR4D element
	Negative consequences	Consequences, risks	Risk assessment and mitigation strategy	Legitimacy, credibility
Outputs	Communication	Methods & tools[1]	Relevance for target audience	Effectiveness
	Enabling environment	Awareness, understanding	Appropriate positioning and targeting	Effectiveness
	Networking	Multi-stakeholder engagement[1]	Adequate and inclusive	Legitimacy, Effectiveness
	Policy linkages	Policy makers engagement[1]	Appropriate and targeted	Legitimacy, Effectiveness
	Scaling readiness	Multi-stakeholder engagement	Contribution to development outcomes	Effectiveness
	IPG generation	Positioning for uptake and impact[1]	Broadness of applicability	Credibility, Effectiveness

[1] These indicators were used in the 2020 CRP reviews.

Interviews with key informants highlighted the importance of including project design, especially for complex projects (such as multi-crop projects) and improved qualitative ability to evaluate gender issues and capacity building. The importance of methodological integrity was also emphasized. Finally, the need for output indicators that captured positioning for use prior to the actual generation of an output was noted.

R3: Expanding the list of qualitative indicators in Table 3 beyond those used in the 2020 CRP reviews will require pilot testing prior to their use in future reviews and evaluations in One CGIAR.

The 2020 CRP reviews combined qualitative and quantitative indicators in a mixed-methods approach to evaluating QoS resulting in a credible, balanced, and comprehensive outcome. Although publications are best evaluated using quantitative indicators such as bibliometrics, many inputs, processes, and outputs are best evaluated using qualitative indicators (Table 3). Other criteria such as capacity building and communication can be subject to both quantitative (e.g., numbers trained, numbers of methods/tools) and qualitative (e.g., useful to planned activities, relevance to target audience) assessment.

R4: Science-Metrix recommends continuing and even expanding the mixed-methods approach used in CRP reviews as part of One CGIAR evaluations of QoS. Science-Metrix considers that the current qualitative criteria and OICR narratives are strong components of evaluation practices at CGIAR. OICRs are crucial for capturing mid-term and longer-term societal outcomes of CGIAR research, which may well be missed otherwise. Further bibliometric developments should be made with an aim to complement and build on this qualitative capacity, and should in no way displace peer reviews, existing qualitative indicators, or OICRs collection.

5 Approach to Bibliometric Analysis: Analysis and Recommendations

5.1 Rationale for an Extended Mixed-Methods Approach, Including an Expanded Contribution of Bibliometrics

Responses by key informants indicate that current bibliometric approaches in the evaluation of QoS in CGIAR research have both strengths and limitations, suggesting areas in which the evaluations could be improved. Significant improvements could be made first by prioritizing indicators based on One CGIAR's development of focused definition(s) of QoS/QoR (which may differ across research endeavors) and bibliometric expertise on the responsible use of some indicators. In prior evaluations of QoS in CGIAR, selecting bibliometric indicators without the benefit of this focus and an understanding of the appropriateness of indicators for each context may have led to the amplification of existing shortcomings and flaws in metrics used in CRP reviews. Science-Metrix considers that greater engagement with responsible use of research metrics is necessary in CGIAR evaluation processes if bibliometric methods are to be used at all.

Peer review and other qualitative evidence streams remain crucial today for assessment of societal outcomes of research (Belcher & Hughes, 2020; Koier & Horlings, 2015; McLean & Sen, 2019; Tahamtan & Bornmann, 2020; Traag & Waltman, 2019). The Science-Metrix team has found (based on prior experience working with research programs oriented toward societal outcomes) the quality of work performed by CAS and the Centers in characterizing non-article publications and societal outcomes of CGIAR research, most notably through the OICRs, to be exceptionally high. This work is crucial to the mandate of CAS, and an extended use of bibliometrics for One CGIAR assessments can complement and build upon those achievements rather than displace them. Additionally, bibliometrics, when used correctly, can be fueled and improved by qualitative input, while simultaneously fostering interrogations and lines of inquiries for follow-up qualitative work.

To give just two examples of proposed approaches for deeper integration of bibliometrics and qualitative outcomes harvesting work from the list of recommended novel and improved indicators (see Section 7), bibliometrics case studies and bibliometric studies of CGIAR partner outcomes could be conducted. The first would focus on completing the OICR outcomes narratives by characterizing the research teams or partnerships that have driven the outcomes of interest. Bibliometric assessments could highlight specific CGIAR practices of cross-disciplinarity or South-North collaborations that have been particularly conducive to certain classes of outcomes. Secondly, bibliometric investigations could aim to characterize how CGIAR collaborations with a particular university or institution in the Global South has led to unique achievements for that partner institution (comparing CGIAR-collaborative publications from that institution with its average publication profile outside the collaboration). The latter approach might be particularly useful for evaluating partnerships as part of future evaluations of CGIAR initiatives. Before a more final design proposal can be made on those two approaches, a initial assessment of data availability would need to be conducted, notably to determine the volume and yearly and subfield distribution of publications available in OICR-related aggregates or from collaborations with close partners.

5.2 Organization of Bibliometric Expertise and Production within CGIAR

Deploying an improved mixed-methods approach requires not only new bibliometric skills and competencies but also a slightly enhanced organization of the evaluation of QoS work itself. In prior bibliometric exercises, CAS-engaged analysts have sometimes provided SMEs conducting QoS reviews with disaggregated paper-level bibliometrics scores. Going forward, Science-Metrix recommends that indicator computations for various aggregates be centralized by the CGIAR system itself, notably the PPU[7] with the new CGIAR digital system, PRMS (see Section 7). That is, SMEs will coordinate with CAS/Evaluation-engaged analysts to define the aggregates and analysis needed for a specific review, but the PPU, the MEL professionals, and/or CAS/Evaluation-engaged analysts rather than SMEs should conduct the computations necessary to arrive at the final findings. This will ensure uniformity of methods, although CAS/Evaluation should also share the methods used to ensure that the interpretation of bibliometric findings is as of high quality as the production steps. **In no way is Science-Metrix suggesting that metrics replace peer review by SMEs. Rather, Science-Metrix suggests that metrics computation be conducted within CGIAR rather than by those SME consultants engaged in external evaluations of programs and projects. SMEs would still be ultimately responsible for analysis and interpretation of these findings in addition to their other review tasks. SMEs should also be involved upstream when defining the panel of indicators to be deployed as part of a specific evaluation (or set of evaluations) of QoS.**

To improve on the mixed-methods approach of evaluating QoS, evaluations should better integrate subject matter expertise with support from experts in bibliometrics and other quantitative and qualitative research evaluation methods, whether they are fully conducted by CAS/Evaluation or with external experts. Monitoring and evaluation (M&E) experts with a specialty in the field of quantitative assessment of research outputs and outcomes can provide additional support to SMEs in using indicators in ways that are both theoretically and statistically valid and appropriate, while considering the context and priorities of each research endeavor. While bibliometrics offer benefits in increasing the transparency of evaluating QoS/QoR and allowing other disciplinary experts to contribute their knowledge of AR4D context and practices, they should be done in consultation with those who can contribute to construct validity. This is because bibliometric phenomena have been shown to follow non-parametric distributions and are often ill fitted for analysis using some of the methods from established statistical traditions that can be heuristic methods for many researchers (Schneider, 2015). SMEs can more readily identify which evaluation questions are of most interest in each QoS evaluation project, but M&E experts should oversee the formulation and computation of the indicators that best answer the questions, although ideally both sets of experts should work closely together throughout this process.

More broadly, Science-Metrix stresses the existence of a specialized field of research on bibliometrics and quantitative assessments of research whose findings and analytical proposals can be found in journals such as *Quantitative Science Studies*, *Scientometrics, Journal of the Association for Information Science and Technology, and Journal of Informetrics* (to name just a few). Others that focus less strictly on bibliometrics include *Research Policy, Research Evaluation,* and *Science and Public Policy.* Specialized bibliometrics researchers take great pains to investigate and characterize the shortcomings of quantitative research indicators and strategies. For instance, and despite the fact that the indicator is widely used in academic governance, the h-index is almost never used by so-called "dedicated

[7] Working title for Project Performance Unit in CGIAR, as of January 2022.

bibliometrics organizations" such as Science-Metrix, the Dutch Centre for Science and Technology Studies (CWTS), or the Norwegian Nordisk institutt for studier av innovasjon, forskning og utdanning (NIFU) (Jappe, 2020). Jappe reviewed 138 bibliometric evaluations of research organizations and funding instruments produced by dedicated bibliometrics organizations between 2005 and 2019 and found that only 4% made use of the h-index. Lutz Bornmann, arguably the bibliometric authority on the h-index, considers that the indicator can be safely used only after several of its drawbacks have been controlled for, including "dependency on field-specific publication and citation cultures… [that] only researchers with a similar academic age should be compared… [and that] researchers should have been active in similar periods" (Bornmann et al., 2022). Engagement with the specialized bibliometrics literature would help identify such limitations and, where possible, improved specifications or mitigation strategies.

SMEs who are unable due to time constraints to engage with this specialized literature should defer to the expertise of the PPU or CAS/Evaluation-engaged analysts and work actively with them in reaching their conclusions as part of CGIAR evaluations and assessments. Where bibliometrics indicators are being deployed in CGIAR M&E related to QoS, CAS/Evaluation-engaged analysts should make sure to deliver findings that have been field- and possibly year-normalized and that use comparable analytical groups. The invited analysts should seek to align their provision of bibliometrics findings with indicator-level best practices found in the specialized literature. The PPU/CAS/Evaluation-engaged analysts should be able to support SME experts with the relevant literature on the h-index, or other similar indicators, when required.

R5: Science-Metrix recommends that the PPU, MEL professionals, and/or CAS/Evaluation-engaged analysts be put in charge of formulating and computing all bibliometrics used as part of One CGIAR evaluation of QoS. SMEs should interact with CAS/Evaluation/PPU to identify the indicators that are most relevant in the context of their specific review. CAS/Evaluation should provide support to SMEs in making robust use of the indicators provided through performance-monitoring systems and beyond and clarify methodological steps conducted that may bear on the interpretation of their results.

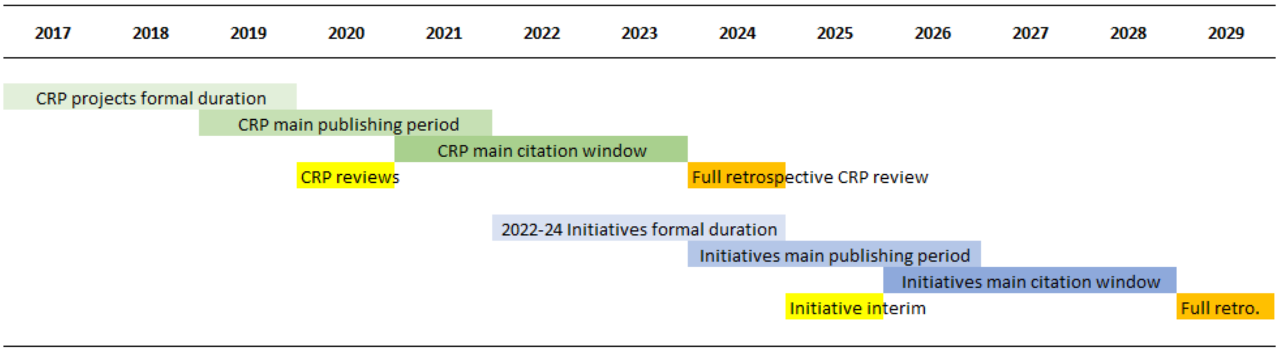
5.3 Shortcomings of Short-Term Evaluation Periods

Science-Metrix has noted that the analytical periods deployed in the 2020 CRP reviews (characterizing in the year 2020 CRP publications issued between 2017 and 2019 or 2017 and 2020) may be problematic from the standpoints of both quantitative and qualitative assessment (see Annex 6 in CAS Secretariat, Synthesis 2021b).

Some outputs and most societal outcomes of research have been shown to typically take 5, 10, or even 20 years to be fully realized (Langfeldt & Scordato, 2015). Research articles and other documents produced in a research project are typically published starting two years after program initiation as well as following project conclusion, often up to two years afterwards (Science-Metrix, 2018). Citations take time to accrue, and at a bare minimum a period of two years after a publication's release year can elapse before measuring citation achievements (Aksnes et al., 2019). Consequently, to be reliable and fully retrospective, measurement of a project or program's citation impacts should be conducted at least four years after a project's conclusion (see Figure 5). For highly novel research, even longer analytical periods are required, since novel research can reach higher citation achievements than less novel research but typically does so starting only three or four years or more after publication (Stephan et al., 2017). Therefore, highly novel research, arguably including research with a strong orientation toward societal outcomes such as development agriculture, is most robustly evaluated five years or later following project or program conclusion.

