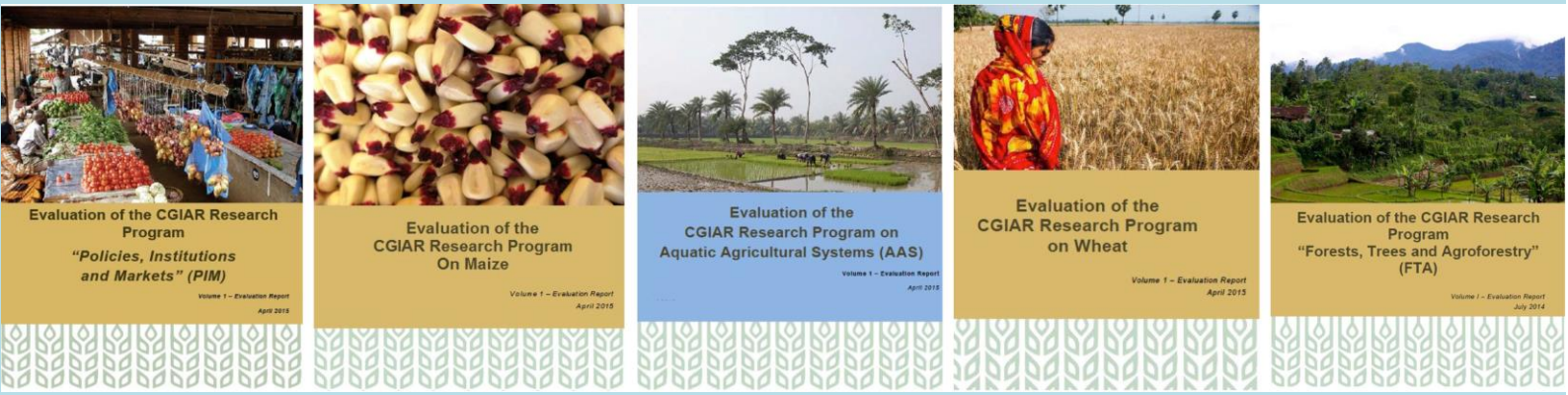


Synthesis and reflections from five CRP Evaluations

Report on a meta-analysis of five CRP evaluations

March 2016

Amended 12 May 2016

Christian H. Roth

Sophie Zimm

CGIAR | Independent Evaluation Arrangement

This Study has been commissioned by the Independent Evaluation Arrangement (IEA) of the CGIAR.

Correct citation: Roth, C. and Zimm, S. (2016): Synthesis and reflections from five CRP evaluations - Report on a meta-analysis of five CRP evaluations. Rome, Italy: Independent Evaluation Arrangement (IEA) of the CGIAR.

http://iea.cgiar.org/

Table of Contents

Tables

Figures

About the Authors:

Christian Roth has over 30 years of research experience in tropical land and water management. He is currently working with CSIRO's Land and Water Flagship based in Brisbane, Australia. Over the last twelve years, his main focus has been designing, commissioning and conducting inter- and transdisciplinary research for development programs and projects in Australia, South and South-East Asia in water resource management, climate change adaptation, smallholder farming systems and conservation agriculture. He has also led or participated in a range of research project and program evaluations in Australia and Asia. His main research interests reside in integration of social sciences and biophysical research to influence the research for development agenda in South and South-East Asia, specifically in the domains of climate adaptation and agricultural development. He has published his research extensively in about 180 publications and research reports. Christian has a PhD in soil hydrology from the University of Göttingen.

Sophie Zimm works as an Evaluation Analyst for the IEA in Rome, Italy. She has been supporting data processing and analysis in several CRP evaluations, including quantitative analysis of indicators relating to quality of science. She also worked very closely with the Team Leader on all aspects of the Review of the CRP Governance and Management. Before her job at IEA, she gained several years of experience in evaluation units of other international organizations. Among other projects, she coordinated the data collection for the impact evaluation of a large World Bank education programme as well as the implementation of a private sector development project (UNIDO) in Mozambique. Sophie has a MA in International Business from the Vienna University of Economics and Business.

Abbreviations

AAS	Aquatic Agricultural Systems (CRP)
ARI	Advanced research institute
BoT	Board of trustees
CIMMYT	Centere Internacional de Mejoramento de Maiz y Trigo
CGIAR	Consultative Group for International Agricultural Research
CIFOR	Center for International Forestry Research
CRP	CGIAR Research Program
FP	Flagship project
FTA	Forests, Trees and Agroforestry (CRP)
ICARDA	International Center for Research in the Dry Areas
ICRAF	World Agroforestry Center
IDO	Interim Development Outcomes
IEA	Independent Evaluation Arrangement
IFPRI	International Food and Policy Research Institute
IP	Impact pathway
M&E	Monitoring and evaluation
ME&L	Monitoring, evaluation and learning
NARS	National agricultural research system
NGO	Non-government organisation
PAR	Participatory action research
PIM	Policies, Institutions and Markets (CRP)
QoS	Quality of science
RinD	Research in Development
SLO	System Level Outcomes
ToC	Theory of change

Executive Summary

This study was commissioned by the Independent Evaluation Arrangement to provide a synthesis of the first five completed evaluations of CGIAR Research Programs (CRPs). It focusses on the results of these evaluations against the main evaluation criteria, with emphasis on relevance, quality of science, effectiveness and results. The five CRPs forming the basis of this synthesis are AAS - Aquatic Agricultural Systems; FTA - Forests, Trees and Agroforestry; MAIZE; PIM - Policies, Institutions and Markets; and WHEAT.

The synthesis framework was organized around key dimensions of the evaluation criteria to extract findings from the evaluation reports, with a particular emphasis on quality of science. All five evaluations placed a high emphasis on assessing program relevance. Quality of science was given the greatest relative emphasis in AAS, MAIZE and WHEAT. Varying levels of emphasis of evaluations mean that synthesis results in this report do not always relate back to all five CRP evaluations.

All five evaluations conclude that either the Programs or their research agenda are highly relevant, but for different reasons: CRPs are i) addressing key global commodities and their trajectories, ii) bringing together unique expertise to address CGIAR system level issues, iii) targeting relevant geographies, or iv) addressing new thematic domains relevant to the CGIAR.

Program level portfolios of work tend to be coherently aligned with CGIAR priorities. However, the theories of change and impact pathways were seen by all evaluations to be in need of much more refinement. There was a general consensus across evaluations that the CRPs have significant comparative advantage, albeit to different levels, for different reasons and not always evenly spread across Program portfolios. All evaluations were of the view that the calibre of senior scientists is generally good or excellent and all five CRPs were seen to have a good cadre of such highly productive senior scientists. Research design is also a key aspect of quality of research input and arguably it is a critical determinant of the quality of science, yet the one that yielded mixed results across the evaluations. This is in part because research design is the most complex aspect of quality of science to assess, and hence it was treated quite differently across the evaluations.

The types of outputs generated by the five CRPs range markedly, including scientific publications such as papers, books/ chapters and research reports; technical manuals on methods and protocols; new production technologies; germplasm; web-accessible databases; and policy briefs and blogs. All CRPs produce scientific journal publications, but in addition, some also produce a wider range of non-journal publications such as policy briefs and other communication products depending on intended audience (FTA, PIM), models (PIM), as well as germplasm (WHEAT and MAIZE).

Production of high quality journal papers is in some CRPs reliant on a smaller number of very high performing senior scientists. The wide ratio between senior scientists and more junior scientists in some Programs also presents a similar problem, in that constraints in mentoring younger scientists also curtails publication output and science quality and impinges on the quality of science thinking.

A common observation by all evaluations was that the bulk of the publications output in each CRP relates to legacy work brought into these programs. This is not surprising, considering the time point at which Programs were evaluated and the time it takes for completed research to be published in scientific journals. Hence it is still hard to discern any significant signatures of the Programs with regard to the nature and quality of scientific journal publications.

Irrespective of whether outputs are context-specific and targeted or constitute international public goods, all CRPs have achieved many planned outcomes, some of which were deemed very significant. However, in most cases these outcomes were achieved by building on or taking legacy work to conclusion, although there are some examples of outcomes (e.g. value chain improvements conducted by PIM) that can be directly related to work initiated in the CRPs.

In general the evaluations conclude that the evidence base for development outcomes is still weak and adoption studies and impact assessments need to be strengthened, ideally being linked to the CRP's theories of change and impact pathways.

Conclusions were also drawn from the three crosscutting criteria. The nature of the partnerships is clearly in transition and varies, with some CRPs actively broadening the partnerships to include a wider range of boundary partners in policy and the development sector relevant to achieving outcomes. Awareness and acceptance of the need to include gender dimensions into research and attempts at establishing a more effective gender mainstreaming approach is evident across all five CRPs. PIM and AAS were judged to be at the forefront and insights from both could provide valuable lesson for other CRPs.

In terms of management performance, a very consistent pattern observable across all five evaluations is the negative impact that the unpredictability of W1/W2 funding at the Consortium Office level is having on Programs and partners. This significantly impinges on CRPs' ability to strategically utilise W1/W2 funds. Reporting systems were generally seen as weak and in need of strengthening.

The **main lessons** derived from this synthesis are:

- *Relevance and effectiveness* – strengthening theories of change and tightening the articulation of impact pathways;
- *Quality of science and research design* – evolving frameworks to better capture systems science dimensions and modalities where integrative science is necessary;
- *Partnerships* – moving towards more strategic and effective partnerships;
- *Gender* – transcending the present (narrow) focus on gender mainstreaming to include dimensions of social equity and inclusion;

  Monitoring, evaluation and learning – promoting a stronger learning culture and strengthening researcher capacity to engage in reflexive processes.

