Experiences from Five Years of Performance Measurement System:
Research-related indicators

Lessons drawn by the Science Council
Final Draft 25 September 2009

Key lessons from the Science Council – Research-related indicators

- Each indicator must create incentives to improve performance in the area it is intended to measure. For example, one measure of *outputs* is the achievement of planned output targets. Yet, if good performance is equated with high success rate, such an indicator can create perverse incentives for innovative science in the quest to ensure high achievement rates.
- Peer reviewed publications are a universal measure of research-based *outputs* but care is needed for not focusing on quantity at the cost of quality.
- Compared to the current PMS, a balanced set of indicators for output should include measures that are appropriate for management and mission, such as capacity strengthening and management of data as an IPG, and program specific outputs.
- Tracking successful outcome cases is a reasonably reliable measure of outcome performance reflecting progress in the part of the impact pathway for which Centers/programs and their partners can be held accountable but not assuming that all research can lead to outcomes. However, research results—achievement of deliverable outputs and outcomes in particular, are better monitored at intervals of a few years rather than annually.
- It is not feasible to use actual impacts as an indicator to measure of impact performance of institutes. Impact culture, however, is a reasonable proxy and other means of fostering impact may be added to what is currently included in the PMS.
- It is important to set benchmarks for each indicator for promoting self improvement of Centers and programs that are monitored. Performance information should be used by scientists and managers for internal program adjustments, not only for reporting.
- It is important that investors do not use only the annual performance indicators for decision making; they should combine information from other monitoring and evaluation sources for informed decisions. Mechanistic links between indicator information and funding should not be promoted.
- Annual variability in indicator results can be expected in research and multi-year rolling averages give a more reliable picture of improvement of performance over time.
- It is important that there is consistent interpretation of indicators and common understanding of what constitutes good performance. Management analysis of the results is often needed for bringing about the desired improvements.

- Indicator results can provide a useful source of performance data for periodic reviews but the current set, due to its limited coverage, needs to be complemented by other means of monitoring and peer review.
- In the new CGIAR it is important that performance measures stimulate high performance and do not stifle ambitious and novel research that can have high impact if successful.
- Of the current indicators, publications may be useful for monitoring Center as well and program performance. Other aspects may be needed for monitoring program output performance. Monitoring outcome performance is relevant for the programs that are the implementation mechanisms for CGIAR research. Monitoring enhancement of impact may be needed both at the Center and program level. *Ex post* impact assessment of sufficiently large research interventions remains essential.
- Performance measurement of results represents a systematic activity to collect data and information about Center/program products and the relevant data should be made available for multiple uses.
- At least part of the performance indicators for programs should be adjusted according to the program and the nature of its components.

1. Executive Summary

The Science Council (SC) has been involved in the design and implementation of PMS indicators for CGIAR Centers that relate to research results. These have included indicators for outputs, outcomes and impact. In the four years of PMS, several modifications have been made reflecting experiences and feed-back from Centers. The SC has sought to develop the indicators so that they provide incentives for Centers to constantly improve their performance against defined benchmarks and thereby fulfil their accountability imperative and the purpose of program/organisational improvement.

The context in which the indicators were designed was set by a) external (donor) demand for the PMS and submission of indicator results to external users rather than primarily an internal (Center/Alliance) tool for monitoring b) the use from the start of the indicator results for funding decisions, i.e. for Center comparison; and c) the need for the indicators to apply equally to all the Centers irrespective of their very different disciplinary activities and modes of operation, and be as unambiguous and fair when used for Center comparison.

The current PMS includes two quantitative indicators for outputs, the major one being on publications and one indicator for both outcomes and impact (addressing impact culture). Publication in high quality venues is a universally accepted measure of scientific output. The CGIAR needs to be a credible scientific institute to be a preferred partner by NARS and ARIs. Although this indicator covers output performance only partly it has not been possible to develop others that would meet the conditions listed above for uniform applicability.

Monitoring output alone is not sufficient for a mission-oriented institute like the CGIAR; there is an expectation that the Centers' research results are used and that Centers monitor their main areas of research for uptake of research results as part of their outcome and impact-oriented culture. The outcome indicator adds both a measure of the diligence of the Centers in monitoring and documenting outcomes across their research portfolio and, also, a measure of documentable and significant outcomes derived from a part of Centers' research and capacity building activities.