These findings sit in direct tension with the demands of research evaluation exercises, which typically require input on project or program performance even before their conclusion, to create learning that will help adjust the design of follow-up initiatives. The trade-off between evaluation timeliness and evaluation robustness is always difficult to address and optimize, but Science-Metrix attempts to outline one potential solution to this issue in the next section.

Figure 5. Typical timelines for scientific project duration, publishing of related publications, and associated citation impact realization windows, applied to the CRP and 2022–44 Initiatives programs



R6: Science-Metrix strongly recommends that bibliometric components of evaluating QoS be conducted at a minimum three years after the completion of a group of projects or a portfolio (non-pooled or initiative projects). Preliminary QoS can be assessed in the year following program completion, but this should be restricted to qualitative assessments (possibly including some OICRs) and a small selection of bibliometric indicators (only indicators with a "+3" time rating in Table 6 and Table 7). See also the recommendation for a staggered approach in the next section.

5.4 Deploying a Staggered, Two-Pronged Approach to Bibliometric Data Production

The distinct problems of (1) demand for multiple analytical periods (shorter and longer periods having distinct advantages); and (2) demands for both highly curated publication data and more generic, normalizable, and comparable data on the very same set of publications can both potentially be addressed through a two-pronged bibliometric assessment approach.

Science-Metrix recommends that CAS/Evaluation deploy two formats or designs of related evaluation in the future. Short-term-impact reviews (that follow the temporal approach also used in the CRP reviews) could be complemented by more in-depth, long-term reviews. For instance, Science-Metrix recommends that a comprehensive review of the new Initiatives (which replaced the prior CRPs and platforms labels), with a strong bibliometrics component, be conducted in 2025. However, to complement the 2017–19 CRP reviews that have already been conducted, 2025 would also be an appropriate time point to conduct a full retrospective bibliometric evaluation of mid-term outcomes from 2017-19 CGIAR research. This evaluation period will allow the full realization window for higher-risk research and many (but not all) societal outcomes to elapse.

For One CGIAR initiatives that follow three-year funding cycles throughout 2022–30, but also to capture longer-term outcomes from the 2017–21 cycle, interesting evaluation points from the bibliometric perspective include the following:

- 2025: February or March of that year is the earliest time point when (1) optimal citation impact analysis of riskier research and (2) evaluation of societal outcomes with mid-term realization times can be conducted for 2017–19 CRP reviews publications.
- 2026: February or March of that year is the earliest point when robust citation impact analysis for all publications produced as part of the 2017–21 portfolio projects can be conducted.
- 2029: February or March of that year is the earliest point when robust citation impact analysis for all publications produced as part of the 2022–24 portfolio projects can be conducted.
- 2032: February or March of that year is the earliest time point when (1) optimal citation impact analysis of riskier research and (2) evaluation of societal outcomes with mid-term realization times can be conducted for the publications from the 2022–24 funding cycle.
- 2038: February or March of that year is the earliest time point when (1) optimal citation impact analysis of riskier research and (2) evaluation of societal outcomes with mid-term realization times can be conducted for the publications from the full 2022–30 One CGIAR funding cycle.

If the deployment of two distinct series of evaluations using distinct time windows is deemed unfeasible (the first being interim evaluations in the year following the completion of Initiatives, the second, fuller evaluations taking place five years after Initiatives completion), Science-Metrix strongly recommends including publications from predecessor programs when computing findings as part of future program evaluations, to make sure that (1) long-term (and likely substantial) QoS outcomes are adequately captured, instead of focusing solely on short-term outputs; and (2) bibliometric assessment of trends in research achievements can be obtained. As one KII highlighted, OICRs used in the 2017–19 CRP reviews captured outcomes that were realized over that period, but those outcomes built on prior research that might have been conducted in earlier periods. In a similar approach to the one used for OICRs, the 2017–19 CRP reviews might have added bibliometric findings on relevant CRP articles published between 2014 and 2016 (notably with a fuller citation profile analysis), alongside (but not combined with) the bibliometrics already presented based on 2017–19 publications.

A second aspect of the two-pronged approach concerns the tension between the highly curated character of metadata currently maintained for CGIAR publications by MEL/PPU and the potential need to conduct normalization operations and deploy comparative designs by processing metadata of much lower quality on other publication sets.

R7: For comparative analyses using MEL metadata where equivalent metadata on non-CGIAR publications is of much lower quality (geographical aspects of research, multi-national collaborations, gender equity in publications, and others), Science-Metrix recommends that the same analysis be repeated using both datasets and that findings be reported next to one another in a multidimensional panel of indicators. It might well be that the metadata reported from a commercial provider will *not* coincide with the findings reported from the curated MEL metadata within CGIAR. However, the lower-quality commercial metadata will allow comparisons with external institutions and programs, and the shortcomings of the metadata should apply equally across all comparison groups and in principle should not favor one over the other.

Table 4. Mock illustrative table for presenting bibliometric findings from two parallel metadata curation approaches

Analytical group	% of pubs with a woman as corresponding author (CGIAR curated)	% of pubs with a woman as corresponding author (software curated)	Cross-disciplinarity score
CGIAR publications	55%	40%	2.5
Northern university Y	n/a	52%	1.1
Southern university X	n/a	45%	0.7
South-North collaboration program Z	n/a	30%	1.6
AR4D world level	n/a	32%	1.2

R8: For the 2022–24 Initiatives, Science-Metrix recommends that a 2025 interim evaluation using QoS criterion be conducted based on qualitative assessments and a restricted set of bibliometric indicators. Science-Metrix recommends that a comprehensive and targeted evaluation of QoS in the 2022–24 Initiatives be conducted in 2030, allowing for all Initiative publications, as well as mid-term societal outcomes, to materialize (such as uptake of publications in policy-related documents, mid-term societal outcomes captured by OICRs), or for citation impact profiles of transformative research articles to accrue. A comprehensive evaluation of QoS for 2017–24 research should be conducted in 2025 to complement findings from the 2017-2019 reviews with a fuller assessment of mid- or longer-term outcomes of research from those years. Given the reorganization of CGIAR research over that period, the assessment should be conducted at the full CGIAR portfolio level; and possibly at the Center level in addition.

Only by deploying these comprehensive reviews about six years following CRP or Initiative conclusion can CGIAR truly measure the full range of the societal outcomes of its portfolio. However, interim evaluations can help quickly identify suboptimal practices or processes that need to be quickly addressed before they become entrenched and support timely decision-making.

5.5 Dealing with Metadata Errors

Current bibliometric assessments produced at CAS/Evaluation benefit from painstaking curation produced by MEL in collaboration with (some of) the Centers. This practice ensures that the metadata currently produced by MEL on the affiliations found in CGIAR publications, as well as the gender of authors of these publications, reach a high level of quality and robustness. In fact, these metadata are likely of a much higher quality than those found in bibliographic databases and studies. These high-quality metadata cannot, however, be used in comparative exercises since they may unduly favor CGIAR achievements over that of others.

Science-Metrix would argue that a certain margin of error is an inevitable but manageable drawback of all bibliometric exercises, especially those conducted at scale and with a high volume of publications.

If CGIAR stakeholders prefer not to use comparative approaches using a lower quality of metadata than what they have been used to so far, it would be possible to conduct ad hoc investigations that would quantify the margins of error associated with using commercial metadata rather than curated CGIAR metadata. For instance, random samples of publications in the AR4D field could be characterized to determine the error rate in affiliation data or gender data and to compute the margin of error associated with bibliometric scores computed from the related metadata.

Future bibliometric exercises with these metadata could then report on the specific margins of error associated with each score obtained. Science-Metrix has employed this approach in the ERA progress reports when reporting shares of publications with at least one woman author (PPMI and Science-Metrix, 2019).

6 Normalization and Comparative Strategies: Analysis and Recommendations

6.1 Normalization and Controlling for Confounding Factors

Normalization is a cornerstone of current evaluative bibliometrics (Waltman & van Eck, 2018). Normalization allows for assembling groups of publications more robustly by different scientific subfields and from different years. It also helps control for confounding factors or biases including

- secular temporal trends,
- subfield-level variety in scientific practices, and
- other confounding factors, such as language of publications (Archambault et al., 2006).

Table 5 illustrates some of the subfield-level differences in citation patterns that can bias citation impact findings when un-normalized citation counts are used. It shows how in the field of development studies, articles published in 2018 received an average of 7 citations, whereas articles published in 2008 received an average of 18 citations, surely reflecting the much longer citation window available to them since time of publication. A citation impact analysis mixing development studies from different years would be biased in favor of older publications.

Even within a single year, articles from different fields that may be of similar levels of quality and intellectual interest might perform differently, because propensities to cite and publication volumes are field dependent. That is, a 2018 article in *Ecology* with 9 citations can be considered to have achieved the same level of interest for its respective specialized audience as an article with 7 citations in *Agronomy & Agriculture*. Again, a citation impact analysis working across subfields or fields would risk giving undue advantage to those publications in a subfield with shorter experimentation and publication cycles or a tendency to include more publications in their reference lists.

Table 5. Average paper-level raw citation counts in various subfields from the Science-Metrix classification, 2008 and 2018

Science-Metrix subfield	Average citation count for articles from publication year	
	2008	2018
Anthropology	17	3.5
Development studies	18	7
Agronomy and agriculture	20	7

Science-Metrix subfield	Average citation count for articles from publication year	
Plant biology and botany	27	8
Ecology	34	9
Bioinformatics	63	13

The normalization approach used by Science-Metrix is to compute the average score for an indicator in a given year and for a given subfield in the Science-Metrix classification (Archambault et al., 2011; Rivest, Vignola-Gagné, & Archambault, 2021). For citation impact, these are the scores presented in Table 4. A paper belonging to the same combination of year-subfield-document type will then see its score divided by the paper-level average found for that combination. For instance, a 2008 publication assigned to the bioinformatics subfield in the Science-Metrix classification with 21 citations (as of 2021) would have a normalized citation score of 0.33 (21/63). Another 2008 bioinformatics paper with 126 citations as of 2021 would instead have a normalized score of 2.0, or "twice the world level" in 2008 bioinformatics.

Also note that the normalization "carries over" into the aggregate, or, to put it differently, normalization becomes "embedded" in the scores once applied. For example, if you have an aggregate with one development studies paper with a normalized citation score of 1.5; one forestry paper with a normalized citation score of 1.0; and one plant biology paper with a normalized citation score of 0.5, the aggregate of these three publications has a "normalized" score of 1.0. But in fact the normalization is only applied at the paper level and does not need to be applied at the aggregate level once applied at the paper level.

The main challenge for CGIAR related to the implementation of normalization will be to develop a novel data architecture for this purpose. This involves computing cross-cutting average scores on indicators for all publications in a given subfield and year in which CGIAR publications can also be found. This is unlikely to be feasible within the resources of CAS/Evaluation or PPU, but commercial bibliometrics providers should be able to provide paper-level average scores by year, subfield, and document type (the latter variable being less crucial) at modest cost (estimated cost of US$5,000 or less for all subfields and all required years; even less if only a select number of subfields is required to adequately cover CGIAR publications); to then be used as denominators by M&E by CAS/Evaluation or PPU in normalization procedures.

One further complication is that in the year-dependent normalization approach used by Science-Metrix, it is necessary to adjust scores every year, even retrospectively for past years. However, if CAS/Evaluation orders new paper-level averages every year for publications from the previous year, it should be able to obtain corresponding averages from the preceding years as well as part of the modest costing already mentioned.

R9: Science-Metrix strongly recommends that the commissioning CGIAR entity (CAS/Evaluation or PPU) provides normalized versions of indicators of citation impact, and possibly of other indicators, to SMEs or external reviewers evaluating QoS at CGIAR. Given current limits to the large-scale retrieval of publication records as part of the MEL data architecture, Science-Metrix recommends that commercial providers of bibliometric data be engaged to compute and provide the average scores on indicators of interest per subfield, year, and possibly publication type to be used by CAS/Evaluation/PPU in the normalization process.

Alternative approach: It is theoretically possible to normalize bibliometric measurements against any aggregate of publications. If CAS/Evaluation or PPU implements a reference thematic set of "development agriculture" articles as discussed below in section 6.2.2, it would be possible for example to normalize bibliometric measures against paper-level averages computed from that reference set instead of ordering average scores from commercial providers. However, for this strategy to work robustly, CAS/Evaluation/PPU would need to have high certainty that publication and citation practices are highly homogeneous within CGIAR publications—that is, that subfield-specific research. It would need to be certain, for example, that the "economics of development agriculture" research follows the same experimentation and publication cycles and practices as the "genomics of development agriculture" research. This premise appears highly uncertain given the explicit objective of fostering cross-disciplinarity within CGIAR research.

6.2 Comparison Strategies

In the focus group and in additional interviews, Science-Metrix has often noted reticence with respect to the use of comparative designs to measure relative levels of achievement for CGIAR research. There appears to be a fear among CGIAR reviewers and stakeholders that CGIAR components may be pitted one against another in a metrics-driven "race for the prize." One key informant noted the nuance of development agriculture research as separate from non-development agriculture research, and how these key characteristics of CGIAR research would need to be heavily incorporated into the selection of any comparison groups.