1. Purpose and scope of the synthesis

In the CGIAR agricultural research for development is implemented by 15 research Centers and their partners through CGIAR Research Programs (CRP; for simplicity, in the remainder of this report denoted as Programs). Ultimately, the Programs are aimed to contribute to the system level objectives of CGIAR: i) reduced rural poverty; ii) improved food and nutrition security for health; and iii) improved natural resources systems and ecosystems services. The Independent Evaluation Arrangement (IEA) of the CGIAR is responsible for external independent evaluations of the Programs and for developing a coordinated and harmonized evaluation system within the CGIAR.

IEA commissioned this synthesis study of the first five completed evaluations of the CRPs listed below. For a synopsis of each CRP please refer to Annex 1.

- AAS – Aquatic Agricultural Systems
- FTA – Forests, Trees and Agroforestry
- MAIZE
- PIM – Policies, Institutions and Markets
- WHEAT

The synthesis study focuses on findings and conclusions, to highlight patterns across the evaluations and to draw lessons learned where possible for the System as a whole. This study focusses on the results of these evaluations against the main evaluation criteria, with emphasis on relevance, quality of science and results as well as performance in the areas of gender, partnerships and capacity building. The synthesis report is the first consolidated input to inform the second cycle of CRPs starting in 2017, and also provides inputs to the synthesis of the full set of CRP evaluations, which is to take place in 2016.

2. Synthesis Methodology

The development of the study methodological framework was conducted in consultation with the IEA in September 2015. This framework (see Annex 2), which was organized around key dimensions of the agreed evaluation criteria[1] used by the evaluations was then used to extract findings from the evaluation reports, with a particular emphasis on quality of science. It also determines the structure for this report. This was followed by a detailed document review of the five evaluation's terms of reference, inception reports, final reports and management responses to the evaluation reports during September to November 2015, mapping key findings from these documents against the evaluation criteria.

A preliminary synthesis of the main findings (including emerging issues) of the five evaluations was iteratively developed with IEA in early November 2015. This was followed by a presentation of key findings in relation to quality of science to the IEA Workshop on Quality of Science in December 2015[2], which also drew on a more detailed cross analysis of bibliometric and survey results conducted by the five evaluations. Feedback from IEA arising from the above workshops was incorporated into the draft synthesis report before the report was finalized in February 2016.

The five Program evaluations synthesized here represent the first suite of Program evaluations conducted by the IEA, and although there was some degree of fine-tuning of the evaluation approaches by the IEA in the course of the evaluations, they all conformed to the IEA standards and guidelines. All five evaluation reports generally were structured around the evaluation criteria, with the exception of AAS, which followed a slightly different format. There was some variation in depth and methodology across evaluations because evaluations took place at different stages of the Programs (e.g. pre or post extension phase) and due to differing additional evaluation questions, resulting in varying emphasis with which each evaluation criterion was addressed (see Table 1).

All five evaluations placed a high emphasis on assessing Program relevance and with the exception of FTA, also addressed quality of science in depth. Conversely, FTA placed a high emphasis on organizational performance and much less on quality of science - a bibliographic analysis of scientific journal publications was not carried out in this case. Quality of science was given the greatest relative emphasis in AAS, MAIZE and WHEAT. These varying levels of emphasis mean that synthesis results in this report do not always relate back to all Program evaluations, but may only reflect some evaluations.

[1] These evaluation criteria are set out by IEA, and are described in http://iea.cgiar.org/sites/default/files/Standards.pdf

[2] The summary report of the workshop can be found under: http://iea.cgiar.org/news/evaluating-quality-science-cgiar

Table 1. Relative emphasis (%)[1] of the five evaluation reports against evaluation criteria. Bold font denotes top three criteria emphasized.

Evaluation criteria	AAS[2]	FTA	MAIZE	PIM	WHEAT
Relevance	**19**	**23**	**16**	**18**	**18**
Quality of science	**23**	9	**31**	**18**	**26**
Effectiveness	14	**21**	**15**	12	**18**
Impact / value add	9	3	9	7	11
Cross-cutting: gender	5	2	4	12	6
Cross-cutting: partnerships	5	2	6	8	5
Cross-cutting: capacity building	3	1	7	4	6
Organizational performance	**22**	**38**	10	**21**	10

[1] Derived by page count of each evaluation criterion as a percentage of total report page count

[2] Structure of AAS did not as stringently follow the evaluation criteria as did the other evaluations, so assigning of emphasis against evaluation criteria is more approximate

Other factors which influence the ability to extract general patterns across Programs, or enable the clear identification of causes for differences across Programs comprise:

- A lack of a more detailed regional/geographic breakdown of evaluation results in cases where programs were geographically structured
- Evaluations did not consistently assess monitoring and evaluation (M&E) procedures or frameworks; hence it is still difficult to ascertain how well Programs are on track to achieving their stated outcomes.

3. Main findings against evaluation criteria

3.1. Relevance

3.1.1. *Alignment with CGIAR, global and national priorities*

At a general level, all Programs were found to be aligned to CGIAR priorities, based on a set of relevant Intermediate Development Outcomes (IDO) contributing to the System Level Outcomes (SLO). However, the SLO are articulated at such a high level, that invariably all Programs can argue relevance. The actual articulation of how the Programs will achieve their contributions to SLOs through their respective pathways to impact underpinning their IDOs varies, as does the depth of analysis of impact pathways conducted by the evaluations. Partly this is a result of the timing of the evaluation (e.g. in some cases, Programs had already developed more explicit theories of change (ToC) and impact pathways (IP) in their extension proposals), partly this depends on the methodological emphasis of evaluations. As a result, linking pathways to impact back to Program IDOs on the basis of the evaluation results is not always straightforward, and most evaluations noted that theories of change (underpinning the impact pathways) needed to be made more explicit and specify more clearly the underlying assumptions and the conditions necessary to support impact and how they contribute to the IDOs, as well as being more strategic and targeted about choice of boundary partners.

Nonetheless, all five evaluations conclude either that the Programs (FTA, MAIZE, PIM, WHEAT) or their research mandate (AAS) are highly relevant, but for different reasons:

- *Programs are addressing key global commodities and their trajectories* – the WHEAT and MAIZE evaluations argue relevance primarily on the basis of the global significance of these commodities for food security. While the relevance of MAIZE is also predicated on the importance of maize as a global staple, the MAIZE evaluation suggests greater attention needs to be given in a future pathway to impact analysis on the dynamics of maize markets and differentiation of the Program from developments in the private sector.
- *Programs are bringing together unique expertise to address CGIAR system level issues of poverty alleviation, food security and sustainable use of resources* – while this is the case for all Programs evaluated, this was highlighted in particular for PIM (e.g. informing policy).
- *Programs are targeting relevant geographies* – this received less attention in WHEAT and MAIZE, as their geographies are defined by where wheat and maize are primarily produced. The regional aspect of relevance was more explicitly evaluated for FTA and AAS. In the former, while generally targeting relevant regions, the evaluation pointed out a bias towards Southeast Asia and humid forest environments. In the case of AAS, the issue was more of how location-specific work (in the right environments) could be usefully generalized beyond Program hubs.
- *Programs are addressing new thematic domains relevant to the CGIAR* – this was highlighted in the AAS evaluation. Its framing around the new concept of aquatic agricultural systems was seen as highly relevant (in fact innovative for the CGIAR), but with no such clear conclusion about the AAS as a Program being relevant.

Some evaluations indicated that relevance of the Programs was diminished because of weaknesses in the research design. Broadly, there are three types of factors leading to reduced relevance:

- Programs that have implemented case study or hub approaches are experiencing difficulties in generalising (scaling out and up) place-based research and subsequently generating IPGs. Programs that were seen to have this problem were FTA and AAS.
- Some evaluations also noted that some of the work was too theoretical and high level to be relevant (FTA), or the balance of research too skewed towards upstream, discovery-type research as opposed to downstream, more delivery oriented research (PIM).
- Donor driven demand (through W3 and bilateral funding) does not necessarily equate to developing country or target community priorities and hence loses national or local relevance. This was observed in the PIM and AAS evaluations.

3.1.2. *Program coherence and use of W1/W2 funds*

Program level portfolios of work tend to be coherently aligned with CGIAR priorities and the SLOs. Coherence does not seem to depend so much on the number of CGIAR Centers involved in a Program but the extent to which Program management has an ability to influence Program design, dependent on the degree of independence and level of empowerment of Program management, the size and use of the Window 1/Window 2 (W1/W2) envelope[3] (which ranges from about 20% - MAIZE, to 50% - AAS), and the theory of change (ToC) and impact pathway (IP) framework being co-developed and shared amongst the partners within a Program. The ToC and IPs were seen by all evaluations to be in need of much more refinement, which was observed to have partially commenced in the extension proposals.

For example, in the case of the WHEAT evaluation, the analysis of pathways to impact led to a view by the evaluation that coherence and congruency with WHEAT and CGIAR objectives could improve through refinement of program strategies and better alignment, prioritization and sequencing of outputs. Similarly to WHEAT, the MAIZE evaluation argues that coherence in MAIZE could be further strengthened by improving the impact pathways for flagship projects, particularly concerning their inter-linkages and the assumptions that relate to the doubling of productivity in the target regions.

Generally, coherence of Programs was strengthened in the extension proposals by building on better defined pathways to impact and by merging previous themes or flagships into larger, more interlinked clusters of work within FP (except for AAS, where evaluation did not consider the extension proposal).

In relation to using W1/W2 funds to help build Program coherence, a consistent issue emerging from all evaluations was that of unpredictability (delays) of W1/W2 funding (see also 3.5.3). Delays were seen to occur at several points, but mainly comprised the transfer from the CGIAR Fund to the lead CGIAR Centers, and the transfer of funds from lead centers to Programs. This was seen to significantly impinge on Programs' ability to strategically utilize W1/W2 funds.