The SC did not consider it feasible to develop an indicator for actual impact for the following important reasons: (i) The ultimate impacts on CGIAR goals are not in the control of the Centers and they can be very far down the path from the research activity and determined by individual circumstances; (ii) Attribution of impact to a Center is difficult and not necessarily desirable when research is conducted in partnerships; (iii) Centers' impact pathways are very diverse due to the very different research areas they engage in; (iv) Centers' different age influences the volume of impact that may have resulted from their work; (v) Impact does not reflect the research performance of the Centers to-date as it occurs with long time lags. An indicator for impact culture was developed instead to measure Centers' efforts to document impact from past research and to measure Center efforts to institutionalize impact culture among their own researchers and partners. Strong commitment to documenting impacts *ex post*, is likely to be correlated with the actual impact of a research Center.

The PMS has been in use for 5 years. *Has the PMS improved Center performance*? There is no clear trend to show that Centers have consistently improved over the period. Some Centers have on average been on a steadily improving or stable trend of good performance but, in most cases, there has been considerable year to year variation. With the current benchmarks all Centers have scope to improve. Internal analyses by Centers of results - and actions to improve in areas where weaknesses have been identified – are both needed for gradual improvement over time.

There are however good indications that the process has led to better practices in some Centers in such areas as planning (MTPs); linking outputs from that planning to outcomes; recording results; and allocating resources to outcome and impact evaluation and documentation. The outcome and impact culture indicators involve qualitative assessment by the SC and other external peers. The Center-Science Council interaction in the process, involving feed-back to Centers on the results, may have been beneficial helping the Centers to improve their planning and outcome/impact pathway monitoring. The improved analysis and discussion of research results and effectiveness would be a very positive result and forms an important part of performance monitoring.

What has been learned over the 5 years? An important lesson was learned when output performance was initially measured against set output targets. This measure seemed logical as it linked performance (regarding deliverable results) to the institutional mission. However it failed because of the inherent risks of perverse incentives when institutions are cognizant of being compared with others for their results and rewarded for higher results. The very high rate of achievement of planned outputs uniformly

across Centers suggested that targets were not ambitious, potentially stifling the risk-taking behaviour essential to the scientific endeavour. This experience also suggests that creating incentives for constant improvement in the actual performance and rigor of self-monitoring may not be compatible with the goal of providing information to donors when this information is used directly for funding decisions.

In the new CGIAR, the SC considers it important that performance will be monitored at the Center and program levels through appropriately designed, even tailor-made indicators that are integrated to other internal monitoring and external evaluation mechanisms. Any performance monitoring should allow more refined and analytical processes to take place than with the current PMS. This is required at the different levels for improving performance, learning lessons, enforcing internal and external accountability, and providing strong incentives for management and researchers. It is important that the research performance measures make internal sense in the program research context and are useful for decision-making; that a certain degree of risk and uncertainty is accepted and responses to performance information are clear regarding changes in direction, corrective measures and acknowledgement of excellent performance.

2. Background

Design of the PMS indicators

In 2003 the CGIAR initiated the development of indicators for performance measurement. The initiative for the PMS was laid when the World Bank decided to start to allocate a part of its funding based on the results of annual performance indicators. The design process of the PMS was led by a Working Group on Performance Measurement. The Science Council, when it was established in 2004, was tasked to design and operationalize results-based performance measurement indicators relevant for the research component. The PMS was launched as a pilot in 2005.[1]

The WGPM intended the PMS to serve multiple purposes, but the two main ones were promotion of high Center performance and accountability towards achieving goals. The SC perceived that the main use of the PMS indicators would be for the Centers to constantly improve their performance against defined benchmarks. Both the WGPM and the SC put emphasis on wider and more continuous use of self-assessment. There was concern however that the PMS would be used for other purposes with direct implications on Center funding. In 2009, the SC provided a Guide to the use of the results-based PM indicators.[2]

The SC's aim from the start was to design indicators that would relate closely to what the CGIAR is trying to accomplish, i.e. research results for development impacts. This required indicators for both good science and for outputs, outcomes and impact on the