Two other key informants, however, have greatly emphasized the need for greater recourse to comparison. One emphasized the difficulty of correctly interpreting the reports of outcomes without a reference level available to contextualize such numbers. Another respondent, a representative for a donor organization, emphasized the demands for transparency and for highlighting value for money that politicians and even taxpayers placed on their organization. The potential risk identified here was that the donor would face difficulties justifying investments in CGIAR without such evidence available.

Science-Metrix holds that comparisons indeed offer great utility when interpreting bibliometric findings but also that comparison need not become an end in of itself. Comparisons are often used with a descriptive aim to highlight unique features of different organizations or social groups, most notably in qualitative social sciences (Jasanoff, 2005). Comparisons need not be purely metrics-driven and can also integrate substantial portions of qualitative insight of the type found in the OICR reports.

With these considerations in mind, Science-Metrix has aimed to design proposals for comparative analyses that could be used and that are more likely to find approval within the greater CGIAR community. These include the following:

- Use of time series
- Comparisons with a thematically appropriate reference set of publications
- Comparisons with partner institutions to tease out the added value of CGIAR collaborations for these partners
- Comparisons between different international co-publication partners to highlight features or achievements of different bilateral relations—for example, Asian-African collaborations or African-American collaborations—to highlight the features or strengths of these collaborations or the differential outcomes of the collaborations for the partners involved, against the rest of their work
- Comparisons with external organizations or institutions with a similar research focus, roughly similar organizational profile, and similar focus on development—for instance, with Northern universities with sizable research portfolios in development agriculture

Each type of comparison will be treated separately in the lines that follow.

R10: Science-Metrix recommends that CAS/Evaluation/PPU increase the use of reference levels and comparative strategies in evaluation of QoS to support more robust interpretation of bibliometric findings. At a minimum, the use of a thematic reference set and comparisons with external organizations should be deployed.

If resources and the number of observations allow (i.e., publication volumes at 30 or above), the other comparison modes presented here could also be deployed as part of ad hoc investigations rather than standard monitoring.

6.2.1 Time Series

Time series, by presenting yearly changes in measurements on a given dimension over time, can succeed in capturing the impact of a support intervention of a given research group, especially when considering specifically the support period under investigation. Time series are particularly useful for measuring progress toward a specific quantitative target—for example, "we want 50% of authorships[8] on our publications to be held by women by 2030"—after a baseline measurement has been made and agreed upon.

However, it should be kept in mind that time series are not able to control for local and global trends in a field or sector that might equally affect all groups or actors. For instance, international co-publication shares[9] have been increasing steadily in almost all countries and areas of science in the past 30 years. An increase of this indicator for a given group must be appraised against those of a comparable organization to determine whether the change observed is causally linked to a hypothetical intervention or whether this increase would be fully explained by the secular trend.

Within the context of One CGIAR, time series are probably less useful for evaluating outcomes from the fast-paced Initiatives, and their use should probably be restricted to measuring progress on overarching targets that can only be achieved over five years or more and by mobilizing efforts across multiple Initiatives, possibly at the Action Area or Portfolio level.

6.2.2 Comparisons with a Thematically Appropriate Reference Set of Publications

A common strategy to assist interpretation of bibliometric findings is to compare measurements against a paper-level average of scores in the full bibliographic database (all articles for a year and in the relevant field, regardless of the institution or country where they were produced). Science-Metrix uses such an approach in most of its reports, in the recent past notably to:

- Retrieve research thematically related to the European Union FP7 Strategic Grand Challenges and Key Enabling Technologies (Science-Metrix and PPMI, 2021)
- Identify research thematically aligned with the United Nations Sustainable Development Goals (Rivest, Kashnitsky, et al., 2021)

[8] This refers to the unit of author contributions within an aggregate of publications. In an aggregate of 10 publications, each with three authors, there is a total of 30 authorships.

[9] Shares of international co-publications within an aggregate of publications are considered a measurement of international collaboration levels. Publications are considered international co-publications when they include at least two different authors affiliated with institutions in two different countries.

- Delineate an AR4D thematic set to track the impact of GIZ-funded research within CGIAR (Pinheiro et al., 2020)
- Delineate comparable publications falling into a combination of subfields including climate science, environmental sciences, ecology, and sustainable development to mirror the mixed portfolio covered by Belmont Forum funding (Technopolis Group & Science-Metrix, 2020)

Of course, in an aggregate where all publications fall within a thematic category used for indicator normalization, assembling a reference set would be strictly redundant with the normalization step described in Section 6.1. In practice, aggregates of publication rarely fall within a single subfield category used for normalization, and so it is often necessary to establish a reference set that cuts across several of the subfields used for normalization.

Reference sets include publications from a wide variety of geographic locales and institutions, and therefore scores registered at this level capture average rather than good or very good scientific achievements. Nevertheless, comparison with reference sets can be useful to identify areas of research practice in need of dire attention and intervention (where scores are below reference level). Of course, a reference set can also be purposefully built to include only publications by high-performing research groups.

Reference sets are often built in consultation with stakeholders of an evaluation/review to ensure that shared and accepted definitions of a thematic area are achieved. Thematic reference sets are typically built by defining keyword-based queries that are then applied to title, abstract, and keywords of articles. While this strategy appears simple, keyword selection needs to be carefully tested and curated manually and individually, possibly by subject experts. Typically, certain keywords may bring in unwanted articles at the same time as they capture articles of interest. For instance, the area of "crop genomics" is most likely of interest to understand CGIAR research, but complex Boolean queries would be required to make sure that only non-medical research with genomic assays are captured for a reference set.

For CGIAR's needs, a reference set on "development agriculture" could be developed using the approach outlined above. Recall tests would need to be performed to make sure that most CGIAR publications fall within the thematic definition obtained (the rate of false negatives should be low). Precision tests would also help determine whether non-CGIAR publications are correctly included in the reference set (set includes true positives and few to no false positives, typically only 5% or less rate of false positives).

R11: CAS/Evaluation, PPU, or MEL professionals should plan the development of thematic queries to delineate a foundational AR4D publication set that will enable comparative assessments of CGIAR research against external institutions. This study should be part of a pilot, expanded bibliometric assessment of CGIAR research to collect initial empirical experiences in deploying a more complex bibliometric pipeline at CGIAR. Support from a commercial bibliometric provider is likely to be necessary during development of this pilot set of thematic publications, but it might be possible for CAS/Evaluation/MEL to develop this capacity in house following this pilot phase.

6.2.3 Comparisons with Partner Institutions to Assess the Added Value of CGIAR Collaborations for These Partners

Partnerships with local research institutions in the South are a recognized priority for CGIAR CRP programs as well as for the upcoming CGIAR 2030 Research and Innovation Strategy, notably as part of the Regional Integrated Initiatives. In that spirit, it may be appropriate to try to showcase the added value and structuring effects that collaborations with CGIAR researchers create for external research partners.

To capture the added value of collaboration with CGIAR, it would be possible to compute bibliometric achievements for the specific group of CGIAR-partner publications. Scores obtained for those co-publications can then be compared with measurements obtained for publications by the partner institutions obtained outside the collaboration.

To achieve this robustness in such an analysis, the aggregate of collaborative publications should contain at least 30 articles. Structuring effects for the partner institution can be expected to be durable and of broad enough breadth only if CGIAR-collaborative publications account for a sizable portion of the partner's publication activities. As far as we know, the optimal portion of publications has not been computed, but it is clear that a collaboration leading to 30 CGIAR-partner co-publications is more likely to register as impact for a partner with 300 publications than one with 3,000 publications or 30,000 publications.

Conducting such a comparison requires that PPU/MEL prepare metadata for bibliometric analysis on the partners' overall publications (beyond publications already captured because they are CGIAR-collaborative articles). Science-Metrix expects that the records kept as part of the CLARISA database can help make this work step more efficient.

6.2.4 Comparisons between Different International Co-Publication Partners to Highlight Features or Achievements of Different Bilateral Relations

As part of the focus group organized to collect CGIAR stakeholders' input on bibliometric assessment, one proposal was made to measure the outcomes of specific bilateral relations, such as bibliometrically characterizing CGIAR Asia-Africa co-publications, potentially with a view to detecting signals of knowledge transfer.

Once country or other geographical attributions have been made to CGIAR publications, such analyses should be straightforward to conduct. Using comparisons will enable analysts to obtain a clearer signal on the unique contributions of these collaborations. Inter-geography collaborations need not be directly compared one with another, but they should be compared against the average outcomes observed throughout CGIAR publications, and possibly against a potential development agriculture reference set.

6.2.5 Comparisons with External Institutions

When the idea of adding a comparative dimension to CGIAR research evaluation for benchmarking purposes was mentioned during engagements with key informants, several CGIAR stakeholders felt that it would be particularly helpful, particularly in cost-benefit analyses of research. Institutions with a similar research focus, a roughly similar organizational profile, and a focus on development—for instance, Northern universities with sizable research portfolios in AR4D—could help identify the unique achievements realized in CGIAR research, point to areas in need of improvement, and potentially satisfy donors' demands to highlight value for their investment.

By way of example, the reader can consult the aforementioned study performed by Science-Metrix for the Belmont Forum (Technopolis Group & Science-Metrix, 2020). This study used external comparisons to help determine whether Forum-supported projects held added value for the Forum's stakeholders (the NSF, the European Commission, UKRI, and other funders) as compared with their core funding mechanisms. To do so, Forum-supported publications were compared with NSF-, EC-, or NERC-supported publications in the subfields of climate and environmental sciences. The findings showed that Forum-supported publications reached higher levels of citation impact, international co-publication, cross-disciplinarity, and some dimensions of policy and online engagement than the publications from the main programs of contributing funders. Therefore, there was clear added value for national funders to invest in the Forum as a supporting mechanism, in addition to their core national award mechanisms.

Comparisons with external institutions also carry a non-negligible risk of false conclusions if they suffer from design flaws, such as comparisons between research groups that are not directly comparable. Selecting research institutions to be included in a comparative exercise is still an art as much as a science, and there is currently no single method or database that offers support for this process. Researchers most commonly select institutions based on prior personal knowledge, desk research, or their own perceptions of the leaders in their own fields. Despite this, factors such as size of the research staff available in each institution, thematic orientation, access to external funding, and many more factors are all likely to affect the degree of comparability of institutions. In an ideal scenario, values for all of these factors should be at roughly the same level or follow similar profiles for institutions to be compared, or otherwise controlled for. In practice, these measurements are seldom publicly available, and so careful interpretation of findings and transparency about limitations are the best safeguard against bias.

As part of evaluation or monitoring workstreams in CGIAR, designing external comparisons will most notably require acquisition of the metadata from relevant publications. Analytical periods should be identical for all publication sets, and distributions of the publications across subfields should be comparable. To help with the latter, it might be useful to restrict comparators' publications to those falling within a development agriculture reference set or other such thematic delineation.

Using normalized findings for all institutions and publications included in the comparison greatly helps to control for any confounding factor not identified ahead of the analysis.

R12: Science-Metrix also strongly recommends that any comparison with Northern institutions include only publications from the latter that have been written in South-North co-authorship. Given that levels of access to funding and specialized infrastructure are completely different along the least- to most-developed-country divide, Science-Metrix believes that any direct comparison of Southern-based research with Northern-based projects are likely to be heavily biased against the former.

The main challenge in conducting these analyses will be combining publication metadata from commercial bibliographic databases with CGIAR-curated metadata, which may be of higher quality, but cannot be extended to external comparators. As already mentioned (section 5.4), Science-Metrix recommends that where comparisons are conducted, CAS/Evaluation/PPU and CGIAR stakeholders accept a duality of findings: (1) curated CGIAR outcomes in descriptive, internally oriented findings and (2) less-curated CGIAR findings to be used specifically for comparative exercises.

6.2.6 Difference-in-Differences (against Comparators)

Comparing a treatment group's performance before, during, and after an intervention (a new programming orientation; a new grant; a new collaboration mechanism; a new instrument or support mechanism to support goals such as gender equity or cross-disciplinarity) provides a measurement of the maximum potential difference brought about by the intervention. Benchmarking against comparable groups that did not receive the intervention of interest but that would have been affected by the same or similar local and global trends allows analysts to isolate more robustly the precise magnitude of changes introduced by an intervention of interest. This strategy is called calculating a difference-in-differences (DID) and is often used in program evaluations (Langfeldt & Scordato, 2015).

Using difference-in-differences approaches, however, requires access to high-quality data on both the intervention of interest and the group it targets, and comparable interventions performed on comparable target groups. Notably, the volume of observations available for all groups included in the comparative analysis needs to be high enough to support robust conclusions (at least 30 publications for each group, but ideally many more).