Programs employed a range of strategies through which they utilized and allocated W1/W2 funds:

- Topping-up of W3 and bilateral projects (e.g. to enhance Program coherence; to achieve full recovery of Program research costs; to cover off on CGIAR Center infrastructure costs; to leverage new W3/bilateral projects)

[3] For more information on CGIAR funding mechanisms and the different funding windows, please see: http://www.cgiar.org/who-we-are/cgiar-fund/

- Funding gender mainstreaming (but making gender work potentially vulnerable to budget cuts)
- Paying for Program management
- Supporting grant schemes to involve external partners (e.g. universities, NARS partners)
- Funding strategic work otherwise not funded (e.g. Sentinel Sites – FTA; Global Science Scaling hubs – AAS; 'discovery' research – PIM).

Across all five evaluations, there was also a view that in the pursuit of W3 and bilateral funding, donor agendas tend to drive Program priorities more than CGIAR or partner country priorities. This in some cases works against a more coherent research program, and detracts from the strategic use of W1/W2 funds in those instances where W3 and bilateral funds do not fully cover infrastructure costs, or donor priorities are not aligned with Program priorities. This tension between strategic programming and opportunistic responses to donors is likely to be exacerbated in the context of current funding cutbacks.

3.1.3. *Comparative advantage*

Commensurate with the findings in relation to alignment with CGIAR, global and national priorities, there was a general consensus across evaluations that Programs have significant comparative advantage, albeit to different levels, for different reasons and not always evenly spread across Program portfolios:

- *Regarded as neutral world class scientific research organisations in their domains ('trusted advisor')* – this was generally stated for all five Programs, but was particularly highlighted for FTA and PIM, where the assessments were underpinned by views of external stakeholders (e.g. FTA, PIM).
- *Programs hold in trust and have access to unique genetic resources for major global commodities* – this was stated strongly for MAIZE and WHEAT, making this the primary comparative advantage for these two Programs. However, while the MAIZE and WHEAT evaluations conclude that these Programs have a strong comparative advantage because of the unique genetic resources at their disposal, their excellent breeding research facilities and their considerable breeding capacity, MAIZE needed to reassess its comparative advantage in relation to the rapidly evolving private sector providers, particularly in relation to commercial breeding of hybrids.
- *Programs can muster larger and more diverse research networks, enabling more comprehensive data collection* – this was highlighted in the PIM evaluation, where reference was made to significant global datasets established by PIM (though largely reflecting IFPRI's role). This capability positions Programs such as PIM to conduct global analyses. In Programs with fewer participating Centers (AAS, MAIZE, WHEAT), the comparative advantage of being able to establish more diverse and wider ranging networks is less pronounced.
- *Programs are better positioned to produce a wide range of IPG* – this is more the case in MAIZE, PIM and WHEAT, where there have either been stronger synergies (combining wheat and maize germplasm expertise and collections), or broader integration and complementarities with multiple Centres (e.g. value chain research; PIM). Conversely, in AAS and FTA, where the evaluations pointed to weaknesses in the ability to scale up from case study or place-based research, full capitalisation on the potential comparative advantage to generate IPG has not yet been achieved.
- *Programs have assembled significant intellectual critical mass in their respective domains* – this was recognised for all five Programs, but this assessment was qualified in most evaluations in that critical mass is uneven, with some domains showing equal comparative advantage vis-a-vis other large international research organisations (especially in relation to quality of research), or weaknesses in specific domains (e.g. scaling).

3.2. Quality of science

IEA uses a framework looking at various aspects and including a number of indicators to support the assessment of quality of science. These are listed in Table 2, together with a qualitative assessment of how the different evaluations addressed these indicators, showing that there was some variation. This section focuses on those aspects that received the greatest attention in the evaluations.

Table 2: A framework to define the need for systems science – building on the nexus between research practices and impact considerations

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

Source: Authors. () denotes indicators covered under other evaluation criteria (e.g. effectiveness - FTA), or where there is overlap between QoS indicators (MAIZE)

3.2.1. *Quality of inputs*

All evaluations were of the view that the calibre of senior scientists is generally good or excellent and all Programs were seen to have a good cadre of such highly productive senior scientists. This is generally backed by the results of the h index analysis conducted for four of the Programs (Table 3). None of the evaluations formally benchmarked these measures against comparative organizations outside the CGIAR system, but were still of the view that leading Program scientists had benchmark or higher h indices (>20) and that there was evidence of scientists in all Programs publishing in reputational journals and also in a few instances, in some of the eminent journals such as Science (AAS) and Nature (WHEAT) (see also Table 4). The WHEAT evaluation explicitly mentions the quality of national agricultural research systems (NARS) scientists collaborating with CGIAR scientists together being among the best available experts in each region of interest; the FTA evaluation also confirms involvement of strong non-CGIAR research partners. Conducting an assessment on the basis of the h index alone does however not necessarily appropriately cover the 'thought leader' dimension of lead scientists, and hence all evaluations to varying degrees complemented this with a qualitative assessment of scientists.

Table 3: Summary results of h index analysis of lead Program scientists (using Scopus)

	AAS	FTA	MAIZE	PIM	WHEAT
Role of researchers selected		FTA leadership group	Researchers with a supervisory role	Leaders of W1/W2 activities	Researchers with a supervisory role
N		25	38	36	45
H index <0	n/a	4%	8%	15%	7%
H index 1 to 5		4%	11%	**46%**	22%
H index 6 to 10		24%	32%	24%	22%
H index 11 to 20		**40%**	**42%**	13%	**33%**
H index >21		28%	8%	3%	16%

Source: Authors, based on bibliometric analysis in CRP evaluations.

In some evaluations it was argued that the ratio of senior scientists to early career scientists was too small. In AAS, the allocation of PhD level staff against each hub and theme was viewed as being too limited and with less than optimal deployment of experienced scientists across the program. The PIM evaluation noted that there was a strongly skewed distribution of scientists with a high performance; 24 out of 136 principal scientists producing 5 or more publications in 2013-2014, versus nearly 50% of all PIM researchers not having published a single publication in the same period. In MAIZE (and to some extent in WHEAT), there has been a rapid expansion of new staff particularly within CIMMYT's Conservation Agriculture program, with 60% of the staff deployed having been with CIMMYT for less than 2 years. In AAS 52% of staff responding to the survey had only joined since 2011.

The evaluations also commented on the skills mix of Programs. In the case of FTA it was observed that there was a lack of key skills in economics, financing and policy analysis. The PIM and MAIZE evaluations highlighted the high proportion of social scientists[4], noting that IFPRI traditionally has had a strong set of social scientists, while the integration of social scientists in MAIZE and WHEAT still requires more effort.

Related to this observation are statements that systems science is still underdeveloped, particularly in some aspects of WHEAT and AAS, where critical mass was not high enough to support systems science. Partly this is due to insufficient integration of disciplines and weak linkages with other Centers beyond those in Program in question. AAS is 90% dominated by Worldfish with weak linkages to other Centers, whereas WHEAT has similar skills originating from CIMMYT and ICARDA, but complementing geographies. Use of integration tools like modelling is also weaker in these CRPs. This contrasts with PIM and to a lesser extent FTA, where a wider range of skills from more complementary CGIAR Centers (PIM – 13 Centers) is inherently more supportive of systems approaches and integration.

Research design is also a key aspect of quality of input and arguably it is a critical determinant of the QoS, yet the one that yielded mixed results across the evaluations. This is in part because research design is the most complex aspect of QoS to assess, and hence it was treated quite differently across the evaluations:

[4] Although it would seem the term 'social' scientist in the case of PIM was used in a broader sense, including economists and socio-economists in addition to pure social scientists

- AAS the only evaluation that tried to unpack research design and develop a systematic approach at assessing it
- In FTA, research design was largely subsumed in the discussion of effectiveness
- In the other evaluations, treatment of research design lies between AAS and FTA

The variation in how research design was approached is probably also due to the fact that this aspect cuts across several other evaluation criteria, i.e. relevance; effectiveness; partners; organizational management. Moreover, research design also intersects with multiple QoS dimensions: science leadership and quality of thinking; facilities; processes; research partnerships. Presumably because of the lack of common approach to assess research design that can apply to various research dimensions and modalities of the CGIAR, research design as a critical aspect might not be clear enough to support greater consistency in evaluations.

Good data management was highlighted in the cases of PIM and WHEAT, primarily based on the staff survey results. The WHEAT and MAIZE evaluations also placed a particular focus on research infrastructure supporting advanced breeding methods, commenting that breeding methods employed in WHEAT were more traditional, but that shuttle breeding across locations has been successful and is being extended. In the case of MAIZE there is also scope for deployment of modern breeding methods based on molecular technologies, but that will have to be complemented by improvements in data sharing and analysis, as well as strengthening plant-based physiology and modelling to increase high throughput screening for traits. The need for improvements to data management and to infrastructure was also raised by AAS scientists.

3.2.2. *Science quality assurance and research management*

Several dimensions of quality assurance were discussed in the evaluations. A culture of acceptance for innovative ideas and taking risk (FTA, WHEAT) and acceptance and encouragement for learning from failures (AAS) are examples of Programs having in place a culture that stimulates scientific thinking and questioning. However, adaptive cycles of learning were less evident (see section 5.2).

Peer review processes seem to be more controlled by Centers where they exist than by Programs. Respondents to the staff survey in PIM pointed to the stronger quality assurance processes in IFPRI compared to other Centers in PIM. In the case of AAS there was dissatisfaction with internal feedback (perhaps reflecting the too thinly spread senior scientists – see 3.2.1); given that the majority of respondents are from WorldFish, this also constitutes more of a CGIAR Center signal. Related to feedback on quality assurance are repeated statements by staff about their dissatisfaction with insufficient incentives, but the evaluations did not provide enough detail to ascertain what form of incentives was being referred to.

Evaluation results against these aspects of QoS draw mainly on the results of the staff survey, with little triangulation by additional evaluation methods. MAIZE as an example where there is a discrepancy between the survey and the evaluation team's assessment in relation to data management. Ambiguity between roles of Centers and CRPs in managing quality assurance is perhaps a reason this aspect being not comprehensively covered in evaluations.