[1] Although the PMS was launched for piloting in 2005 the World Bank had already began to allocate its funding in 2004 based on a set of preliminary indicators agreed with the Center Directors' Committee.
[2] http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/SC_documents_to_ExCo_16/SC_Report_to_ExCo_Members_18_May_2009.pdf

mission of the CGIAR. The SC sought to design indicators that provide an accurate reflection of research performance rather than those things that can be easily counted. In his contribution to CGIAR News in September, 2004, the SC Chair (Per Pinstrup Andersen) commented on this concern in the design of the PMS. He warned that *"a focus on what can easily be counted and compared across Centers and programs may also reduce the incentive to take risks in breaking new ground and seeking new solutions, using a learning and feedback mode, which is so important in innovative applied research"*. Subsequently, the SC put its emphasis on seeking ways to measure progress towards agreed goals (outputs, outcomes and impact) and linking the PMS to research planning through the Centers' Medium Term Plans, (MTPs).

In the early years the PMS included two components, "Results" and "Potential to Perform", that contained indicators related to research (see Annex x). The results-based indicators covered outputs, outcomes and impacts. Of these the indicator for outputs was designed to be self regulating. The indicators for outcomes and impact required subjective assessment by peers.[3] The latter component included indicators on publications.

It was foreseen at the very start that the progress to a fully established system would be long due to the complexities of developing a PMS and the need for consultation and buy-in at all stages. Early on the SC initiated a joint annual meeting with the Center research directors (later the Alliance Deputy Executive) where the experiences from the PMS on research-related indicators were discussed (among other issues) and improvements were considered. The SC's Standing Panel on Impact Assessment (SPIA) through its regular contact with the Centers' Impact Assessment Focal Points (IAFP) held consultations regarding the indicators for impact. Each year the SC has published its feed-back with the discussion of the results and experiences from the research-related indicators.[4]

In 2007, the Science Council organised a workshop in Rome to discuss the lessons learned from the first 2 years of the PMS. This workshop led to the strong recommendation to remove the indicator for Outputs[5] that was based on the

[3] The initial indicator for outputs was directly linked to the Medium-Term planning and in the outcome indicator there was a requirement of linking outcomes to research planned in a specific MTP.

[4]

- http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Performance_Measurement/SC_Suggestions_to_Implement_the_CGIAR_PM_System.pdf.
- http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Performance_Measurement/PM_2005_moving_forward_Oct31.pdf. Joint document with the CGIAR Secretariat.
- http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Performance_Measurement/SC_Comments_on_PM_Results_indicators_2007_FINAL.pdf
- http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Performance_Measurement/SC_Feedback_on_the_2008_PMS_exercise.pdf
- http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Performance_Measurement/PMS-SC_assessment_of_the_research-related_indicators.pdf

[5] Outputs are the products of research with a defined time line, contributing to reaching the Center goals by offering solutions to problems identified during the planning process.

achievement of output targets[6]. Monitoring of this indicator relied on Center self assessment of how many output targets planned for the year of reporting had met. It seemed the most direct indicator of program performance but in fact created perverse incentives and inherently discouraged engagement in challenging research.

Following a subsequent workshop in Washington in 2008, several changes were made to the PMS for submission of 2008 data. There was agreement on the way output performance would be reported and further developed (discussed below). Changes were also agreed for the outcome and impact related indicators. A clear rationale was provided for each indicator and there was agreement to set benchmarks and performance targets for the indicators.

The following provides a discussion of the design and implementation of each of the indicator elements that directly concern the Science Council. The lessons and conclusions are presented in the context of forthcoming changes in the CGIAR.

3. Evolution of the Indicators and Lessons Learned

Outputs

Design

As mentioned above the initial *Output indicator* was linked to the MTP. In the rolling plans three-year logframes project multi-year outputs and annual output targets[7] for each of the major projects. The indicator was based on the logic that output performance in research should be measured against the expected results from Centers. The indicator was self-reported achievement of output targets declared for the first planning year and calculated as % of total output targets planned. The achievement of output targets reported by each Center was audited by an external auditor on sampling basis. The SC, in its commentaries of Center MTPs provided guidance on the clarity and substance of the output targets and subsequently provided a commentary on the PM submission of output targets. This process led to a considerable improvement in the clarity and quality of the planning logframes and descriptions of outputs, output targets and outcomes.