Obtaining a difference-in-differences score involves making the following computations:

[(*score of treatment group [presumably a CGIAR aggregate] for period p on indicator i)* **–** *(score of treatment group for pre-intervention period q of equal length in years to p on indicator i)]* **–** *[(score of control group for period p on indicator i)* **–** *(score of control group for pre-intervention period q of equal length in years to p on indicator i)]*

The pre- (q) and post- (p) intervention period should be carefully selected as the effective implementation of interventions sometimes lags behind their official announcement. Control groups should be selected from suitable external comparators, with comparators ideally sharing several features with CGIAR research, including focus toward AR4D, funding volume, international collaboration profile, and size as measured by effective full-time equivalent (FTE) of researchers.

In the case of the Belmont Forum study, one confounding factor likely to have affected the analysis was the tendency for researchers to engage in cross-disciplinary, cross-sectoral, and participatory projects to apply for the Forum's support (Technopolis Group & Science-Metrix, 2020). The question arose as to whether good performance on these dimensions could be traced back to a specific effect of Forum funding or were due solely to "self-selection of transdisciplinary researchers" in applications to the Forum's competitions. Science-Metrix compared Forum-funded publications with non-Forum-funded publications by the same authors, and compared both sets of publications with prior publications by the same authors, to tease out the specific differential gain in performance for Forum support on top of the already good performance brought in by the "selection bias" in program competitions.

6.2.7 Comparing Analytical Breakdowns to a Main Publication Set

It is also possible to compute breakdowns of a publication set by categories such as subfields, publications with a high proportion of women's authorships, publications with a high proportion of Southern authorships, and so forth. The indicators identified in the Data Collection Matrix (DCM) can then be applied to these breakdowns to obtain differential findings by group. For example, it is possible to compute indicators on achievements in any dimension included in the DCM for breakdowns, such as:

- The subset of publications
- The subset of highly cross-disciplinary publications
- A subset of collaborative publications with a specific partner or selection of partners
- A subset of publications from a specific geography

This list is just a few illustrative examples. Using these breakdowns allows analysts to answer questions such as "Have women-led CGIAR publications achieved high levels of cross-disciplinarity?" or "Have publications written in collaboration with NGOs led to high levels of policy-related uptake?" Comparisons are implemented insofar as the set of women-led publications is compared with the full publication set, to follow up on the example provided above.

7 Data Collection Matrix (DCM) and Menu of Novel and Improved Indicators: Analysis and Recommendations

Based on the analysis and triangulation of evidence, within the QoR4D framework and consistent with the approach to applying QoS evaluation criterion, the following indicators are recommended for consideration for M&E of quality of One CGIAR research. As one key informant noted, the appropriate indicators differ based on ex ante or ex post status of analyses. The indicators listed here focus on summative evaluations. CAS/Evaluation/PPU may want to consider additional priorities and guidelines for the use of bibliometrics and other indicators in ongoing monitoring and formative evaluation of CGIAR research, including in evaluability assessments, especially if CGIAR wants to meet the needs of research programs (as revealed through engaging with key informants) for real-time QoS monitoring.

7.1 Prioritization of Quantitative Indicators

The focus group, interviews, and surveys provided important feedback by CGIAR stakeholders on the usefulness of current CGIAR indicators in evaluating QoS, as well as thoughts on newly proposed bibliometric indicators. As one key informant noted, an important step in improving the indicators used for assessing QoS in CGIAR research is to prioritize the most important indicators. Therefore, this section includes a description of some of the common themes regarding the usefulness and importance of indicators as identified by key informants, followed by a list of priority indicators. Although not all indicators are mentioned in this section, they are included in the comprehensive DCM (see Annex 4, table A1).

Multiple experts in the evaluation of QoR4D who served as key informants during the development of the technical note, agreed that CGIAR will have to determine if the focus of evaluation as described in this technical note will be QoS or quality of research (QoR) based on its research priorities and where CGIAR research now sits on the impact pathway. It will need to subsequently decide which elements to include in this piece of the overarching QoR4D framework.

Selection of indicators can also be conducted together with the formulation or revision of a theory of change (ToC) elaborated for a given program or project of interest. While ToCs can provide crucial input in the prioritization process, it must be remembered that not all steps or components in a ToC are amenable to bibliometric assessment; and simultaneously, that individual bibliometric findings indicators often offer findings of relevance to multiple components of a ToC. For the design of bibliometric assessments to be informed by a program or project's ToC, it is sometimes necessary to refine the ToC and add details of the specific technical or organizational instruments that are being deployed and implemented by research teams or research administrations.

Important note: 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. To take a hypothetical example, indicators of cross-disciplinarity cannot be said to be of higher priority for program evaluation than citation impact indicators across all cases. For some programs, however, indicators of cross-disciplinarity are of higher relevance to capture program outcomes and the realization of objectives than citation impact indicators are (Technopolis Group & Science-Metrix, 2020).

All new or improved indicators suggested by Science-Metrix fulfill a basic level of relevance for capturing desirable outputs or outcomes as part of the One CGIAR strategy/portfolio. The prioritization of indicators suggested below has been made primarily based on operational, implementation, and cost considerations. In short, all new or improved indicators suggested by Science-Metrix would be relevant for evaluating QoS/QoR4D as part of One CGIAR, but their systematic implementation is likely to result in excessive workloads or costs for CAS/Evaluation/PPU. Given these constraints, a more discriminating assessment of relevance is necessary to select a panel of indicators available in different cost or workload breakdowns.

Furthermore, the panel of indicators provided here has not been designed to evaluate individual researchers or to be used as direct input into formulas for funding decisions, such as those deployed in some national performance assessments (Hicks, 2012). Science-Metrix always advises users to interpret findings from its indicators alongside qualitative and/or peer review evidence as part of a comprehensive triangulation process.

The sections below include tables with indicators of high-, mid-, and low-priority indicators for inclusion in bibliometric analysis to evaluate QoS in the context of One CGIAR. A system of abbreviations has been used to synthesize the large amount of information required for this exercise. Box 1 defines the columns and abbreviations used in these tables.

7.1.1 Level 1 Priority Indicators

Table 6 contains the list of indicators that Science-Metrix suggests are of the highest priority for the evaluation of QoS within One CGIAR. To repeat, assignment in a given priority level is based on a determination of the degree of relevance and alignment of a given indicator to the objectives of One CGIAR, based on the key informant interviews (see section 7.1.3) as well as the review of the One CGIAR and CRP documentation. Assessment of relevance to One CGIAR objectives is balanced against operational considerations and effort or cost considerations, particularly considering the need for CAS/Evaluation/PPU to implement and produce a good portion of these indicators in house. Again, high performances on high-priority indicators are not “better” than high performances recorded for lower-priority indicators.

Within the list of high-priority indicators, inclusion of indicators of equal gender participation (L23 and L24); of altmetrics mention within blogs (R30); of shares of publications that are academic-private co-publications (R34); and of cross-disciplinary integration of the social sciences and humanities (SSH) within publications (R46) have all been directly supported by KIIs. The remaining indicators of normalized citation impact, cross-disciplinarity, and South-South or South-North co-publication are supported by strategic orientations in the CGIAR 2030 Research and Innovation Strategy and the QoS and QoR4D frameworks (CGIAR System Organization, no date). All the dimensions measured or partially measured with the high-priority indicators were previously deployed in the CRP review exercises, through either qualitative indicators (most often at the project rather than publication level) or different quantitative ones.

Except for some normalized citation impact indicators (R38 and R39) and cross-disciplinarity indicators (R42 to R46), all indicators on the list follow a similar basic logic: computing shares of a publication set displaying a certain feature of interest. 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 (all publications from a 2022–24 Initiative or, say, all publications from the Action Area Resilient Agrifood Systems). 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 one women co-author or 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. Most of the work in these examples involves preparing the publication metadata on authors or affiliations. Once this metadata coding work is done and maintained, computing different variations of these indicators to answer specific questions or assess specific intervention logics is highly efficient.

Note that all indicators in the high-priority list involve relatively low levels of effort for in-house computation by CGIAR (the PPU, MEL community, or CAS/Evaluation-engaged analysts) or could be computed by Science-Metrix (or presumably by another commercial bibliometrics provider, although obviously Science-Metrix cannot guarantee this) at low cost for a given set of DOIs. Including these indicators in a comparative strategy may require higher efforts or costs, but those efforts or costs would be traced back to the exigencies of metadata harmonization for robust comparison rather than to indicator computation strictly defined.

Box 1. Glossary and Abbreviations

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.

Table 6. Data Collection Matrix (DCM) for level 1 prioritized indicators

ID	Title	Implementation	Time	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)
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

ID	Title	Implementation	Time	Limits
				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
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

7.1.2 Level 2 Priority Indicators

The level 2 (medium-priority) indicators listed in Table 7 are just as relevant to the evaluation of QoS in One CGIAR as those in the high-priority list. These indicators, however, either

- require more or much more effort (or costs, if subcontracted) to compute robustly than those in the high-priority list (E10, L25, L28, L30, R31, R33, R35, R36, R37, R47, R50, R51),
- have less certain levels of robustness because they are pilot indicators with no published evidence on their use or because Science-Metrix has not used them before (L30, L33, R32, R33, R47, R48, R50, R51),
- require longer outcome realization periods (R29), or
- are somewhat redundant with indicators already presented in the high priority list (R27).

For more details on these indicators, the reader is referred to the full DCM presented in Annex 4 to this report. Selecting a few illustrative examples from the list can nonetheless help provide a sense of both the utility and challenges in using these potential indicators.

The cost-effectiveness of research investments in CGIAR research, as measured by volume of publications per million euros or dollars (E10), has been mentioned in KIIs as highly relevant for some stakeholders to justify continued support to CGIAR's mission. Although this is a strong rationale for computing this indicator, there are multiples constraints to the production of such findings. For all CGIAR publications, including collaborative publications with external co-authors supported by their own funding sources, the fraction of support attributable to each source would need to be precisely determined to allow analysts to assign the corresponding fraction of the publication to each funder involved. This might require extensive survey answer collection or administrative data processing. Other important limitations to the approach include potential recall differences when the indicator is used in a comparative strategy or intrinsic, field-level differences in typical cost of research (Pinheiro et al., 2020). Of course, based on financial and publication volume information that will be readily available at the level of Action Areas and

Initiatives, it will be possible to arrive at a coarse version of this indicator. Science-Metrix suggests that such a coarse version may be sufficient for advocacy and auditing purposes if all limitations of the approach are made clear.

With regard to indicators on co-publications with authors from non-academic sectors (R35 to R37), coding affiliation metadata for comparator institutions whose recurrent partners have not been coded in CLARISA may be prohibitively resource-intensive in a comparative setting. Even commercial providers such as Science-Metrix must conduct these coding procedures with manual curation and a minimum of input from automated methods to ensure acceptable levels of precision and recall for these findings.

The share of publications cited by policy-related documents (evidence syntheses, grey literature, reports from organizations like the UN or World Bank) can be computed robustly using the new database Overton, but policy-related citations typically take three to four years after article publication year to fully accrue (Pinheiro et al., 2021).

An indicator capturing the balance in the seniority of authors of CGIAR publications has been mentioned in a KII as being highly desirable. While bibliometric measuring of author seniority can be reasonably approximated by counting the number of years since an author's first recorded publication, it can be argued that the optimum composition of a research team in terms of seniority is not clear at this point. Somewhat dated findings (obtained from publications issued between 2000 and 2007) showed that PhD students alone contributed to roughly 30% of publications in the natural sciences (Larivière, 2011). If master's students and postdoctoral fellows are included in a similar analysis, the proportion of overall publications with contributions by at least one early-career researcher is likely to be very high, particularly in recent publications. Given these observations, it is currently uncertain how an indicator of balance in authorship seniority might yield non-trivial results. Deploying a bibliometric indicator of balance in seniority would therefore first require some pilot work to establish typical distributions in seniority across research subfields and years. The indicator would then possibly be more useful to raise red flags on problematic situations that deviate from the optimum if it is found that most publications globally are already written by co-authorships with rather diverse seniority levels.

Finally, indicators such as bibliometric companions to OICRs (R33) or share of publications that capture transformative research (R48) must be further specified at a later time, possibly in consultation with CGIAR stakeholders. For instance, the share of transformative publications would include publications that perform particularly well on more than one dimension of cross-disciplinarity or societal impact. This could be computed for publications with a certain ratio of women authors **AND** with a certain ratio of Southern authors **AND** reaching a certain threshold of multidisciplinarity, to take just one possible example. However, as Science-Metrix has never computed such an indicator or specification, some exploration of the distribution of the so-defined publications within research more globally would be desirable to obtain an initial grasp of the features of such transformative publications (concentration in certain subfields or countries, for instance).