3.2.3. *Quality of outputs*

The types of outputs generated by the five Programs evaluated range markedly, including scientific publications such as papers, books/ chapters and research reports; technical manuals on methods and protocols; new production technologies; germplasm; web-accessible databases; and policy briefs and blogs.

In the following we focus on the quality of scientific publications, while the wider range of outputs are discussed in section 3.3.1.

All CRPs produce scientific journal publications, but in addition, some also produce a wider range of non-journal publications such as policy briefs and other communication products depending on intended audience (FTA, PIM), models (PIM), as well as germplasm (WHEAT and MAIZE).

Four evaluations conducted a bibliographic assessment of a subset of selected publications; a summary of results is presented in Table 4. Of note is the wide range of total outputs, with a maximum of 1400/700 in the case of FTA, and 262/238 for MAIZE. Partly this is due to differences in period evaluated and the way in which Programs manage their databases, but there may also have been errors in reporting or attribution of papers to a particular Program. The FTA evaluation specifically noted problems in reporting and attribution of results to the Program.

To keep the analysis of citation frequency comparable, the period of assessment was held the same (2012 – to mid-2014). Patterns across AAS, MAIZE and WHEAT were quite similar, with about 13 – 15% of papers having received citations above 20. AAS and WHEAT also produced papers with a citation above 30. The performance of PIM on these measures would appear to be lower than the other Programs. However one needs to bear in mind that impact factors and citations in social science and economics journals tend to be lower than those of biophysical journals/papers and cannot be compared in absolute terms.

Table 4: Summary results of bibliographic analysis conducted on a subset of Program publications

Criterion	Unit	AAS	FTA	MAIZE	PIM	WHEAT
Total publications	Period	'09 - mid '14	'11 - '13	'12 - mid '14	'12 mid '14	'12 mid '14
	No	599	1400	262	370	333
Scientific journal papers	No	214	700	238	167	291
	%	36	50	91	45	87
Impact factors	N	27	n/a	166	173	180
	Most frequent journal	2.009	n/a	2.474	1.085	3.658
	Highest impact journal	31.027	n/a	11.808	n/a	38.597
Citation analysis (2012 and 2013)	N	27	n/a	166	173	180
	0	19%		14%	29%	12%
	1 to 10	52%		57%	57%	62%
	11 to 20	15%		12%	9%	14%
	21-30	11%		11%	4%	8%
	Above 30	4%		5%	1%	3%

Source: Authors, based on bibliometric analysis in CRP evaluations.

Nonetheless it is noteworthy that PIM did have the lowest output in journal papers. As noted in 3.2.1, this is in part due to a large proportion of researchers in PIM not having published during the evaluation period. Partly this is also because there is a greater emphasis on other publications and communication products (e.g. policy briefs) given the greater linkage of PIM to policy makers. The high relevance and quality of policy briefs emanating from PIM was lauded in the PIM evaluation, but in this case it was also noted that researchers were uncertain about PIM management expectations regarding publications.

3.2.4. *Constraints to quality of science*

Production of high quality journal papers is in some Programs reliant on a smaller number of very high performing senior scientists. This was explicitly noted for PIM (backed by metrics; see 3.2.1), but the wide ratio between senior scientists and more junior scientists (often nationally recruited and some without a PhD), was discussed mainly in the AAS and MAIZE evaluations. These evaluations also highlight constraints in mentoring younger scientists which ultimately also curtails publication output and science quality and impinges on the quality of science thinking. In some evaluations there was also commentary (based on staff survey responses) that senior scientists often are busier with the additional transactions associated with engaging across the CGIAR Center versus Program matrix, further diminishing their ability to act as mentors and thought leaders. The AAS evaluation goes further to note that insufficient presence by senior researchers in the regional hubs to engage with communities is compromising the ability of the PAR/RinD paradigm under which AAS is operating to translate into scalable outcomes, as well as ensuring the appropriateness and quality of the research designs.

None of the Programs were reported to have formal science quality assurance mechanisms or strategies in place to enhance quality and volume of science output; rather, one can infer that responsibility for quality of science still lies more with the Centers, where control over appointments sits, and procedures and incentives vary.

Other constraints noted by the evaluations but that relate more to individual Programs comprise:

- Need for improved staff training where there has been an influx of new more junior scientists (MAIZE)
- Extending learnings across projects within Program (WHEAT)
- Increase investment in improved data management (MAIZE, WHEAT)
- Support more synthesis and systems approaches (AAS, WHEAT)
- Increase use of analytical and quantitative methods (AAS)
- Broaden use of models as a key tool for integration, facilitating inter-disciplinarity and supporting system science (AAS, MAIZE, WHEAT).

3.3. Effectiveness and results

3.3.1. *Key outputs*

As already mentioned in section 3.2.3, all Programs were seen to produce a wide range of outputs, with scientific publications considered as a key output common across all five Programs. While all five evaluations qualitatively assessed outputs[5] in depth, an actual breakdown of the relative importance of different kinds of outputs produced across each Program could not be provided; nor was there a systematic analysis to how well Programs had delivered on stated output objectives and targets (with the exception of FTA). In part this may be due to the difficulty in easily extracting the related information from the Programs' output databases.

A common observation by all evaluations was that the bulk of the publications output in each Program relates to legacy work brought into the Programs. This is not surprising, considering the time point at which Programs were evaluated and the time it takes for completed research to be published in scientific journals. Hence it still is hard to discern any significant signatures of the Programs with regard to the nature and quality of

[5] Scientific publications were also assessed quantitatively, see section 3.2.3.

scientific journal publications. Non–journal publications (e.g. policy briefs, blogs, fact sheets, reports) are more frequently associated with results from Program initiated research. These outputs are important in terms of influencing boundary partners and achieving desired outcomes, and in general the evaluations recommend strengthening or retaining a strong focus on such targeted outputs.

In addition to production of new knowledge through publications, the development of new germplasm constitutes a major high quality output from WHEAT and MAIZE; this was a strong focus in these evaluations. WHEAT has produced an impressive 200,000 novel lines per year for all types of wheat, with the best performing of these lines being released to the International Wheat Improvement network as IPG. MAIZE has produced a molecular atlas of more than 30,000 maize accessions and breeding germplasm with more than 2,200 genotypes, as well as increasing the output of varieties (mainly hybrids) from 90/year pre-CRP to 160/year presently. Coupled to the production of new germplasm is a vast and robust body of work on innovations in molecular and breeding methods, which is captured in high impact journal publications.

The production of databases was also recognised by some of the evaluations as a key output. This relates both to databases with regard to germplasm (MAIZE, WHEAT), as well as relating to compilation and provision of access to policy relevant statistics (e.g. gender statistics in PIM) or ecological, economic and social characterisations of the state and transition of representative study sites (e.g. Sentinel Landscapes in FTA).

New technologies that lead to enhanced productivity, increased sustainability or improved livelihoods also constitute a key set of outputs, and relate primarily to AAS, MAIZE and WHEAT. The AAS evaluation was more focussed on how the technologies might be transferred, assessing the effectiveness of the AAS participatory action research (PAR) approach and discussing less the technological innovations *per se*. The MAIZE and WHEAT evaluations provide more commentary on the nature of the technologies and their relevance (e.g. Conservation Agriculture techniques), while also reflecting on how well these Programs were linking outputs to next and end users.

3.3.2. *Potential for outcomes and production of IPGs*

In general terms, one can broadly distinguish between two pathways by which outputs lead to outcomes:

1. Mostly technical (practices) or germplasm (varieties) related outputs that lead primarily to *local* development outcomes, mainly involving local boundary partners (government, private sector, NGOs) or next users (farming households). In aggregate these outcomes contribute towards the Program development outcome targets.
2. Outputs such as knowledge products (e.g. journal publications, technologies, genes, databases) that have much *wider (global)* relevance and applicability, provided they are readily accessible to everyone. These outputs constitute International Public Goods (IPGs). The ability of Programs to generate IPGs constitutes their purported comparative advantage, but depends not only on the production of such outputs, but having in place delivery mechanisms to relevant next users (which intersects with effectiveness and partnerships).

The boundary between these two categories is fuzzy, and the implication is that not all outputs are IPGs. However, the extent to which Programs attempt to transfer place or context specific outputs into more generic IPGs is a pertinent question, as is the balance between the two. AAS was the only evaluation that reflected more deeply on how well or not the Program was positioned to deliver on IPGs. In the case of MAIZE, the extent to which outputs constitute IPGs was primarily looked at through the lens of comparative

advantage[6]. At the other end of the spectrum, the PIM evaluation did not explicitly take into consideration the IPG dimension of outputs, but did evaluate outputs more generally. The other two evaluations conducted an intermediate level of assessment of IPGs.

Irrespective of the above distinction between context-specific/targeted outputs and IPGs, all evaluations concluded that Programs have achieved many planned outcomes. Some of these outcomes were deemed very significant, e.g. PIM's Country Strategy Support Programs in Bangladesh and Ethiopia. However, in most cases these outcomes were achieved by building on or taking legacy work to conclusion, e.g. uptake of Conservation Agriculture practices in South Asia (WHEAT) or the adoption of new maize varieties (MAIZE).

There are some examples of outcomes (e.g. value chain improvements conducted by PIM) that can be directly related to work initiated in the Programs, despite the short duration of most of the Programs at the time of evaluation. Mostly there was a view that Programs were positioned to capitalise on the work initiated during the Programs' life time, in particular as they made their impact pathways more explicit and operationalised these through a more strategic broadening of partnerships, but that it was unrealistic to expect major outcomes arising out of work initiated under the Programs. The nature of the partnerships evolving in the Programs' development of the extension phase was deemed as providing increased prospects for achieving planned outcomes and impacts in the future. This assessment however was qualified in the case of the AAS and FTA evaluations, where weaknesses in the research design to support scaling of site-based research outputs to broader application domains hindered the ability of these Programs to generate IPGs (see also section 3.2.4.).