However, the experience over three years suggested that linking the monitoring of output target achievement to the PMS creates perverse incentives. The results over three years were uniformly high averaging at about 90% success and in many cases there was full achievement of the output targets set. However, Center research ought to involve strategic, innovative and even high risk projects where targets are difficult to achieve but if achieved have potentially very high impact on intended goals. The expectation of 100% success therefore was unrealistic and counter productive. Furthermore, the indicator did not facilitate the research process where a) failure to achieve the expected results may provide important understanding and feedback about the research problem;

[6] The output targets that are defined as: the annual deliverables, defined by quantity and type, expected in a specific year and contributing to achieving the MTP Project Outputs.

[7] Output targets are planned in the five categories: materials, policy strategies, practices, capacity, and other kinds of knowledge.

b) serendipitous results can have high value, and c) scientists should have the incentive to drop a line of activity that they can tell is going to be unproductive, and shift their attention to something that their research has indicated is more promising. Thus, measuring % achievement of output targets made the production of pre-specified output targets more important than the process of discovery inherent to research. In summary, the SC's arguments to replace the output targets as the basis for an output indicator in the PMS were:

1. The measure provides the incentive to specify lower targets (even if the Centers are working toward greater targets);
2. The measure provides the incentive to pursue targets even if they are discovered to be no longer appropriate;
3. Information on achievement or research results can be managed and better understood through internal monitoring and external review; and
4. Relevance of research and likelihood of success are better assessed at the planning and early monitoring stage when corrective measures are possible.

The SC concluded that the self-monitoring of output targets should continue—not linked to the PMS but in the on-line database, CGMap, where MTPs are stored and retrievable and where output target achievement data complements the planning data and is available for internal and external monitoring. This self management role could be reinforced by Program committees of Boards (or their equivalent).

The subsequent challenge was to provide more effective annual indicators for output performance for use in the PMS. Three areas of activity were identified: 1. Publications; 2. Capacity building; and 3. Data management. These three measures of results apply to all Centers and could therefore be developed as a fair measure of important aspects of each Center's results-related performance. Thus the publication indicators were shifted from the old "Quality and Relevance of Current Research" element of the PMS to a new component of the PMS "Indicators of Results". Indicators for capacity strengthening and database management were to be added to the complement of output indicators and their development begun in 2008. The rationale for this change includes:

Publications of research results in a form that can be assessed by peers and used by clients and other peers are the universal hallmark of good science. The CGIAR System is a "mission based" research organisation focused on delivering international public goods (IPGs). That knowledge can have maximum outcomes when it is made available to, and influences the maximum number of users. Precisely because publications are vehicles for transmitting output information, they can be a useful proxy indicator of quality and quantity of the output.

The target audiences for the publication are also important. The CGIAR is a system of excellence in research for development and therefore aims to be both a preferred partner by advanced research institutes and by NARS. For the former Centers need to publish in well recognised journals; for the latter it may be desirable to make the information available in more targeted peer-reviewed venues, such as national and regional publications and specialised books.

The indicator needs to provide incentives for an appropriate publishing strategy and therefore set realistic targets for each of the publication venues. The current output indicator based on publications is a composite indicator. It includes three differently weighted measures with a performance target for each component so as to set incentives for both quality and quantity. The components are:

A. Number of peer-reviewed publications per scientist in 2008 that are published in journals listed in Thomson Scientific/ISI (50% weight). *Rationale: This measure reflects the contribution of knowledge by the Center to a wide international audience and the quality and usefulness of that information as determined by peers from an internationally recognized journal database.*

B. Number of externally peer-reviewed publications per scientist in 2008 (excluding articles published in journals listed in the Thomson Scientific/ISI) (20%). *Rationale: This measure reflects the contribution of peer reviewed knowledge and information by the Center for targeted stakeholder audiences (not including major international journals)*

C. Relative rating of Center's best publications (30%). *Rationale: The CGIAR Centers aim to be Centers of excellence in agricultural science to address complex issues of relevance to the poor. As a system of excellence the CGIAR is more likely to attract new research partners. This measure reflects the quality and originality of the Center's research shown by ability to reach top quality journals with a proportion of all publications.*

Capacity building is an important activity of the CGIAR particularly to enhance likely outcomes from Center activities. Currently the only indicator associated with capacity measures the extent to which Centers publish jointly with developing country partners. This indicator although inadequate by itself could become one component in a composite indicator for performance in capacity building in the future.[8]

Implementation

The publications indicator for components A and B is based on Center self-reported number of publications that qualify for the indicator and the number of scientists in terms of Full-Time Equivalent (FTE).[9] The third component, relative rating of Center's best publications, is based on the journal Impact Factors.[10] The calculation of all indicator components is purely mechanistic and there is no peer assessment involved.