Table 7. Data Collection Matrix (DCM) for level 2 prioritized indicators

ID	Title	Implementation	Time	Limits
E10	Count of publications per million euro	CGIAR+	+3	Un-normalized; complex and uncertain funding data acquisition; imperfect proxy; cleaning
L25	Share of publications with explicit conceptualization of gender dimensions	Pilot	+3	Imperfect proxy (gender equity in knowledge production); limited knowledge base; limited technical deployment (access to full texts)
L28	Average publication-level diversity of nationality	CGIAR+	+3	Un-normalized; cleaning; unknown optimum; imperfect proxy; does not capture balance or equity
L30	Normalized index of relative multinational diversity	Pilot	+3	Metadata errors (affiliation data); limited knowledge base (novel indicator); imperfect proxy (equity in multinational integration)
L33	Share of publications with a balanced mix of early career and senior authors	Pilot	+3	Imperfect proxy (equity in seniority); limited knowledge base (novel indicator)
R27	Share of publication with Wikipedia mentions	CGIAR+	+5	Imperfect proxy (public engagement and knowledge transfer); limited knowledge base; metadata errors
R29	Share of publications cited by policy documents	CGIAR+	+7	Geographic and language coverage biases; imperfect proxy (science-policy engagement and knowledge transfer); limited knowledge base; metadata errors
R31	Thematic alignment with SDG-relevant topic	CGIAR+	+3	Imperfect proxy (knowledge transfer for development); limited knowledge base; metadata errors
R32	Google Scholar citations to local-oriented publications *	CGIAR+	+3	Un-normalized; imperfect proxy (impact and outcomes); unknown reference (comparison); restricted characterization of the data source
R33	Bibliometric companions to OICR	CGIAR+	+3	Design complexity; potentially low number of observations

ID	Title	Implementation	Time	Limits
R35	Share of publications that are academic-NGO co-publications	CGIAR+	+3	Un-normalized; extensive cleaning; complex categorical definition; imperfect proxy (technology transfer)
R36	Share of publications that are academic-policymaking co-publications	CGIAR+	+3	Un-normalized; extensive cleaning; complex categorical definition; imperfect proxy (technology transfer)
R37	Share of publications that are academic-governmental research center co-publications	CGIAR+	+3	Un-normalized; extensive cleaning; complex categorical definition; imperfect proxy (technology transfer)
R47	Relative contributions of SSH subfields to cross-disciplinarity indices	Pilot	+3	Imperfect proxy (collaborative disciplinary integration); bias toward novel and radical disciplinary diversity
R48	Share of transformative publications	Extern	+3	Imperfect proxy (knowledge transfer for development); limited knowledge base (novel indicator proposal)
R50	Share of publications first issued as preprint	Pilot	+3	Imperfect proxy (open science practices); complex data acquisition; difficult to normalize
R51	Share of publications associated with an open data release	Pilot	+3	Imperfect proxy (open science practices); complex data acquisition; difficult to normalize; novel and unproven indicator

7.1.3 Key Informant Contributions to the Prioritization of Indicators

Several key informants recommended that CGIAR maintain a focus in developing a framework for assessing QoS using indicators. As one KI noted, it would be better to have a few quality indicators than many that would not be realistic. Another noted that there should be enough indicators so that it would be difficult for researchers to use the metrics for evaluation as targets (i.e., Goodhart's law) but not so many that researchers become frustrated in trying to engage in quality research.

Additionally, a common theme across CGIAR stakeholder input was the question of how to best evaluate impact as research moves down the impact pathway. As noted earlier, applied research can be a priority in AR4D, and this may determine whether the evaluation goal is to understand the QoS in contrast to quality of research. Several informants noted the importance of recognizing partnerships in both publications and other outputs. This input has led Science-Metrix to suggest indicators not only of research excellence but also of relevance and credibility in its updated DCM.

Input collected by key stakeholders through the focus group, interviews, and surveys revealed indicators that have been useful for assessing QoS previously, as well as those that were less useful. Most key informants pointed to journal impact factors as a useful indicator for understanding research impact, although several key informants, including the CRP reviewer for FISH, noted the need to field-normalize this indicator to increase validity for comparison (which was done during CRP reviews through assessment of journal impact factor in quartiles). Some key informants (reviewers for the RTB and WHEAT CRPs) found looking at journal impact factor in quartiles to be particularly useful. Other key informants, however, believed this indicator was not as useful, specifically in the context of reviewing the RICE and GLDC CRPs. It was mentioned that some journals with low journal impact factors nevertheless have an important audience of key stakeholders in developing countries.

Science-Metrix's recommendation takes stock of these diverging views among CGIAR stakeholders on the issue of whether potential CGIAR research with clear development and societal outcomes but with comparatively low achievements in excellence is considered desirable or not.

R13: One CGIAR evaluations of QoS should always simultaneously deploy a panel of bibliometrics that together cover dimensions of credibility, effectiveness, legitimacy, and relevance. Using a mix of bibliometrics that capture multiple dimensions of research excellence and societal engagement in CGIAR research will help achieve a holistic assessment of CGIAR outcomes and help determine whether any trade-offs have been necessary in reaching certain achievements.

Two KIs emphasized the priority of cross-disciplinary research because they viewed cross-disciplinarity as a key characteristic of research that engages in problem solving for engaging policymakers. One of these informants referenced another's work in the discussion about prioritizing cross-disciplinarity in assessing both effectiveness and scientific credibility. Four others emphasized gender as one of the key parameters for assessment. One key informant noted a particular need for indicators for impact assessment (IA).

Notably, at least three key informants separately mentioned the importance of QoS being assessed specifically with reference to the goals of the research. As one key informant who has worked on a CRP review noted, AR4D programs focus largely on applied research, meaning the outputs should be focused within the context of the intended audience (e.g., policymakers instead of other researchers). Another key informant used the example of their own institution, which seeks community engagement and real-world impacts over publications.

7.1.4 Cost Considerations in the Prioritization of Indicators

CGIAR stakeholders and key informants have queried Science-Metrix about standard costs for delivering bibliometric indicators, so as to be able to include these cost levels in planning for future QoS/QoR4D evaluations.

Science-Metrix specializes in highly customized bibliometric analysis, and many of the indicators suggested in the present study would require some pilot development before systematic deployment in evaluations. For these indicators, a "standard" price cannot be given at this point, since R&D work is first required to compute the indicator at scale.

For more established indicators, the costs of indicator computation at scale are very low if a robust list of DOIs is provided to Science-Metrix or another commercial provider. This cost is likely to be around US$5,000 for a dozen or so pre-computed indicators, if no harmonization work needs to be performed by the commercial provider and if no reporting or further analysis is required by CGIAR (that is, if the provider delivers only unprocessed findings at the paper level or simple aggregates for, for example, CGIAR portfolio Action Areas or Initiatives). A common misconception about bibliometrics is that indicator computation amounts to the bulk of effort in producing findings. In most bibliometric studies, rather, costs come from the following sources:

- careful reporting that highlights the limitations of the indicators and the comparative strategies;
- development of communication materials (complex visualizations, slide decks, reports, etc.) targeted to the primary intended audience and their background knowledge on bibliometric analyses; and
- harmonization work and other manual curation necessary to (1) obtain stratifications with high precision and recall figures and (2) allow robust comparison between analytical breakdowns.

Where CGIAR can conduct high-quality harmonization work in house and is able to transmit lists of DOIs for CGIAR publications with associated paper-level metadata to perform stratifications, it will be able to greatly save on costs if dealing with commercial providers of bibliometric data. In such a case, however, CGIAR will itself have to accomplish the work of robust reporting and discussing of limitations in bibliometric analyses, an important task whose demands in terms of level of effort should not be underestimated.

7.1.5 Implementation of the Expanded DCM

Even if implementation is restricted to level 1 priority indicators, Science-Metrix expects that the use of novel data curation practices, normalization and comparative strategies, and novel indicators by CAS/Evaluation, PPU or MEL professionals will amount to a substantial workstream in months and years to come, leading to a significant amount of organizational complexity. Science-Metrix strongly recommends that CGIAR deploy an iterative, step-by-step, and reflexive approach to bibliometric capacity development to ensure that preliminary plans can be corrected and adjusted in view of emerging technical and organizational challenges over the course of implementation.

R14: Science-Metrix recommends that a bibliometric pilot study be conducted specifically to ease the implementation of the recommendations contained in this report pertaining to bibliometrics expertise and production. The practical experience acquired in such a pilot would be useful in developing in-house expertise and capacity at scale, while keeping some amount of flexibility in the implementation process. The pilot should include comparative and normalization strategies to build familiarity with these approaches. To benefit from the learning obtained as early as possible, the pilot should be conducted on short-term outputs of recent CRP publications (2019–21) and longer-term outcomes of 2010–18 CRP publications. Curated publication metadata should be used wherever possible, but if retrospective publication records are difficult to obtain from Centers, they can also be identified through affiliation and acknowledgment metadata for the purposes of the pilot (in which case it will not be possible to aggregate publications by specific CRPs). The bibliometrics pilot study would also be an appropriate context for developing an AR4D thematic publication set.

R15: Science-Metrix suggests that CGIAR, with support from the PPU in alignment with PRMF, explore the possibility of setting quantitative targets to guide bibliometric assessments. Science-Metrix particularly recommends that such targets be explored for equity, diversity, and integration dimensions such as gender balance or Southern-Northern authors balance in CGIAR publications. Targets for research excellence indicators are best set once initial (baseline) bibliometric measurements have been made. Such targets can in principle be set at any level of stratification (Portfolio, Action Area, Initiative), although in a pilot approach it might be best to work at the Action Area level to reduce the amount of customized development work or to select a subset of Initiatives.

7.2 Additional Recommendations on Qualitative Assessment and on Bibliometric Indicators for Publications Not Indexed in the Main Bibliometric Databases

7.2.1 Expanding the CAS/Evaluation Qualitative Toolbox with RQ+ Contextual Factors

Qualitative methods remain the cornerstone of efforts to capture and assess non-journal outputs of research projects and programs. Science-Metrix's assessment of CGIAR and CAS efforts to engage in qualitative assessments of non-journal research outputs is overwhelmingly positive. The mix of narrative case studies, subject matter assessment (including careful reading of many project proposals, output publications, and other materials), and other qualitative evaluation is impressive. Science-Metrix considers that CAS and CGIAR are currently doing better or much better in this respect than other similar organizations it has come across.

Nevertheless, several informants noted the need to have better indicators for capacity development (an area in which Science-Metrix has no prior experience). On the basis of Science-Metrix's less-than-expert reading of the IDRC RQ+ assessment instrument, it would appear CAS/Evaluation could deepen its application of this framework by implementing not only its quality dimensions and subdimensions, but also its tools for assessing contextual factors (maturity of the research field, data environment, organizational research environment, political environment, research capacity strengthening) (IDRC, 2022). Evaluation of QoS may gain in local relevance if CGIAR projects, their outputs, and their outcomes are clustered by contextual profiles (this stratification could in principle be extended to bibliometric assessments as well if contextual profiles are transposed to curated publication-level metadata fields).

R16: Science-Metrix recommends that CAS/Evaluation consider assessing the RQ+ set of contextual factors for One CGIAR projects and then cluster projects along contextual profiles when evaluating QoS. Alternatively, contextual assessments could be tied to specific CGIAR Centers and their local context.

Contextual assessments should also consider what kinds of institutional support are available to individual CGIAR projects or Initiatives. That is because, in Science-Metrix's experience, it is often not enough for funders to declare and present a program as being cross-disciplinary; cross-disciplinarity needs to be actively *deployed* in research projects by requiring funding applicants to include researchers from a minimum number of distinct disciplines, by asking individual researchers to move to a research group operating in a different discipline than theirs (Science-Metrix, 2018), or by employing research coordinators or brokers that actively work to integrate different disciplinary questions and methods in a project (Schneider et al., 2019). To take another example, hypothetically low rates of OA publishing may be deemed problematic in any context, but they are cause for particular concern where a policy to systematically reimburse OA costs has been deployed.

In deploying qualitative assessments of project design, potentially in evaluability assessments, Science-Metrix recommends that CAS/Evaluation or SME reviewers actively identify how project designs have deployed these institutional mechanisms for achieving (as opposed to just planning) project characteristics or project outcomes. If institutional mechanisms are not in place for projects to benefit from or implement, the contextual assessment needs to recognize this. For instance, it would be unfair to compare results in terms of intensity of OA publishing for a group of researchers that has benefitted from financial support to cover OA publishing charges, against a group of researchers who have not had access to the same support instrument.

7.2.2 Quantitative Indicators for Non-Publication Outputs

Currently only limited options are available for quantitative assessment of research outputs and/or outcomes outside of journal publications. Whereas bibliographic databases provide centralized records of a sizable portion of journal-based outputs, there is a lack of centralized databases containing a large proportion of research outputs that are not published in journal articles. Prior evaluations of sources and methodologies have recommended sticking to qualitative methods for evaluating non-journal-based outputs of research (Koier & Horlings, 2015; Vignola-Gagné et al., 2021).