3.3.3. *Validity of theories of change*

Four of the evaluations carried out some level of description or assessment of the Program ToC; the PIM evaluation was the exception. The depth and nature of the analysis was in part determined by the timing of the evaluation in relation to whether Programs had started refining their ToC as part of the preparation of their extension proposals. In part there was also divergence in how evaluations interpreted the role of ToC in Program effectiveness. In the case of WHEAT (and to a lesser extent, MAIZE), the ToC was assessed in the context of impact evaluation, rather than effectiveness. In contrast, FTA undertook a more detailed analysis in the context of relevance and effectiveness, providing assessments and recommendations on how FTA needs to proceed with strengthening its ToC. While AAS also described and reflected on the Program ToC in some detail, in this case the evaluation team was critical of the ToC approach in general, based on the view that this approach is too simplistic and not cognizant of an existing body of theory regarding adoption.

Despite these divergences in evaluations' assessments of Program ToC, at a general level it can be deduced that Program, Flagship Project and project level ToC and impact pathways are not yet being systematically used and developed by Programs in a way to clearly articulate underlying assumptions, or not yet providing a clear line of sight to how outputs will aggregate to targets and IDOs, and that they provide insufficient clarity about which and how boundary partners need to be engaged to secure the desired outcomes. In the case of WHEAT this was expressed as a need to incorporate detailed project/program impact pathways or ToC in each adoption study and impact assessment to increase their robustness and transparency. The FTA evaluation gave a much more nuanced assessment: at the time of evaluation the impact pathways and the results framework had not yet been integrated with each other and were not yet further developed into a

[6] MAIZE and WHEAT produce germplasm that is considered to be the most typical IPG. So at least implicitly the evaluation looked at IPGs although they might not have come to explicitly call it as such.

full FTA ToC. The ToC was seen to be skewed towards conceptualizing highly aggregated outcomes and impacts far beyond the direct influence of the Program, at the expense of focusing on establishing the basis for a results-based management framework within FTA's sphere of influence. Impact pathways and the results framework were unrelated to each other, and somehow detached from implementation practice in FTA's research portfolio, providing a clear rationale for the need to further develop FTA's ToC.

3.3.4. *Evidence base*

In general the evaluations conclude that the evidence base for development outcomes is still weak. For instance in the case of AAS, the evaluation was of the view that claims on the extent of new technologies reaching landless and poor households in Bangladesh were not substantiated by any rigorous impact assessment. Other evaluations also concluded that adoption and impact assessments need to be strengthened. In the case of WHEAT, a critique was that the Program's impact assessments provided limited information on: (a) the scale of adoption; (b) links to relevant publications, databases, or other sources of evidence; (c) a statement about the quality of the evidence provided, including key assumptions and important qualifiers or; (d) extrapolation from specific evidence where the findings are considered generalizable over larger domains than covered in the evidence. In addition, information on the adoption pathway, the projected or realized sustainability of adoption and the timing of the analysis (i.e., the number of years after the completion of the project when the evaluation was undertaken) was also scarce. Similar weaknesses of impact assessments were raised in the other evaluations. In some cases the evaluations also highlighted the lack of procedures to incorporate lessons learnt in subsequent research design (e.g. MAIZE). Overall, the extent of quantitative impact assessments conducted *ex post* is patchy and more ad hoc or opportunistic rather than systematic, and in some cases lacking in rigour and depth (e.g. AAS, WHEAT, MAIZE). This raises the question as to how well Programs are in a position to demonstrate that they are meeting their development targets as encapsulated in the intermediate development outcomes.

3.3.5. *CRP value-added*

Overall, the evaluations convey that all Programs are already generating or have the potential to yield a greater value-added. However, reasons for this and the extent to which the value-added has already materialized varies.

In the case of PIM, evidence for Program value-added resides in the formation of the greatest critical mass in social scientists (and economists) across the CGIAR, positioning PIM to pursue more cutting-edge science. In addition, there was also evidence for additional benefits accruing out of the inter-Center collaboration on value chains. Moreover, PIM is facilitating a more integrated approach by complementing the more upstream policy research of IFPRI with the more delivery focused research arising from some of the commodity oriented Centers (again, primarily in the value chain work). However, some of these gains seem less predicated on Program level mechanisms and possibly are more a reflection of initiatives by individual lead scientists.

The MAIZE and WHEAT evaluations also concluded that these two Programs were already providing value addition to the CGIAR. This was primarily seen as a result of the merging of the CIMMYT and ICARDA wheat mandates in the case of WHEAT, and the merging of CIMMYT and IITA mandates for maize. In both cases this complementarity reduced the risk of duplication and enhanced research critical mass by pooling human and infrastructure resources. Indicators such as increased joint publication across these Centers was an indication of this enhanced coordination and collaboration already occurring.

In some cases the potential value-added was seen as not yet being realized, mainly due to shortcomings in research design. The FTA evaluation expressed some reservations as to whether in its current form the Program would be able to capitalize on the potential benefits of bringing together ICRAF and CIFOR, despite attesting an increased collaboration and coordination across these two Centers. Similarly, in the case of AAS the potential for creating value addition was recognized, based on the introduction of the novel concept of aquatic agricultural systems into the CGIAR, but that as yet the Program has not been able to capitalize on this due to poorly defined and implemented linkages between the aquatic and the agricultural components of such systems (i.e. by linkages to other Centers).

4. Cross-cutting issues

4.1. Partnerships

The nature of the partnerships is clearly in transition and varies across the five Programs. Nonetheless, some general observations can be drawn out from all evaluations. While some Programs (e.g. PIM, AAS) are actively broadening the partnerships to include a wider range of boundary partners in policy (FTA, PIM) and the development sector (NGOs in AAS), across all five Programs the majority of partners were observed to still be other research institutions, with a tendency for Programs to partner with Advanced Research Institutes (ARI) in upstream research (e.g. PIM, WHEAT), while National agricultural research system (NARS) partners seem to be more represented in some of the downstream and application end of research (particularly in WHEAT and MAIZE; or NGO partners in AAS).

The evaluations conclude that the choice of research partners is generally justified and adding value. However, the rationale of choice of partners is not always clearly articulated and could be better linked to impact pathways. What needs to be strengthened in future is a more strategic choice of boundary partners critical in maximising outcomes. Here there is more work to be done, and some evaluations explicitly recommend that IPs and ToCs should be better underpinned by relevant partnerships (i.e. more boundary partners as opposed to research partners). Also, the nature of engagement needs to evolve into being more inclusive or empowering, e.g. allowing some of the stronger NARS to have a greater say in the research agendas (e.g. WHEAT, MAIZE). Some Programs are already moving towards this (e.g. PIM through inclusion of partnership information in annual reports, and a 'Statement of Partnerships'), in other cases evaluations recommend a partnership and engagement strategy to be developed (e.g. WHEAT).

A large number of partners in each Program reflect past partnerships that Centers have had and brought into the newly formed Programs. In some Programs, the complementarity of Centers thus leads to a useful broadening of the overall partnerships. A good example is PIM, where IFPRI historically has had strong partnerships with ARI and universities in upstream research, while non-IFPRI Centers have been more involved with NARS in terms of downstream research (e.g. CIMMYT and NARS in South Asia). In the case of AAS, the evaluation concludes that the balance of partnerships in some FPs is too skewed towards development partners (as a result of research design) and not involving enough universities or ARIs, to the detriment of quality of science. Private sector partners were recognised as important partners for MAIZE by staff.

The ways Programs engage with partners is also variable, both between Programs and between FPs within Programs. In some cases, partners were critical of the way they were being engaged, stating that they had expectations of greater engagement in research agenda setting, research design and joint publications (e.g. MAIZE, PIM, WHEAT). Overall there is scant information in some of the evaluations on how well partners were being resourced through Program funds, or which partners were receiving funds (e.g., ARIs, Universities, NGOs), and which partners were not receiving funds (e.g. NARS). Some evaluations indicated that W1/W2 funds are being used to fund strategic partners (e.g. competitive partner grants in MAIZE and WHEAT).

4.2. Gender

Awareness and acceptance of the need to include gender dimensions into Program research and attempts at establishing a more effective gender mainstreaming approach is evident across all five Programs. However, the effectiveness of gender inclusion varies between Programs.

PIM and AAS were judged to be at the forefront. In the case of PIM there are a number of factors that contribute to this. PIM adopted an explicit gender strategy since 2013 and has established gender specific IDOs in each of its FPs. Implementation is supported through allocations of W1/W2 funds towards explicit gender related research components, resulting in a comparatively high proportion of gender disaggregated data being captured, e.g. PIM's Women's Empowerment in Agriculture Index. This is underpinned by an explicit monitoring framework with gender specific performance indicators. IFPRI's strong reputation in leading gender mainstreaming within the CGIAR and its greater proportion of social scientists were also seen as factors.

AAS has also placed a high emphasis on gender mainstreaming through attempts to build gender into the research design, e.g. through the promotion of its Gender Transformative Approach (GTA), which has been systematically deployed across the different aquatic agricultural systems as well as linking it to the PAR approach. The evaluation considered the gender transformative approaches in AAS to be well considered and potentially effective in better understanding changes in gender norms, perceptions and relations, highlighting this aspect of the Program as one of its main achievements. Insights from both PIM and AAS could provide valuable lesson for other Programs that do not as yet have as well developed gender strategies.

At the other end of the spectrum, WHEAT and MAIZE were still seen to be in earlier stages, initiating some level of gender disaggregation of data, and conducting initial, individual cases of project level activities. Although Program staff in the surveys indicated willingness and interest in engaging more on gender related issues, a constraint emerging from these evaluations was the challenge to build more internal capacity in gender.

The FTA evaluation was the only one that that pointed out the need to transcend the focus on gender to be broader, incorporating social diversity. It also highlighted the challenges faced in scaling gender approaches.

4.3. Capacity development

No formal capacity development plans or strategies have been drawn up by any of the Programs. Capacity development seems to have been more driven from within participating CGIAR Centers, with ICRAF being the only Center reported to have an explicit capacity development strategy in the case of FTA.