The Centre submissions are verified by an external auditor through a process managed by the CGIAR Secretariat. The practice has varied between all publications being verified

[8] The SC, the ADEs and the Center training focal points began developing a composite indicator in 2008. The activity was placed on hold during the transition in the CGIAR.

[9] FTE is tailor made for each Center to include all internationally recruited staff except those in purely administrative role; regionally recruited scientists; and nationally or regionally recruited staff if the author is the first or only author of the publication.

[10] Journal impact factors are accessible in the Thomson Scientific Science and Social Science Editions that several research institutions subscribe to and otherwise are available at a cost.

or Centers being sampled for verification of one or two indicators. In 2009 the Center-specific verification results were published for the first time. Both this year and in previous years there has been considerable amount of mis-reporting.

The indicator of joint publishing with developing country partners is based on Center self-reporting of the percentage of publications in components A and B (described above) that were published jointly with developing country partners. The results have been verified annually either through a comprehensive process or sampling. The indicator involves no external assessment.

Results

Outputs: Publications

There has been considerable year to year variation with each the Centers in the components of the publications indicator making it difficult to conclude what kind of performance to expect and whether progress has been made over the past years. The results for 2006 and 2007 data are available at the CGIAR PM Web site[11], and for 2007 and 2008 data are presented in the SC's feed-back report to Centers (see footnote 3).

The mean score across all Centers for publishing in internationally recognised journals (Component A—articles in Thomson index journals) increased slightly from 0.84 to 1.14 over the years. About half of the Centers showed rather consistent improvement and four Centers maintained a relatively high publishing rate. None of the Centers has reached the set target level (2 journal articles per staff) and most remain far below it. However, in 2007 and 2008 a reasonable portion of the publications have been published in top quality journals in each field of Center activity (as reflected by 11 and 13 Centers respectively receiving ≥70% of top score in component C). The mean score for publishing in other peer-reviewed publications (Component B) was at its highest (1.29) in 2005 and has since fluctuated. However in 2009 the publication target of one publication per FTE had been met by nine Centers.

The results for the composite indicator can be observed over two years. Ten Centers improved their score in 2009 from the previous year. It is likely that awareness of the performance targets and weights given to different components help the Centers to incorporate these aspects of performance into their publishing strategy.

It would be more appropriate to consider rolling averages over three years for monitoring publishing performance than considering only annual scores. This is because annual fluctuation can be due to factors in the normal cycle of research such as completion of major programs that result in multiple publications simultaneously; differences in peer-review time; and achievement of results (negative results being more difficult to publish than positive results). There is also need to constantly verify all submissions in order to encourage Centers to record their publications more carefully.[12]

[11] http://cgpms.cgiar.org/

[12] The fluctuations in the results over four years may have been partly due to lack of verification of some records in some years.

Outputs: Joint publishing with developing country partners

The results in four years show considerable year to year fluctuation for several Centers. There was no performance target established for this indicator which implies a goal of 100%. This is not realistic or desirable. In 2009, 12 Centers published 40-60% of the publications with a developing country partner, which may represent a more appropriate target range. For this type of performance a diagnostic indicator defining the lower and upper levels of optimal performance would be more appropriate.