R17: Science-Metrix recommends that upcoming developments be monitored on the use of databases such as Google Scholar and altmetrics databases to capture societal outcomes linked to a larger set of documents than just journal-based publications. Indicator R31 in the DCM raises the possibility of capturing citations of documents in languages other than English and in non-journal formats, following a methodology by Martín-Martín et al. (2018), although caution is very much required as these databases do not allow for normalization or even assessment of subfield in which a given document falls. The Overton database of policy-related documents indexes citations between policy documents. The field is currently under development, and advances in quantitative monitoring may occur soon, although one should also be careful to distinguish between research advances and advances usable in an evaluation context.

Concerning publications in local journals with high relevance to regional communities but not indexed in the main bibliographic databases, the main issue remains control and filtering out of predatory journals. For this matter, Beal’s list remains the authoritative filter to be applied in collating data on these outputs.

8 Annexes

Annex 1: Study Methodology note

Development of the Technical Note followed modified Terms of Reference of the CGIAR Advisory Services (CAS)/Evaluation, based on the proposal by Science-Metrix. Science-Metrix was selected through a competitive process. A consultative and codesign process to develop this Technical Note was implemented from October 2021 through January 2022.

Following an initial briefing meeting with CAS/Evaluation and consultants, Science-Metrix engaged in a documentary review of core documentation and literature on CGIAR governing bodies and stakeholders; documentation on prior evaluations of CGIAR Research Programs; upcoming structure based on the CGIAR reform of 2020 which restructured CGIAR's partnerships, knowledge, and operations to create One CGIAR; and current efforts to manage data on CGIAR research and activities. A detailed outline and an initial Technical Note were drafted based on document reviews. To solicit input and feedback for incorporation into a second draft of the Technical Note Science-Metrix and the CAS/ Evaluation Function then conducted the following data collection activities with stakeholders and subject matter experts (referred to herein as "key informants"):

- a focus group discussion,
- interviews, and
- surveys

CAS/Evaluation consultants Jillian Lenne and Stefania Sellitti supported co-development of the interview guides, participated in the interviews and led implementation of the survey.

A final validation meeting was conducted with CAS, consultants, and other subject matter experts. At the same time, a second draft was then reviewed by Guy Poppy of the CAS Evaluation Reference Group; Enrico Bonaiuti (ICARDA Research Team Leader – Monitoring, Evaluation, and Learning); and CAS consultants Jillian Lenne, Paolo Sarfatti and Stefania Sellitti, prior to revision of the final Technical Report.

Annex 2: Interview and Focus Group Protocol

A subset of the following questions was asked of subject matter experts and other stakeholders regarding the use of bibliometrics in evaluating QoS in CGIAR research.

Introduction

Good morning/afternoon and thank you for taking the time to meet with us. My name is […], and my colleagues and I are collaborating with CGIAR to conduct research on ways to improve the assessment of Quality of Science within CGIAR. We are conducting a series of interviews to collect opinions and information from experts in the subject, to integrate the suggestions into a Technical Note that will be used to develop Guidelines that will be used to evaluate QoS in One CGIAR.

1. Before we begin, do you have any questions?
2. May we have your permission to begin?

Core questions

3. Please tell us briefly about your background, as well as current and past involvement with assessment of Quality of Science and, if relevant, with CGIAR.
4. What do you think are the most important qualitative and quantitative criteria or indicators for assessment in agricultural research for development (at CGIAR but also elsewhere), and why?
5. In what ways have you deployed these indicators and criteria in your project/program/institute review work? What are the main criteria/indicators that you have used – qualitative and quantitative?
6. What have been the most robust indicators to work with, and why?
7. What have been the least useful indicators to work with, and why?
8. What limitations have you found to the value of some/all these criteria/indicators, if any?
    - Do limitations tend to come more in design or from data quality and availability issues?
9. What suggestions do you have to improve the most important criteria/indicators, if any?
10. Do you find that indicators or findings are equally valued by all audiences and stakeholders? Which indicators have a more restricted audience or impact, if any, and in what ways?
11. Have you had any experience in working with researchers to increase awareness for and include QoS principles in their research activities? If so, could you expand on this experience?
12. One recommendation by Science-Metrix is very likely to work with subfield- and year- normalized quantitative indicators, which will increase the validity for using these indicators in research assessment. But that means working with more abstract indices rather than volumes, counts or proportions. What do you think about these proposed changes?
13. At what level of aggregation do you find indicators and findings to be most relevant? By portfolio initiatives, CGIAR centers, CGIAR countries, any other? If you had to choose one?
14. How would you react to the use of comparative strategies in CGIAR research assessment? Would it make sense to compare CGIAR's degree of women authorship or developing country authorship in research publications against the equivalent figure for say Cornell University's work in agricultural research for development?
15. Respondent-specific questions (WHEAT CRP)
16. Page 15 of the WHEAT CRP review– alignment of research questions and methods: Did Victor and Donna read the papers themselves? Support from peer review?
17. Did the WHEAT CRP reviewers get support from CAS in computing the indicators used, such as IF5? I see Max and Enrico in the acknowledgments; how did this relationship work? Did Victor request the indicators he needed and set requirements? The CAS bibliometric indicators used in the review:
    - IF5
    - H-index
    - Altmetrics
    - Share of OA papers
    - Network analysis
18. There are other indicators in use by CAS, such as diversity of nationalities in papers and diversity of disciplines – why were these not included in the WHEAT CRP review?

19. You discussed gender in the review. Would an indicator of gender equality in publications have been interesting here, and why or why not?
20. What are reasons to conduct a CRP review between 2017 and 2019, when most scientific outputs such as citations have yet to fully accrue and cannot be robustly computed? Are there obstacles to using a longer evaluation window (such as 2012-2017 for bibliometrics specifically)? What are your thoughts about calculating the h-index just for 2017-2019 papers?
21. The WHEAT CRP review included that "Further, WHEAT and MAIZE biometricians delivered free software to support breeding decisions." – was there any evidence of adoption of this software? If so, please explain.
22. The happy seeder story is super interesting. Could we have had more details on legitimacy aspects within the story and how they connect to effectiveness aspects – how were policymakers and farmers and entrepreneurs engaged exactly by the research team, what kind of data collection was done jointly, and was there any co-creation? Did these linkages require inter-disciplinary brokering?
23. You emphasize the reliance of WHEAT on collaborators – were top collaborators and recurring co-authors identified? What was the share of collaborative papers among WHEAT CRP publications, compared to other CRPs? You mention too much reliance on co-authorship; were there many collaborations that do not lead to publications?

Optional questions

24. How have partnerships and collaboration evolved in the agricultural research for development field? How do you collaborate yourself?
25. What are the most relevant CGIAR mechanisms to support specific research approaches from recent years? What about the context as it relates to mechanisms to increase cross-disciplinarity or gender equality, such as requiring projects to have a minimum number of disciplines or women authors? How are such objectives concretely supported for implementation?

Annex 3: Survey Instrument

The following questions were asked to subject matter experts and evaluators to the 2020 CRP reviewers through a short online survey. All questions are open-ended.

1. Name and Last Name
2. In which CRP review did you participate?
3. What were the most robust qualitative and/or quantitative indicators used to evaluate QoS and why?
4. What were the least useful qualitative and/or quantitative indicators used to evaluate QoS, and why?
5. What limitations, if any, were identified based on data quality and availability or design?
6. What suggestions do you have, if any, for additional indicators to evaluate QoS?
7. What suggestions do you have, if any, as to how to improve the most important indicators?

Annex 4: Comprehensive Data Collection Matrix with current, improved, and novel indicators

A comprehensive data collection matrix (DCM) includes both indicators already in use by CAS for CRP reviews and the CGIAR online dashboard (***Table A8***). A system of abbreviations has been used to try and synthesize the large amount of information required for this exercise. The columns and abbreviations used in the DCM are as follows:

Box 2. Glossary and Abbreviations

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.

Table A8: Comprehensive data collection matrix for One CGIAR, with current and recommended indicators

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
C01	Desc	Current	Peer-reviewed papers by top journals	Five-year normalized impact factor and distribution in quartiles	Field biases; imperfect proxy (credibility, discrimination in peer-review); unknown optimum or reference (comparison)	yes	yes
C02	Desc	Current	List of top partners institutes co-publishing peer-reviewed papers with CGIAR research Centers and CRPs.	Count of co-authorships between CGIAR and external researchers	Nothing major when used descriptively rather than for benchmarking		yes
C03	Quant	Current	IF5 or Average of relative CiteScores	Considers the likely prestige of journals in which a set of publications are found as well as being a potential early indicator of future citation impacts for those papers that have the smallest citation windows.		yes	yes
C04	Qual	Current	Project(s) objectives are aligned and significant to CGIAR Impact Areas	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C05	Qual	Current	Project(s) objectives are	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially		

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
			clear and appropriate		resource-intensive; potentially limited scale		
C06	Qual	Current	Project(s) design is coherent, appropriate, and rigorous	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C07	Qual	Current	Project(s) design plans for linkages, synergies, and complementarities	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C08	Qual	Current	Project(s) internal disciplinary skill base is appropriate and adequate	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C09	Qual	Current	Project(s) external disciplinary skill base complements the internal skill base	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C10	Qual	Current	Project team(s) has an appropriate diversity of nationality	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C11	Qual	Current	Project team(s) has an appropriate diversity of	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
			geographical locations				
C12	Qual	Current	Project team(s) achieve trans-disciplinary integration of disciplines	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C13	Qual	Current	Project(s) internal research infrastructure is appropriate and adequate	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C14	Qual	Current	Project(s) external research infrastructure is appropriate and adequate	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
C15	Qual	Current	Project(s) funding model is implementable	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
E01	Qual	Current	Project management cycle	Planned activities carried out as scheduled	Out of scope for Science-Metrix		
E02	Desc	Current	Total peer-reviewed papers	Number of CGIAR research papers published in peer-reviewed journals	Un-normalized; unknown optimum (comparisons needed)	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
E03	Desc	Current	Knowledge product published	Number of knowledge products published	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
E04	Desc	Current	Online publications	Number of online publications	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
E05	Desc	Current	Research products disseminated	Number of research products disseminated	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
E06	Desc	Current	Datasets	Number of datasets generated	Field biases; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
E07	Desc	Current	Thesis	Numbers of master/PhD thesis developed through the research project	Field biases; unknown optimum or reference (comparison)	no	yes
E08	Qual	Current	Project(s)' design plans integration with communities of practice	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
E09	Quant	Current	H-index	Maximum value of number publications holding the same number of citations of more each	Un-normalized; Imperfect proxy (quality and intellectual achievement)	Pilot	yes
L01	Desc	Current	Diversity of age	Subject matter or peer review assessment of project proposal	Risk of tokenism; unknown reference (comparison)	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L02	Quant	Current	Gender balance in team participation	Percentage of male and female team members	Possible discrepancy between formal team and participation in activities (incl. publications)	no	yes
L03	Desc	Current	Team-level diversity of nationality	Number of countries/nationalities represented within the team	Possible discrepancy between formal team and participation in activities (incl. publications); un-normalized; unknown optimum; imperfect proxy; does not capture balance or equity	no	yes
L04	Qual	Current	Scientists' diversity	Number of scientists coming from different fields and with different expertise	Possible discrepancy between formal team and participation in activities (incl. publications); un-normalized; unknown optimum; imperfect proxy; does not capture balance or equity	no	yes
L05	Desc	Current	Number of students	Number of students and other participants in the project research activities	Possible discrepancy between formal measurement and realized intensity; imperfect proxy	no	yes
L06	Desc	Current	PhD students	Number of PhD students participating in the research activities	Possible discrepancy between formal measurement and realized intensity; imperfect proxy	no	yes
L07	Desc	Current	Long-term trainees	Number of academic trainees	Field biases; unknown optimum or reference (comparison)	no	yes
L08	Desc	Current	Short-term trainees	Number of trainees (all other types of trainees)	Field biases; unknown optimum or reference (comparison)	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L09	Quant	Current	Female students trained	Percent share of female students trained	Unknown reference (comparison); risk of tokenism	no	yes
L10	Desc	Current	Exchange	Number of people undertaking exchange visits	Possible discrepancy between formal measurement and realized intensity; imperfect proxy	no	yes
L11	Desc	Current	Partnership count	Number of partnerships established	Field biases; imperfect proxy (multidisciplinarity); unknown optimum or reference (comparison); risk of tokenism;	no	yes
L12	Qual	Current	Sectoral characterization of partners	Type of partners (Academic, NARS, Private, etc.)	Field biases; imperfect proxy (multidisciplinarity); unknown optimum or reference (comparison); risk of tokenism;	no	yes
L13	Qual	Current	Expertise characterization of partners	Partners by main area of expertise (Capacity development, policy, research, delivery, other)	Field biases; imperfect proxy (multidisciplinarity); unknown optimum or reference (comparison); risk of tokenism;	no	yes
L14	Qual	Current	Project team(s)' composition reflects gender inclusiveness, responsiveness, integration and leadership	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L15	Qual	Current	Project team(s) enact equitable recognition	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L16	Qual	Current	Project team(s) recognize contributions from members	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L17	Qual	Current	Project team members are all involved in co-design and delivery	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L18	Qual	Current	Project team members feel roles and responsibilities are clearly defined	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L19	Qual	Current	Project(s) deploy incentives and rewards for quality	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L20	Qual	Current	Project(s) have an adequate risk assessment and risk mitigation strategy	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L21	Qual	Current	Project(s) collaboration networks are appropriate and adequate	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
L22	Qual	Current	Project team(s) have produced IPG with broad application potential	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
R01	Desc	Current	Citation counts to policy diagnostic tools	Number of citations of diagnostic tools in new policy documents	Database limitations (altmetric); imperfect proxy; un-normalized	no	yes
R02	Qual	Current	Contribution to System-level outcomes	Number of contributions to the three CGIAR System-level outcomes	Out of scope for Science-Metrix	no	yes
R03	Desc	Current	Altmetric.com composite score	Measure the number of total mentions of CGIAR publications (including those that are not peer reviewed) in online media, providing an indication of reach and influence.	Un-normalized; imperfect proxy (knowledge transfer); technical shortcomings	yes	yes
R04	Desc	Current	OA scientific publications	Number of scientific publications in an OA environment	Un-normalized; imperfect proxy (knowledge transfer)	yes	yes
R05	Desc	Current	OA technical reports	Number of technical reports published in open access (OA) environment	Un-normalized; imperfect proxy (knowledge transfer)	no	yes
R06	Desc	Current	Communication material	Number of new communication material produced	Un-normalized; imperfect proxy (impact and outcomes);	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
					unknown optimum or reference (comparison)		
R07	Desc	Current	Guidelines	Number of guidelines produced	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R08	Desc	Current	Policy briefs	Number of policy briefs produced	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R09	Qual	Current	Lessons learned	Number of outcome stories and lessons learned from producers	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R10	Desc	Current	Maps	Number of assessments and maps completed	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R11	Desc	Current	Sharing events	Number of knowledge and sharing events	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R12	Qual	Current	Innovations	Number of research and development innovations	Un-normalized; unknown optimum or reference (comparison)	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R13	Desc	Current	Digital and innovative output	Number of digital and innovative outputs identified and utilized for providing training	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R14	Qual	Current	Adoption of technology	New technology was adopted	Un-normalized; unknown optimum or reference (comparison)	no	yes
R15	Desc	Current	Download	Knowledge products download	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R16	Desc	Current	Downloads of reports	Number of downloads of research reports and resources from webpages on selected websites	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R17	Desc	Current	Downloads of datasets	Number of visits/downloads of datasets and tools by country NARS	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R18	Qual	Current	Reachability	People that received the disseminated knowledge products	Un-normalized; imperfect proxy (impact and outcomes); unknown optimum or reference (comparison)	no	yes
R19	Qual	Current	Contribution to System-level outcomes (1, 2 and 3)	Number of contributions to the three CGIAR System-level outcomes	Out of scope for Science-Metrix	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R20	Desc	Current	Outcome Impact Case Reports (OICR)	Number of OICR	Definition complexity; Field biases; unknown optimum or reference (comparison)	no	yes
R21	Qual	Current	Outcome Impact Case Reports (OICR)	Narration of impact, including: partners; authors…	Nothing major when used descriptively rather than for benchmarking	no	yes
R22	Qual	Current	Project(s) research topic alignment to CGIAR Impact Areas	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
R23	Qual	Current	Project(s) have appropriate and realistic research plan	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
R24	Qual	Current	Designs for usefulness and capacity building are appropriate and adequate	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
R25	Qual	Current	Methodological outcomes are appropriate for next users	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		
R26	Qual	Current	Team(s) actively stewards policy adoption of project outcomes	Subject matter or peer review assessment of project proposal	Discrepancies between plans and achievements; potentially resource-intensive; potentially limited scale		