However, capacity development is provided within all Programs, through a range of means and for a number of purposes:

- Training of partners in methodologies required to conduct the research (e.g. value chain analysis – PIM; participatory research methods and gender transformative approaches – AAS; advanced technologies ins pre–breeding research – WHEAT, MAIZE)
- Training of next- and end-users in use of the research results (e.g. value chain actors – PIM; extension services and NARS partners engaged in delivery to farmers – WHEAT, MAIZE).

Overall, capacity development was the evaluation criterion that received the lowest level of attention across the five evaluations (see Table 1).

5. Organisational performance

5.1. CRP governance and management issues

Governance arrangements vary between the five Programs. These have been summarised in Table 5, together with some of the main weaknesses recorded by the evaluations.

In total, the five CRP evaluations include eight recommendations relating to governance and management, of which three were fully accepted, four partially accepted and one was rejected. This is a lower rate of acceptance than for the other recommendations. This somewhat reflects the current tension in the System with respect to governance and management issues.

Table 5: Summary of CRP governance arrangements.

	AAS	FTA	MAIZE	PIM	WHEAT
Main governance body	Program Oversight Panel (POP)	Steering Committee (SC)	Stakeholder Advisory Committee (StAC); Management Committee	Science and Policy Advisory Committee	Independent Steering Committee (ISC) (formerly Stakeholder Committee - SC)
Key weaknesses	More of an advisory role than an oversight role	Limited ability to set strategic directions and allocation of funds; No additional advisory panels	StAC was reporting to CIMMYT BoT; StAC initially had advisory role only; established independence since	Only providing some strategic direction and little oversight; Insufficient mutual accountability	SC was reporting to CIMMYT BoT; no independent oversight of Program until change to ISC
Key recommendations	Strengthen role of POP and tighten link to Worldfish BoT	Strengthen mandate of SC (priority setting, resource allocation)	No recommendation on governance, since MAIZE has been implementing changes in line with IEA Review of G&M	Establish an Independent Steering Committee	No recommendation on governance, since WHEAT has been implementing changes in line with IEA Review of G&M

Source: Authors, based on CRP evaluations.

The AAS, FTA and PIM evaluations concluded that Program governance arrangements needed a greater degree of independence from the BoT's of the respective lead Centers.

Conversely, governance was deemed effective in the case of MAIZE and WHEAT, where the CIMMYT BoT has been proactively involved in MAIZE and WHEAT oversight and fiduciary responsibilities. Changes made or underway include appointment of Program directors responding directly to the lead Center BoTs and increasing independence of Program governance. Related to the above weaknesses in governance

arrangements of AAS, FTA and PIM, evaluations concluded that Program management also needed greater empowerment of Program leaders or directors, in conjunction with improved governance arrangements.

Disbursement of W1/W2 funds by the lead CGIAR Centers to partners in the Program was in some cases perceived to be lacking in transparency (e.g. PIM) and sometimes delayed (e.g. FTA). This creates additional cash-flow problems to those generated by delays in transfers from the CO to the Programs (see 3.5.3), in turn affecting the Centers' ability to appoint or sub-contract partners (e.g. PIM, AAS).In both instances, evaluations noted that this leads to high transaction costs (e.g. FTA, AAS, WHEAT) and demotivation of staff (e.g. MAIZE).

Reporting and planning procedures was also generally seen as an area in which the Programs had weaknesses, although the evaluations assessed reporting requirements as onerous. In some cases there was evidence of incomplete project planning and budgeting (e.g. AAS, PIM, WHEAT), in turn making tracking how projects were delivering problematic. In other cases, reporting systems were in place, but only in terms of tracking outputs rather than how outputs were delivering on stated outcomes (e.g. FTA). It was also observed that more appropriate (i.e. fit for purpose reporting in relation to Program needs) planning and monitoring systems were still under development at the time of the evaluations (e.g. PIM, MAIZE, WHEAT).

5.2. Monitoring, evaluation and learning

The degree to which Programs had established a distinct culture of learning, as opposed to purely reporting and monitoring in response to various levels of reporting requirements (CGIAR, Program, donors) was not given much prominence in the evaluations; in fact some evaluations didn't explicitly assess monitoring, evaluation and learning (ME&L). In the other cases, monitoring and evaluation were primarily analysed with respect to fulfilment of accountability needs (i.e. reporting *sensu strictu,* as discussed in section 2), while evaluation was primarily discussed in the context of adoption studies or impact assessments. These were generally found to be lacking in rigour, and not being systematically conducted. The AAS evaluation was the only case where learning and feedback mechanisms were assessed in more detail, given this Program has an explicit Knowledge Sharing and Learning theme, albeit with little evidence that these mechanisms were contributing to a strengthening of the research design in AAS. The MAIZE evaluation also noted the lack of feedback mechanisms.

It was noted that setting up ME&L systems was still underway, some Programs having only just appointed dedicated resource persons (M&E specialists; e.g. MAIZE, PIM); in the case of the WHEAT evaluation, it was recommended that in future adoption studies and impact assessments there be an explicit link back to Program or FP level IPs and ToCs.

5.3. CGIAR system issues affecting CRP performance

A very consistent pattern observable across all five evaluations is the negative impact the unpredictability of W1/W2 funding is having on Programs and partners, mainly due to delays in fund transfers (and initially, the inability to carry over funds from one financial year to another). This significantly impinges on Programs' ability to strategically utilise W1/W2 funds. In many instances, the resulting cash flow problems require Centers to pre-fund salaries and activities, which can only be done to a certain extent because of fiduciary risks, increasing the opportunity costs of recruitment, time spent on work arounds, and additional transaction costs of senior managers (see 5.3.). The way that W1/2 funds are allocated and disbursed by donors leads to a perception that W1/W2 funds, while they are more flexible, are also more short term in nature, again discouraging a more long term, strategic use of W1/W2 funds.

Partly because of these tensions arising out of W1/W2 funding delays, but also partly due to perceived duplication and complexity in reporting arrangements as well as process overload (e.g. in preparation of Extension Proposals), all evaluations noted that the relationship between Programs and host CGIAR Centers, and the Consortium Office were problematic.

A related problem that the evaluations noted for some Programs was that many donors still do not cover full costs recovery of projects in W3 or bilateral projects (despite this being a CGIAR requirement), reinforcing the use of W1/W2 funds to be used as base funding for unmet infrastructure costs of Centers, further compounding the impact of delays.

6. Implications for CRP Phase 2

The purpose of the previous chapter is to provide an overview of the key findings of the evaluations against a common set of predetermined criteria used by the evaluations, noting that these evaluations commenced in 2013 (FTA evaluation) and concluded at the time CRPs were moving into the extension phase. Despite some variations in the extent to which evaluations assessed Programs against the evaluation criteria, a number of clear headline lessons can be extracted by this synthesis. In this chapter we discuss some of the implications of these headline learnings offering considerations relevant to the design of phase 2 of the CRPs.

The main lessons derived from this synthesis (listed against selection criteria) and requiring further elucidation are:

- *Relevance and effectiveness* – strengthening theories of change and tightening the articulation of impact pathways;
- *Quality of science and research design* – evolving frameworks to better capture systems science dimensions and modalities where integrative science is necessary;
- *Partnerships* – moving towards more strategic and effective partnerships;
- *Gender* – transcending the present (narrow) focus on gender mainstreaming to include dimensions of social equity and inclusion;
- *Monitoring, evaluation and learning* - promoting a stronger learning culture and strengthening researcher capacity to engage in reflexive processes.

These lessons and their implications are explored in greater depth in the following sections of this chapter.

6.1. Strengthening theories of change and impact pathways

One intent of the CGIAR reform process has been to instigate the use of theories of change (ToC) and reorientation towards outcomes based on plausible impact pathways (IP), to determine i) the research priorities, ii) the partners and iii) the key pathways to impacts, enabling a much more rigorous articulation of how Program research will achieve envisaged IDOs. However, as discussed in sections 3.1 and 3.3, while the evaluations (with exception of AAS) note the progress made here, they also identified the need for further improvement. In some cases suggestions are made in the evaluations, which are amplified here:

1. Flagship Project and project level ToCs and IPs need further refinement, in particular to enhance linkages (coherence) and better integration. Linkages between project plans or logframes and ToCs and IPs need to be better captured in project documentation and reporting to facilitate future evaluations of effectiveness.
2. Further orientation of Programs towards achieving outcomes necessitates the development of much more strategic and nuanced partnerships that formally recognise and include relevant boundary partners. Choice of partners, their roles and how they are resourced needs to be more clearly reflected in FPs and project level ToCs and IPs, articulating specific engagement processes to operationalise pathways to impact (see also section 6.3.).
3. Program monitoring and evaluation frameworks as well as impact assessments need to explicitly link back to impact pathways.

6.2. Strengthening research design

There is a growing recognition that there is an increasing need for integrative science to underpin the complex nature of challenges arising from agricultural development, maintaining food security, combating resource depletion and facing climate change. Amongst other objectives, to more effectively address these global challenges and in response to higher donor expectations, the CGIAR established the Programs. It was anticipated that by bringing together more diverse and complementary disciplinary skills sets from across a wider spectrum of CGIAR Centers under the umbrella of a Program, the CGIAR would be better positioned to tackle these challenges in a more systemic way. This opens up as yet an unresolved tension between the when, where and how systems science needs to be deployed to move beyond just interdisciplinary or integrative science, and how this manifests itself in CRPs' research design. In this section we attempt to explore this tension field a little more, offering some additional aspects that need to be considered in helping shape better research designs in phase 2 of the CRPs.

Being able to draw on a wider range of scientific skills within a given Program is a necessary prerequisite to conduct systems science, but the extent to which this translates into an ability of *doing* systems science also depends to a large extent on the right questions being asked, appropriate research designs being implemented, and engagement with relevant partners to span the research – outcome – impact continuum. As the evaluation criteria cut across these dimensions, the issue of systems science (particularly its relevance, quality and effectiveness) was to varying degrees implicitly canvassed by four of the evaluations. Conversely, this tension field was explicitly analysed in quite some depth in the AAS evaluation.