Conclusions on output indicators

1. The self- assessment of planned outputs targets achieved as an indicator of scientific performance was not successful. It provided disincentives for creative science and did not (could not easily) accommodate the probing nature of science.
2. A more robust and universally acknowledged indicator of scientific output is peer reviewed publication. The 2008 Independent Review of the CGIAR felt that the PMS had been strongest on monitoring the number of research outputs in terms of publications. The new component score for publications, with clear performance targets for each measure, is intended to provide incentives for Centers to publish their research results in high quality journals, aim a proportion of their publications to the most prestigious journals in each area of research and publish some of their work in high quality books and other peer-reviewed venues for targeted audiences. It measures quantity. It does not measure relevance and quality only directly through relying on the "venue" of the publications.
3. Although the publication indicator covers a major activity that is directly linked to research outputs, it is insufficient in several important areas of output performance. Other essential elements of output that apply to all Centers include capacity building and data (production, management and utility).
4. There is a diversity of activities that take place under the general rubric of capacity building - many of which cannot be simply measured or calculated and many of which require a qualitative assessment rather than quantitative. Thus the development of appropriate and uniformly applicable indicators is very challenging. In the absence of other measures for capacity building the co-publishing indicator has been reported, but as discussed above the interpretation of results in terms of what constitutes good performance is problematic and incentives for Centers therefore unclear.
5. The SC emphasises that data and its management is a crucial activity in CGIAR research, for which appropriate incentives and monitoring need to be set through performance management.
6. The SC considers that a component indicator for measuring any type of output is more appropriate than choice of a single metric. A component indicator (as discuss above for publications) has a better chance of representing and measuring the key aspects or dimensions of the output area being monitored. If there is no algorithm to combine several separate components (as in the earlier situation with multiple publication measures) they appear as having equal value and yet may measure potentially competing activities resulting in contradictory incentives. Managers and researchers may nevertheless find the component information very meaningful for their own analysis.

Outcomes

Design

Outcomes are defined as the use or adoption of program or Center outputs by intended intermediary or ultimate beneficiaries; use or adoption being a necessary step before longer-term impacts can be achieved.

Rationale: *This indicator measures the uptake and use of the research results by the immediate clients. It is a measure of the relevance of the research by the Center and its ability to monitor and document outcomes from the diffusion of research outputs as the first step toward demonstrating impact. It also reflects the effectiveness of activities by the Center to stimulate outcome, such as capacity building and establishment of partnerships.*

The outcome indicator is based on the SC's assessment of case examples of outcomes described by the Centers. The outputs from which the outcomes have derived can be in research or in capacity building. The indicator has remained the same since the launching of the PMS - but its scoring by the SC, and the number of Outcome cases that are required to be submitted, have changed. As is explained later for impacts, demonstrable outcomes take time to develop and successful outcomes may result from several steps subsequent to the original research, including adaptation to local conditions. Therefore outcomes cannot predictably be expected on an annual reporting basis. Due to both the risk involved in research and serendipity, outcomes cannot be predicted with great certainty and in order to use them as basis of an indicator they need to be documented. The SC therefore developed an indicator that integrates the characterization of a certain number of outcomes annually (not all theoretically possible outcomes) and an obligation for Centers to invest in monitoring and documenting outcomes across their research agenda. There is no expectation that all research should lead to outcomes or that Centers should over-invest in outcome studies at the cost of other activities. There is an expectation that the Centers' research results are used and that Centers monitor their main areas of research for uptake of research results as part of their outcome and impact-oriented culture.[13]

The outcome indicator depends both on a) identifying the recommendation domain and intended users of the research results, and b) collecting data to show that adoption and use of research results have been taking place. The former became a mandatory component of Center Medium-Term Plans (which includes planning of the impact pathway). For the latter, the requirement of credible evidence in support of the outcome indicator was expected to create a strong incentive.

In the first four years of the PMS system the Centers were expected to submit five outcome cases each year. In 2008 the SC recommended that the number of outcome cases

[13] The 2008 Independent Review that viewed this indicator only as a measure of outcome achievement considered that annual comparisons had only limited utility due to the long gestation time that outcomes may take to develop.

should be tailored to the size of the Center. Thus in 2009 the requirement changed to correspond with the size of the Center in terms of budget.[14]

Implementation

In the first years of PMS (2005 and 2006 data) SC based its assessment of the cases on very few criteria (3 and 2, respectively) focused on the clarity of the case description. These criteria did not allow good differentiation between cases of different quality, relevance and magnitude, and many cases that formally fulfilled the criteria achieved a full score. In 2008 the assessment was changed to reflect characteristics and importance of the case itself in addition to specificity of the description and the strength of the evidence (current assessment criteria are shown in Annex x). A more elaborate scoring, including a higher number of criteria to take the quality, relevance and attribution aspects into account, was designed to improve the incentives for Centers to monitor and document their outcomes and improve their submissions.

As the outcome case assessment is done by peers (three peers for each case), it is by nature subjective. However the clarity in the guidelines for submission and in the assessment criteria aims to increase the objectivity of the assessment and reduce variability among the peer assessments. The aim has been to draw fully from the SC members' expertise in different subject areas and to make the assessment manageable regarding time (i.e. spreading the load among assessors). This has made it necessary for a few reviewers to read through all cases and provide the necessary calibration across Centers and cases. A final step has been to rationalise the results in cases where there have been large differences in scoring among assessors.