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
E10	Quant	PUU/CGIAR+	Publications per million euro	Number of peer-reviewed articles produced by a research group per year and per volume of funding. Joint external funding found in acknowledgements should be accounted for and publication volume fractionated for funding fractions.	Un-normalized; complex and uncertain funding data acquisition; imperfect proxy; cleaning	no	costly
E11	Quant	PUU/CGIAR+	Sum of normalized citations per million euro	Normalized citation level for peer-reviewed articles produced by a research group per year and per volume of funding. Joint external funding found in acknowledgements should be accounted for and publication volume fractionated for funding fractions.	Un-normalized; complex and uncertain funding data acquisition; imperfect proxy; cleaning	no	costly
E12	Quant	PUU/CGIAR+	Sum of highly cited publications (top 10%) per million euro	Normalized share of peer-reviewed articles falling into the top decile of most cited publications in its year and subfield, produced by a research group per year and per volume of funding. Joint external funding found in acknowledgements should be accounted for and publication volume fractionated for funding fractions.	Un-normalized; complex and uncertain funding data acquisition; imperfect proxy; cleaning	no	costly
L23	Quant	PUU/CGIAR+	Women participation in authorship	Share of publications with participation by at least one woman researcher	Does not capture balance or equity; may capture tokenism; paying software (NamSor); margin of error (especially for Asian people); un-normalized;	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L24	Quant	PUU/CGIAR+	Gender balance in key authorship	Integration of women as first, corresponding or (last) senior author	Paying software (NamSor); margin of error (especially for Asian people)	yes	yes
L26	Quant	PUU/CGIAR+	Shares of North-South/South-South co-publication	Share of publications that are international co-publication and include at least one southern country in authors' affiliations; or two different southern countries	Un-normalized; cleaning; unknown optimum; imperfect proxy; does not capture balance or equity	no	yes
L27	Quant	PUU/CGIAR+	Southern authors' participation as first, corresponding, or last	Share of publications where the first, last or corresponding author is located in a low or low-middle income country; or where two out of three of these positions are occupied by southern researchers	Error rate in affiliation data; imperfect proxy (south-north equity);	yes	yes
L28	Quant	PUU/CGIAR+	Average publication-level diversity of nationality	Avg count of unique countries within authors' affiliations, averaged at the publication level	Un-normalized; cleaning; unknown optimum; imperfect proxy; does not capture balance or equity	yes	yes
R27	Quant	PUU/CGIAR+	Share of publication with Wikipedia mentions	Based on a de-composition of the altmetrics composite score into constituent dimensions	Imperfect proxy (public engagement and knowledge transfer); limited knowledge base; Metadata errors;	yes	yes
R28	Desc	PUU/CGIAR+	Share of publication with journalistic mentions	Mentions towards peer-reviewed articles in online journalistic articles in sources such as The New York Times, Washington Post, Le monde, the Guardian, The Conversation, Smithsonian; and, in platforms aggregating academic press releases.	Imperfect proxy (public engagement and knowledge transfer); limited knowledge base; Metadata errors;	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
				Ideally requires some curation and filtering of sources.			
R29	Quant	PUU/CGIAR+	Share of publications cited by policy documents	Entity's paper mentioned at least once in a policy document	Geographical and language coverage biases; Imperfect proxy (science-policy engagement and knowledge transfer); limited knowledge base; metadata errors	yes	yes
R30	Quant	PUU/CGIAR+	Share of publication with blog mentions	Based on a de-composition of the altmetrics composite score into constituent dimensions	Imperfect proxy (public engagement and knowledge transfer); limited knowledge base; Metadata errors;	yes	yes
R31	Quant	PUU/CGIAR+	Thematic alignment with SDG-relevant topic	Shares of a given publication set that fall within each SDG	Imperfect proxy (knowledge transfer for development); limited knowledge base; metadata errors	yes	yes
R32	Desc	PUU/CGIAR+	Google Scholar Citations to local-oriented publications *	Google Scholar captures citations links between a wide variety of documents, including books and other non-publication outputs; as well as in non-English languages. Academically developed software packages allow retrieval of these citation numbers at scale.	Un-normalized; imperfect proxy (impact and outcomes); unknown reference (comparison); restricted characterization of the data source	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R33	Quant	PUU/CGIAR+	Bibliometric companions to OICR	Bibliometric characterization of the research groups (and their publications) that have led the realization of the outcomes captured in the OICR narratives. The bibliometric case studies can help characterize how certain research practices support the realization of societal outcomes of research	Design complexity; potentially low number of observations	yes	yes
R34	Quant	PUU/CGIAR+	Academic-private co-publications	Papers published in collaboration with private sector	Un-normalized; Extensive cleaning; Complex categorical definition; imperfect proxy (technology transfer)	costly	costly
R35	Quant	PUU/CGIAR+	Share of publications that are academic-NGO co-publications	Share of publications written as co-authored between a research team and representatives from NGOs. Such publications may capture instances of knowledge transfer towards or co-creation with civil society.	Un-normalized; Extensive cleaning; Complex categorical definition; imperfect proxy (technology transfer)	no	yes
R36	Quant	PUU/CGIAR+	Share of publications that are academic-policymaking co-publications	Share of publications that are co-authored between a research team and representatives from governmental policymaking agencies. Such publications may capture instances of knowledge transfer towards or co-creation with governments and other policymakers.	Un-normalized; Extensive cleaning; Complex categorical definition; imperfect proxy (technology transfer)	no	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R37	Quant	PUU/CGIAR+	Share of publications that are academic-governmental research center co-publications	Share of publications that are co-authored between a research team and representatives from governmental research centers. Such publications may capture instances of knowledge transfer towards applied government-steered research	Un-normalized; Extensive cleaning; Complex categorical definition; imperfect proxy (technology transfer)	no	yes
R41	Quant	PUU/CGIAR+	Average of relative citations (ARC)	Average of relative citation scores (normalized by subfield, year and/or document type) of all the articles published by a given entity	Imperfect proxy (publication quality and intellectual achievement); sensitive to outliers; 30 publications or more required; computable 2 years or more after publication year	yes	yes
R38	Quant	Extern	Share of highly cited publications (HCP)	Share of publications within the top 10% of most cited papers worldwide	Imperfect proxy (publication quality and intellectual achievement); 30 publications or more required; computable 2 years or more after publication year	yes	yes
R39	Quant	Extern	Citation distribution index (CDI)	Composite indicator constructed from each decile's score	Imperfect proxy (publication quality and intellectual achievement); 30 publications or more required; computable 2 years or more after publication year	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R42	Quant	Extern	Index of interdisciplinary integration	Diversity of subfields associated within the reference lists of publications	Imperfect proxy (intellectual disciplinary integration); bias towards novel and radical interdisciplinarity; abstract index most meaningful as part of comparisons	yes	yes
R43	Quant	Extern	Highly interdisciplinary papers	Share of an entity's papers falling within the top 10% of highly interdisciplinary papers in the world	Imperfect proxy (intellectual disciplinary integration); bias towards novel and radical interdisciplinarity	yes	yes
R44	Quant	Extern	Index of multidisciplinary integration	Diversity of subfields associated with the prior publications of co-authors of an article	Imperfect proxy (collaborative disciplinary integration); bias towards novel and radical disciplinary diversity	yes	yes
R45	Quant	Extern	Highly multidisciplinary papers	Share of an entity's papers falling within the top 10% of highly multidisciplinary papers in the world	Imperfect proxy (collaborative disciplinary integration); bias towards novel and radical disciplinary diversity	yes	yes
R46	Quant	Extern	Chord diagram visualization of interdisciplinarity (notably to capture SSH participation)	Visual representations of co-reference relationships between subfields, based on components from the interdisciplinarity index	Un-normalized; Imperfect proxy (interdisciplinary integration); bias towards novel and radical interdisciplinarity; unknown reference (comparison)	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R48	Quant	Extern	Share of transformative publications	Share of publications fulfilling a minimum number of dimensions that contribute to "transformative research": SDG aligned; high interdisciplinary; high multidisciplinary; North-South co-publication; at least one non-academic partner; and/or woman as corresponding, first or last author	Imperfect proxy (knowledge transfer for development); limited knowledge base (novel indicator proposal)	yes	yes
R49	Quant	Extern	Share of publications cited in patents	Share of publications that have received a citation in one or more patent(s), potentially indicating a first step in technology transfer. Requires extensive curation of patent databases.	Requires a citation window of 7 years or more; imperfect proxy (technology transfer)	yes	yes
L25	Quant	Pilot	Share of publications with explicit conceptualization of gender dimensions	Semantic analysis and text mining of research content to identify inclusion of gender-related concepts in the methods and conceptual frameworks of articles	Imperfect proxy (gender equity in knowledge production); limited knowledge base; limited technical deployment (access to full texts)	yes	yes
L30	Quant	Pilot	Normalized index of relative multi-national diversity	Developing a new indicator based on affiliation metadata, that not only considers the variety of countries involved in an international co-publication, but also the relative rarity of these country-pair linkages	Metadata errors (affiliation data); Limited knowledge base (novel indicator); imperfect proxy (equity in multi-national integration)	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
L31	Quant	Pilot	Chord diagram visualization of international co-publications	Chord diagrams illustrating linkages between country pairs in international co-publications colored by degree of relative rarity of the country connections co-authorship link; countries displayed on the circumference of the graph, grouped by continent or income level	Metadata errors (affiliation data); Limited knowledge base (novel indicator); imperfect proxy (equity in multi-national integration)	yes	yes
L32	Quant	Pilot	Share of publications with southern field work or mention of a southern location	Querying for lists of southern cities, regions, or countries to retrieve instances of fieldwork in these locations. Knowledge transfer towards local innovation is expected to require direct engagement with local conditions upstream in the innovation process.	Imperfect proxy (local engagement in southern countries); limited knowledge base (novel indicator)	yes	yes
L33	Quant	Pilot	Share of publications with a balanced mix of early career and senior authors	Share of publications with a desirable balance and diversity in lengths of publication careers by author	Imperfect proxy (equity in seniority); limited knowledge base (novel indicator)	yes	yes
R47	Qual	Pilot	Relative contributions of SSH subfields to cross-disciplinarity indices	Pilot indicator that would need to be developed by Science-Metrix, based on the multidisciplinarity or interdisciplinarity indices	Imperfect proxy (collaborative disciplinary integration); bias towards novel and radical disciplinary diversity	yes	yes