A constraint to answering the question whether systems science is required or not is that there is still a degree of ambiguity of what is meant by the word 'system' in systems science, and defining systems science is a topic of ongoing debate within the CGIAR. In order to better answer the above question in the context of future Program research design, the CGIAR needs to define and provide consistency in the use and understanding of what is meant by 'systems' and 'system approaches'. Dimensions of systems science include (but are not restricted to) questions relating to:

- *What system?* – Food systems (PIM, etc.); farming systems (MAIZE, etc.); livelihood systems (FTA, etc.); value chains (PIM, etc.); socio-ecological systems (AAS, etc.).
- *What systems science?* – Modelling (PIM, etc.); big data (PIM, etc.); multi-scale analysis and scaling (FTA, etc.); integration of biophysical and social science (AAS, etc.).
- *What systems science modality?* – Inter-disciplinarity vs. trans-disciplinarity.

In addition to defining 'what system', the question of where and when systems science approaches are *essential* for delivery of outcomes is likely to need more explicit articulation by the Programs in the design of their programs as they move into Phase 2. There are several frameworks that could facilitate that. One potential approach to facilitate the discourse on 'systems approaches' could be to define the appropriate balance of component (or 'bounded') vs. integrative (or 'systems') science, for example using the framework developed by Stone-Jovicich et al. (2015)[7] and presented in more detail in Annex 3.

The key message is that determining the appropriate modality of research for development depends on the problems being tackled and the types of changes and impacts being sought. 'Bounded' types of research can

[7] Stone-Jovicich, S. J. Butler, L. McMillan, L. Williams, C.H. Roth. 2015. Agricultural Research for Development in CSIRO: A review of principles and practice for impact. CSIRO Agriculture, Canberra, 44pp.

contribute significantly to food security and improved livelihoods and wellbeing of rural communities. But unless that research is embedded in a broader 'systems-oriented' research perspective, the impacts are also bounded (i.e. tend to be either narrow and short-lived, or inappropriate and even harmful to rural communities as demonstrated by decades of conventional research for development applications).

A separate but related issue of system science that has significant bearing on research design is the question of better integration within Programs, in the context of the system being addressed, the balance between 'bounded' and 'systems' science mentioned above, and noting that integration in itself does not necessarily lead to systems science.

Aspects of 'integration' that need to flow into research design are: Clarity on which 'system' is requiring an integrated approach

- Operational integration (e.g. between CGIAR Centers within a Program; between Programs and external partners)
- Science integration (e.g. to address 'system' level problems; methodological integration)
- Scale of integration (e.g. local/household, production unit/value chain, or national/policy).

Related to the above dimension of integration (which for instance was debated in the WHEAT evaluation) is the question of which science mode supporting integration is most appropriate for the Programs:

- Interdisciplinary science: integrates different science disciplines, and is still essentially science driven – core comparative advantage of CRPs?
- Transdisciplinary science: requires the co-development of and acting on science and non-science based knowledge; it involves science *and* non-science actors - is this where Programs should be heading through broader partnerships (for instance the AAS evaluation proposed more transdisciplinary research)? Have they got the processes and resources in place to support this?

Enhancing integration and inter-disciplinarity requires a significant increase in interactions between multiple actors, raising transaction costs and requiring resources. The evaluations provided a cursory analysis of factors magnifying integration, but also raised some of the constraints (e.g. high transaction costs). Dimensions that may need to be explored in more detail in future comprise:

- Strengthening incentives for increased engagement: currently incentive signals to project and science leaders from CGIAR Centers are too weak
- Changing reward systems: promotion and career advancement in many instances still depend largely on publication metrics, and metrics regarding achievement of outcomes are lacking. For the latter to take more effect, in addition to publication metrics, individual performance needs to be linked more to robust outcome and impact frameworks, recognising that outcomes and impacts cannot always be planned in advance
- Adequate provision of staff time and resources in project plans for increased engagement.

In addition to these issues of research design in support of systems science in Phase 2 of the Programs, all of the above issues raise new questions for evaluative processes in the future. Performance assessment of how Programs are embracing system science will require a broadening of QoS metrics, possibly requiring a rethink of research design as a primary dimension to assess.

6.3. Evolving the CGIAR partnership model

There is a growing discourse on the need to transcend traditional modes of dissemination of agricultural research outputs[8]. There is now considerable evidence that agricultural research is most effective in creating solutions to food systems challenges when it is coupled with the efforts of public agencies and private companies. Success is contingent not only on the invention and adoption of single component technologies by farmers for on-farm issues, but the creative dynamic between component technology, business model, value chain and policy innovation.

Using agricultural innovation to support agricultural development is not an issue of science and technology alone - it is an issue of coupling technology with different types and sources of innovation in ways that create more relevant solutions and opportunities. Hence, partnerships are a central mechanism in harnessing agricultural innovation to bring together different technologies, ideas, resources and capabilities and creating the conditions needed to make productive use of them.

Such partnerships transcend the traditional notion of partnership in the CGIAR, evolving from traditional *research* partnerships still prevalent in Programs and described in the evaluations, to innovation platforms and systems that span research, the private sector, civil society and government institutions. The evaluations touched upon this, but more thought by the CGIAR regarding the how, who and where needs to occur. In fact, partnership models are a researchable issue in their own right, providing a theoretical base for magnifying impacts and to validate the ToCs.

The prospect of the partnerships being effective needs a more strategic selection of relevant boundary partners, co-planning in the design phase and adequate resourcing. Questions that need to be asked and aspects of partnerships that need to be considered are:

- Formal (governed by agreements), or informal (loose associations)?
- Constituted largely by external research partners (NARS, universities, ARIs) or comprising development/boundary partners (NGOs, private enterprises, government agencies)?
- Receiving CRP funding, or co-investing?
- Selected based on historical relationships or explicitly targeted?
- What role can the private sector play in augmenting government services to reach out to socially disadvantaged groups?
- How can insights in the above be used to inform NGOs, policy makers and public-private partnerships and lead to better design of and implementation in future policy and development interventions?

6.4. Transcending the focus on gender

As reported in section 4.2., all evaluations assessed how Programs have started addressing the gender agenda. However, arguably this focus on gender is too narrow and has come at the expense of losing sight of wider issues of social equity and adverse exclusion. The CGIAR Programs ultimately lead to new technologies and polices that promote further agricultural intensification, if anything to close the gap between population

[8] E.g. Hall, A., W. Janssen, E. Pehu, and R. Rajalahti. 2006. Enhancing agricultural innovation: how to go beyond the strengthening of research systems. Worldbank Report. Washington, DC.

growth and food security, and the past decades of CGIAR led international agricultural research has been quite effective in that regard. However, agricultural intensification is coming at the cost of an increasing social dichotomy between more affluent (commercial) land holders (or agri-enterprises) and the more socially disadvantaged groups, of which women constitute one, but not the only component (Like landless or marginal smallholders, tribal minorities).

Some of the dimensions that the CGIAR might need to consider comprise:

- How and why are different rural livelihoods affected by agricultural intensification in key agro-ecological settings? Who gains, who loses?
- How do institutional arrangements and power structures mediate access to knowledge and inputs?
- What are the strategies that could lead to increased social inclusion and a reduction in unintended consequences of agricultural intensification?

6.5. Strengthening ME&L

The reporting weaknesses discussed in section 5.1., when seen in conjunction with the lack of or only weakly developed adaptive management and learning approaches as canvassed in section 5.2 points to a major deficiency in the way Programs are operating, calling into question their ability to maximise their effectiveness, let alone being able to credibly demonstrate the achievement of stated outcomes and impacts.

Theories of change are not static, and as unanticipated opportunities present themselves and alternative impact pathways open up, there is a need for Programs to be able to more flexibly adjust and capture these changes in an adaptive management approach. However, this requires a culture of reflexive discourse about what works and what not and why, being able to challenge initial assumptions underpinning IPs and take corrective action. There was some evidence in the evaluations that this change in culture is occurring as a result of establishing CRPs, but this change process needs to be underpinned by improved ME&L systems and an ongoing strengthening of researcher capacity to engage in such reflexive processes. Some of the results of the staff surveys indicate that there is a willingness to embrace this.

ANNEX 1 - Overview of CRPs evaluated

CRP	Lead-Center	Key CRP characteristics and focus	CRP information	Evaluation information
Forests, Trees and Agroforestry (FTA)	CIFOR	An integrated global research initiative that aims to enhance the management and use of forests, agroforestry and tree genetic resources in the developing world as a way to improve livelihoods and sustain environmental values. CRP spans across a wide range of topics, from small-scale production technologies through to international trade and global conventions, with multiple partnerships, and with a diverse set of strategies to achieve impact.	http://foreststreesagroforestry.org/	http://iea.cgiar.org/evaluation/crp-evaluation-forests-trees-and-agroforestry-fta
MAIZE	CIMMYT	Aims at significantly improving the productivity, resilience and sustainability of maize-based farming systems thus contributing to farmer incomes and livelihood opportunities, without using more land, and as climates change and fertilizer, water and labor costs rise. The strategy is three-fold: sustainable intensification and income opportunities for the poor; new maize varieties for the poor; and integrated post-harvest management.	http://maize.org/	http://iea.cgiar.org/evaluation/crp-evaluation-maize
WHEAT	CIMMYT	Aims at building on the input, strength, and collaboration of public and private sector partners to catalyze and head an emergent, highly-distributed, virtual global wheat innovation network to improve productivity of wheat farming systems, address the global threat of stem rust disease, and help wheat farmers in developing countries grow their crops in hotter conditions with less water and less fertilizer.	http://wheat.org/	http://iea.cgiar.org/evaluation/crp-evaluation-wheat
Policies, Institutions and Markets (PIM)	IFPRI	Aims at overcoming challenges and failures that prevent effective functioning of policies, institutions and markets needed to improve the supply of key public goods and services, direct incentives towards agriculture, help stabilizing food prices, and strengthen relationships that create wealth. Also addresses macroeconomic dimensions, environmental inputs and outcomes, and important enabling conditions and aims at improving quality of the policy environment through improved observation and metrics.	http://pim.cgiar.org/	http://iea.cgiar.org/evaluation/crp-evaluation-policies-institutes-and-markets-pim

Aquatic Agriculture Systems (AAS)	WorldFish	Aims at fostering innovation and innovation capacity of the poor and vulnerable that depend upon aquatic agricultural systems to to improve their own well-being in the face of opportunities and challenges. The CRP represents largely a new research agenda in the CGIAR that uses a demand-driven gender approach and operates through participatory action research. Also aims at establishing effective learning and partnership arrangements that can lead to impact at multiple scales.	http://www.aas.cgiar.org/	http://iea.cgiar.org/evaluation/crp-evaluation-aquatic-agricultural-systems-aas

ANNEX 2 – CRP Synthesis – Main dimensions

1. **Relevance**

- Aspects of programme design/coherence (Proportions of W1/2, bilateral funding alignment of funding with IDOs)
- How well are the CRPs aligned with global and regional priorities?
- What is the comparative advantage of the CRP/participating centers?