ID	Type	Implem	Indicator title	Description and data collection	Limitations	Norm	Comp
R50	Quant	Pilot	Share of publications first issued as preprint	Retrieval of preprint information from preprint servers to connect preprints to publications. Greater recall is achieved by using algorithms to match preprints' authors and titles to article records in the full bibliographic database.	Imperfect proxy (open science practices); complex data acquisition; difficult to normalize	no	yes
R51	Quant	Pilot	Share of publications associated with an open data release	Retrieval of open data release on repositories using the Data Monitor aggregator and a fuzzy matching algorithm.	Imperfect proxy (open science practices); complex data acquisition; difficult to normalize; novel and unproven indicator.	no	yes

Annex 5: Key Informants and Focus Group Participants

Name	Position	KII/FGD
1. Bas Bouman	Director of CGIAR Research Program Rice, Agrifood Systems	KII
2. Ben Bennet	Subject Matter Expert at the 2020 review of the CGIAR Research Program Livestock; Professor of International Trade and Marketing Economics and Deputy Faculty Director at Natural Resurce Institute (NRI), UK	KII
3. Brian Belcher	Professor at Royal Roads University, Canada	KII
4. Enrico Bonaiuti	Research Team Leader, Monitoring, Evaluation ad Learning, ICARDA, CGIAR center	KII/FGD
5. Margaret Gill	Chairperson of ISDC until 2019, The University of Aberdeen	FGD
6. Michael Friedmann	Science Officer of CGIAR Research Program Root, Tubers and Bananas (RTB)	KII
7. Paolo Sarfatti	Evaluation Senior Strategic and Technical Advisor at CGIAR	FGD
8. Paul Engel	Evaluator of CGIAR Research Program on Policies, Institutions, and Markets (PIM); Founder of Knowledge Perspectives and Innovation	FGD
9. Rachid Serraj	Professor at Mohammed VI Polytechnic University, Morocco	KII
10. Raphael Nawrotzki	Monitoring & Evaluation Advisor at GIZ, member of the	KII
11. Ravi Kumar	Evaluator of CGIAR Research Program Rice, Associate Professor of Monitoring and Impact at Natural Resurce Institute (NRI), UK	FGD
12. Robert McLean	Senior Program Specialist in Policy and Evaluation at IDRC, Canada	KII
13. Valentina de Col	Agricultural Information System Officer, ICARDA, CGIAR center	KII/FGD
14. Victor Sadras	Subject Matter Expert at the 2020 review of CGIAR Research Program Rice; Affiliate Professor at University of Adelaide, Australia	FGD

Annex 6: Validation Meeting Participants

Name	Position
1. Brian Belcher	Professor at Royal Roads University, Canada; CIFOR Senior Associate Scientist
2. Enrico Bonaiuti	Research Team Leader, Monitoring, Evaluation ad Learning, ICARDA, CGIAR center
3. Guy Poppy	Professor at University of Southampton, CAS Evaluation Reference Group member
4. Helen Altshul	Performance and Partnerships Manager, CGIAR Research Program on Livestock, ILRI, CGIAR center
5. Karen Garret	Preeminent Professor at University of Florida
6. Paolo Sarfatti	Evaluation Senior Strategic and Technical Advisor at CGIAR
7. Plex Sula Aaron	Research Assistant, Plant Pathology Department and Food Systems Institute, University of Florida
8. Thiele Graham	CGIAR Research Program on Roots, Tubers and Bananas, Program Director
9. Valentina De Col	Agricultural Information System Officer, ICARDA, CGIAR center
10. Zenda Ofir	International Evaluation and Design Specialist; Interim Chair, Council of the International Evaluation Academy (IEAc); CAS Evaluation Reference Group member *(Written contribution)*

References

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

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. In B. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (pp. 66–77). http://science-metrix.com/?q=en/publications/conference-presentations/towards-a-multilingual-comprehensive-and-open-scientific

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329–342. http://www.akademiai.com/index/34W733263N36H661.pdf

Beaudreault, A., & Meinke, H. (2021). Constructive criticism: ISDC’s external review of Initiative proposals. CGIAR CAS blog, August 16. https://cas.cgiar.org/isdc/news/constructive-criticism-isdcs-external-review-initiative-proposals

Belcher, B. M., & Hughes, K. (2020). Understanding and evaluating the impact of integrated problem-oriented research programmes: Concepts and considerations. *Research Evaluation, 30*(2), 154–168. https://doi.org/10.1093/reseval/rvaa024

Bornmann, L., Ganser, C., & Tekles, A. (2022). Simulation of the h index use at university departments within the bibliometrics-based heuristics framework: Can the indicator be used to compare individual researchers? *Journal of Informetrics*, *16*(1), 101237. https://doi.org/10.1016/J.JOI.2021.101237

CAS Secretariat (CGIAR Advisory Services Shared Secretariat). (2012). *CGIAR Policy for Independent External Evaluation.* Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020a). *CGIAR Research Program 2020 Reviews: Agriculture for Nutrition and Health*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020b). *CGIAR Research Program 2020 Reviews: Grain Legumes and Dryland Cereals*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020c). *CGIAR Research Program 2020 Reviews: Livestock*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020d). *CGIAR Research Program 2020 Reviews: Policies, Institutions, and Markets*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020e). *CGIAR Research Program 2020 Reviews: Roots, Tubers and Bananas (RTB)*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2020f). *CGIAR Research Program 2020 Reviews: WHEAT*. Rome: CAS Secretariat Evaluation Function

CAS Secretariat. (2021a). *Synthesis of Learning from a Decade of CGIAR Research Programs*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2021b). *2021 Synthesis Report: Annexes*. Rome: CAS Secretariat Evaluation Function.

CAS Secretariat. (2022). *CGIAR Evaluation Policy*. Rome: CAS Secretariat Evaluation Function.

CGIAR System Organization. (no date). CGIAR 2030 Research and Innovation Strategy. Transforming food, land, and water systems in a climate crisis. Montpellier : France.

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).

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

ICARDA (International Center for Agricultural Research in the Dry Areas). (2021). *ICARDA Annual Report 2020*. Beirut, Lebanon.

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

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

Jasanoff, S. (2005). Why compare? In *Designs on Nature: Science and Democracy in Europe and the United States* (pp. 13–41). Princeton, NJ: Princeton University Press.

Koier, E., & Horlings, E. (2015). How accurately does output reflect the nature and design of transdisciplinary research programmes? *Research Evaluation*, *24*(1), 37–50. https://doi.org/10.1093/reseval/rvu027

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

Larivière, V. (2011). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics, 90*(2), 463–481. https://doi.org/10.1007/S11192-011-0495-6

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. https://doi.org/10.1016/J.JOI.2018.09.002

McLean, R. K. D., & Sen, K. (2019). Making a difference in the real world? A meta-analysis of the quality of use-oriented research using the Research Quality Plus approach. *Research Evaluation*, *28*(2), 123–135. https://doi.org/10.1093/RESEVAL/RVY026

Noyons, E., & Ràfols, I. (2018). Can bibliometrics help in assessing societal contributions of agricultural research? Exploring societal interactions across research areas. In P. Wouters, R. Costas, T. Frassen, & A. Yegros-Yegros (Eds.), *Proceedings of the 23rd International Conference on Science and Technology Indicators*. *Science, Technology and Innovation Indicators in Transition*. Leiden: The Netherlands. https://hdl.handle.net/1887/64521

Pinheiro, H., Campbell, D., Vignola-Gagné, E., & Hellwig, J. (2020). *Science Quality and Research Impact Study: Advisory Service on Agricultural Research for Development (BEAF): Final Synthesis Report*. Quebec: Science-Metrix. www.science-metrix.com

Pinheiro, H., Vignola-Gagné, E., & Campbell, D. (2021). A large-scale validation of the relationship between cross-disciplinary research and its uptake in policy-related documents, using the novel Overton altmetrics database. *Quantitative Science Studies*, *2*(2), 1–27. https://doi.org/10.1162/qss_a_00137

PPMI and Science-Metrix. (2019). *ERA Progress Report 2018: Technical Report.* Brussels, European Commission. https://ec.europa.eu/info/sites/info/files/research_and_innovation/era/era_progress_report_2018-technical.pdf

Rivest, M., Kashnitsky, Y., Bédard-Vallée, A., Campbell, D., Khayat, P., Labrosse, I., … James, C. (2021). Improving the Scopus and Aurora queries to identify research that supports the United Nations Sustainable Development Goals (SDGs) 2021. https://doi.org/10.17632/9SXDYKM8S4.4

Rivest, M., Vignola-Gagné, E., & Archambault, É. (2021). Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. *PLoS One*, *16*(5 May), e0251493. https://doi.org/10.1371/journal.pone.0251493

Schneider, F., Buser, T., Keller, R., Tribaldos, T., & Rist, S. (2019). Research funding programmes aiming for societal transformations: Ten key stages. *Science and Public Policy*, *46*(3), 463–478. https://doi.org/10.1093/scipol/scy074

Schneider, J. W. (2015). Null hypothesis significance tests: A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411–432. https://doi.org/10.1007/s11192-014-1251-5

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

Science-Metrix and PPMI. (2021). *Provision and Analysis of Key Indicators in Research and Innovation*. https://doi.org/10.2777/587466

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature*, *544*(7651), 411–412. https://doi.org/10.1038/544411a

Tahamtan, I., & Bornmann, L. (2020). Altmetrics and societal impact measurements: Match or mismatch? A literature review. *El Profesional de La Información*, *29*(1), e290102. https://doi.org/10.3145/epi.2020.ene.02

Technopolis Group, & Science-Metrix. (2020). *Evaluation of the Belmont Forum: Final Report*. https://www.belmontforum.org/wp-content/uploads/2021/03/Belmont-Forum-Evaluation-Report.pdf

Tijssen, R., & Kraemer-Mbula, E. (2018). Research excellence in Africa: Policies, perceptions, and performance. *Science and Public Policy*, *45*(3), 392–403. https://doi.org/10.1093/SCIPOL/SCX074

Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications, 5*(1), 1–12. https://doi.org/10.1057/s41599-019-0233-x

Vignola-Gagné, E., Pinheiro, H., & Campbell, D. (2021). *Provision and analysis of key indicators in research and innovation*. Policy brief E, *Testing the societal outcomes of R&I policies: Altmetric case study: Is cross-disciplinary research increasing the odds of research findings influencing decision-making?* Brussels: European Commission. https://op.europa.eu/en/publication-detail/-/publication/af06a3ed-2b18-11ec-bd8e-01aa75ed71a1/language-en/format-PDF/source-234777110

Waltman, L., & van Eck, N. J. (2018). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, S. U., & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 281–300). Dordrecht: Springer. Retrieved from http://arxiv.org/abs/1801.09985

Wilsdon, J., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., & Wouters, P. (2017). Next-generation metrics: Responsible metrics and evaluation for open science. In *Report of the European Commission Expert Group on Altmetrics*. Brussels: European Commission. https://doi.org/10.2777/337729

Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2019). Performance-based research funding in EU Member States: A comparative assessment. *Science and Public Policy*, *46*(1), 105–115. https://doi.org/10.1093/SCIPOL/SCY041

CGIAR Advisory Services – CAS

Via di San Domenico 1, 00153 Rome, Italy

Email: CAS-Secretariat@cgiar.org

URL: https://cas.cgiar.org/evaluation

Follow CGIAR Advisory Services on social media