2. **Quality of Science**

- Assessment of the inputs (researchers, infrastructure)
- What are the mechanisms for ensuring high quality science?
- Of what quality are the outputs produced by the CRP?
- What are the constraints to high quality of science within the CRP?
- What is the role of the centers in quality of science?
- Has quality of science been maintained or improved?

3. **Effectiveness (more contextual)**

- Aspects of programme design/coherence - design process
- What are the factors which are influencing (positively and negatively) the effectiveness of the CRP?

4. **Impact (focus on processes)**

- Articulation of theory of change
- Impact pathway development/how impact pathways are used for program management and learning
- Evidence of impact from legacy work

Cross cutting

5. **Partnerships**

- How are partnerships managed/how are partners involved in the CRP?
- How do partnerships increase the effectiveness of the CRP?

6. **Gender**

- How is gender integrated into the CRP? What importance is given to gender?
- How does gender increase the effectiveness of the CRP?

7. **Capacity development**

- How does the CRP address capacity development issues? What types of capacity development activities are undertaken?
- How effective are capacity development activities?

8. Organizational performance

- CGIAR general system issues affecting CRP performance
- Positive and negative aspects of the new structure
- What have been good management practices to increase the effectiveness of the CRPs?
- What have been constraints for effective management?
- Monitoring and evaluation and learning
- Observations on Center versus CRP management

9. Value added/conclusions

- Enhanced relevance
- Results orientation
- Increased coordination

ANNEX 3 - The nexus between research & impact: a conceptual framework

An approach to facilitate the discourse on 'systems approaches' could be to define the appropriate balance of component (or 'bounded') vs. integrative (or 'systems') science is the framework developed by Stone-Jovicich et al. (2015)[9] which is presented in brief here.

Figure 1 highlights the multiplicity of research for development (R4D) modes and associated underpinning assumptions about the characteristics of the 'problems' being tackled, the dynamics of change and interventions for impact; and implications for practices around research. The roles for research in these different modalities differ depending on the characteristics of the problems being addressed, and underpinning assumptions about dynamics of change, modes of intervention needed to pave pathways to impact, and types of impacts targeted (vertical axis in Figure 1). These in turn shape the ways in which research is conceptualised and implemented (i.e., research practices; horizontal axis in Figure 1).

Bounded' R4D comprises research approaches that focus on enhancing understanding of (and having impact on) a particular or small set of component(s) of the system via technical and/or discipline-based scientific expertise. The path to impact is commonly conceptualised in terms of transfer of expert/scientific knowledge (e.g. irrigation systems and yield variability; household division of labour and agricultural productivity) to a particular set of actors (e.g., farmers, local agriculture extension office) through coordinated actions (e.g. via training modules) that results in the adoption of 'best' or 'good practices'.

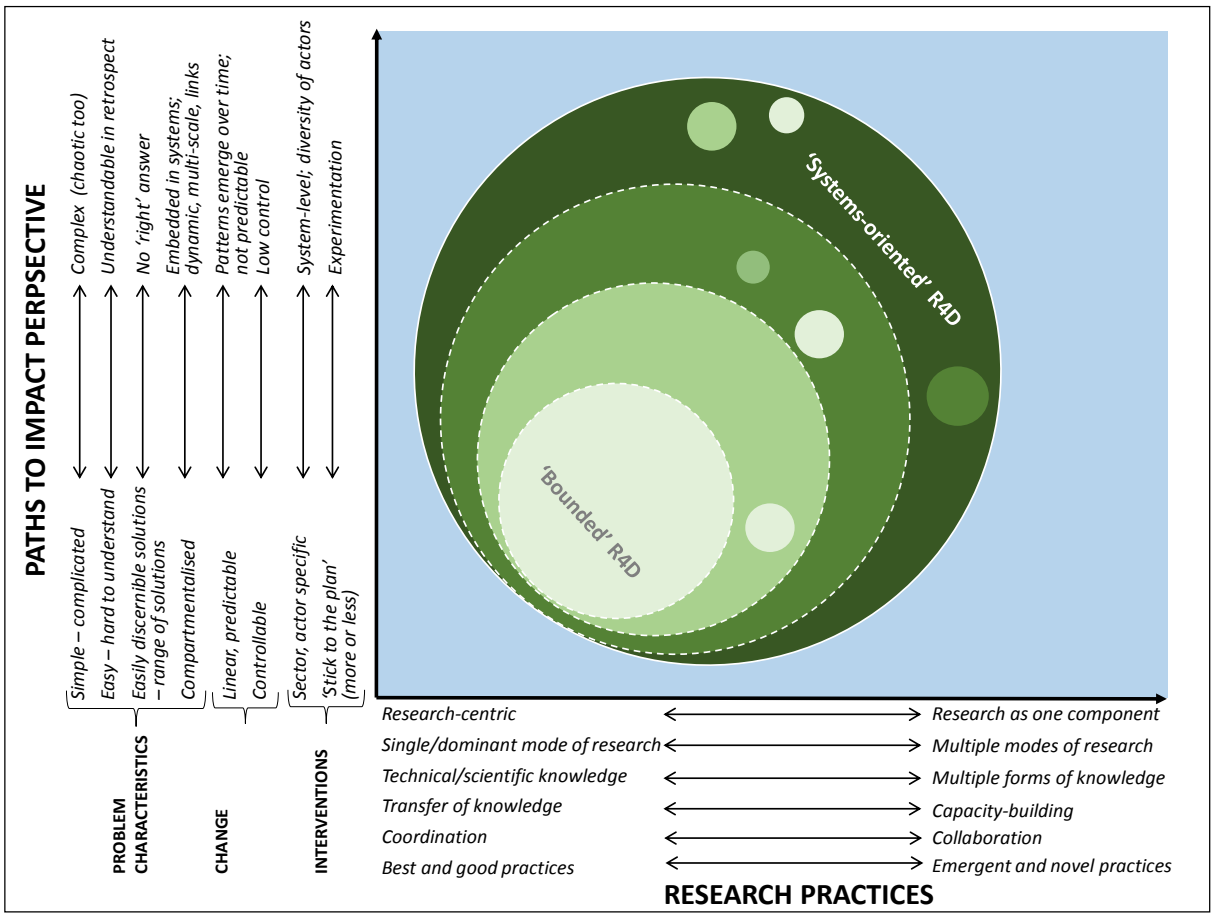
By contrast, 'systems-oriented' approaches presume that 'problems' are always embedded in a wide and dynamic system characterised by multiple interconnections, uncertainty and unpredictability. As such, international agricultural development is seen as inherently complex. This means that interventions for impact require focusing at a systems-level, including across sectors, scales, and actors. Interventions also need to be experimental, flexible and adaptive in nature in order to be able to respond to unexpected and emergent changes and opportunities. Equally deemed critical is the pursuit of emergent, novel and innovative approaches underpinned by active learning (as opposed to 'best practices' and 'good practices' which are argued to have limited impact in complex situations).

The key message in Figure 1 is that the modality of R4D that is most appropriate depends on the problems being tackled and the types of changes and impacts being sought. 'Bounded' types of research can contribute significantly to food security and improved livelihoods and wellbeing of rural communities. But unless that research is embedded in a broader 'systems-oriented' research perspective, the impacts are also bounded, i.e. tend to be either narrow and short-lived, or

[9] Stone-Jovicich, S. J. Butler, L. McMillan, L. Williams, C.H. Roth. 2015. Agricultural Research for Development in CSIRO: A review of principles and practice for impact. CSIRO Agriculture, Canberra, 44pp.

inappropriate and even harmful to rural communities as demonstrated by decades of conventional R4D applications, (see related discussion by Hall et al. 2010[10]).

Figure 1: A framework to define the need for systems science – building on the nexus between research practices and impact considerations



Source: extracted from Stone-Jovicich et al. 2015

The co-centric circles in Figure 1 reflect the wide range of approaches along the 'bounded' to 'systems-oriented' spectrum. The circles do not suggest hard boundaries between these different modalities. The further outward the circle, the greater the emphasis placed on systems and systemic change and impacts. Moreover, 'bounded' and 'systems-oriented' modes of research are not necessarily mutually exclusive. For example, 'systems-oriented' approaches can/often do incorporate, as part of a larger suite of research practices, 'bounded' research (e.g., development of drought resistant crop varieties, ethnographic studies of culture- and gender-based divisions of labour, political economic analyses of agricultural structural adjustment programs).

[10] Hall, A., J. Dijkman, and R. Sulaiman V. 2010. Research Into Use: Investigating the Relationship between Agricultural Research and Innovation. United Nations University - Maastricht Economic and social Research and training centre on Innovation and Technology, UNU-MERIT