Results

Due to the changes in the assessment criteria, it is not possible to assess the changes in the quality of the outcome cases over the past 4 years. Change in 2008 to the more detailed assessment criteria lowered the level of scoring for all Centers. Only three Centers (CIP, IRRI and IWMI) have reached 70% or more of the maximum score in all four years. If 2008 results are considered a benchmark, the results in 2009 were better in several measures: 9 Centers improved their performance; 7 Centers received a relatively high score (≥70% of top score) as opposed to 4 in 2008 and the overall average score was higher than in 2008 (6.8 vs. 6.2 of a possible 10).

For the Centers that have scored consistently quite high this is likely to be an indication that outcomes are constantly accruing from the Center's research and that the Center takes care in documenting uptake of its research results and subsequently the outcomes. The reason for poor scores may be due to one or both of two factors: a) limited attention to monitoring of outcomes and lack of systematic documentation of the use and adoption of research results that has led to lack of cases to report; and b) lack of clear outcomes. The incentives therefore are to make outcome monitoring and documentation a systematic practice required at the level of major programs and projects, and to take

[14] The requirement ranges from 3 to 8 cases, respectively, for a budget of US$ 10 to 50 million.

steps that make research results and capacity building outputs better suited for or more attractive to their intended users.

The reasons for low scores that are related to the process may include: a) misunderstanding of the criteria and reporting requirements; and b) delegation of the preparation of the cases to inexperienced staff. The common weaknesses of the cases have been that: a) the research results are so recent that there has not been sufficient time for use and adoption of the results to take place; i.e. the outcomes are of pilot nature and predicted rather than well documented; b) the cases are very generic reporting outcomes from research that has been going on for years and outcome are mixed with impacts; c) the cases are of minor activities that do not represent a key area of research (or capacity building) by the Center; d) the cases represent anecdotal reference to Center activities and lack credible linkage between Center outputs and the outcomes. In general, it seems that the benchmarks for each Center are clear enough for them to improve their performance in reporting outcomes. If outcomes were reported in the future moving to three year rolling averages (starting from 2007 data) would be justified to allow better monitoring of longer term trends.

Conclusions on outcome indicator

1. The outcome indicator provides useful results-based assessment that integrates research outputs with other activities still under the control of the Center such as capacity strengthening, partnerships, advocacy etc for plausible impact. It reflects relevance of research as demonstrated by the uptake of the research results and a reasonable measure of what Centers and their partners are responsible for to their investors.
2. It has utility because it conveys qualitative information about the crucial interface of successful research and the use of the results by their intended users. Optimally it stimulates the Centers to identify carefully the intended users and beneficiaries of their research results and orient the research accordingly; determine how technologies and other research output are used; analyse which conditions influence the use and adoption of the results, and determine what the Center's role is in the uptake process *vis a vis* its partners. The best cases that score well in the peer-assessment can be used for demonstrating success (by the Center and *de facto* by the System) and for learning lessons for designing outcome pathways in future programs. The obligation to make outcome reports public and subject to donor consideration in and of itself is likely to improve the care with which outcomes cases are prepared and, more importantly, how research is planned for outcomes and the way in which outcomes are monitored.
3. Since the indicator is based on planning and communication of results by the Center and requires credible evidence of uptake, it is less vulnerable to "gaming" and inflation of scores than was experienced with the achievement of output targets as a measure of performance.
4. There remain however areas of subjectivity. The indicator does not aim to cover the full array of potential outcomes from all Center activity and therefore is subject to some "cherry picking". It is not easy to establish a number or proportion of outputs

from which an outcome could reasonably be expected. [15] Judging by the number of outputs[16] and output targets reported in the recent MTPs and assuming that an output takes 4 years to complete, it can be estimated that about 130 outputs come to an end each year in the CGIAR resulting in some deliverable results—even accepting some failures within the output research. Over 1000 output targets are declared annually, and accepting a level of failure about 50% can be expected to lead to deliverable results. Outcome then depends on the complexity of the environment in which uptake should occur (capacity, socio-economic, policy and biophysical constraints) some of which the Center can address during the research process. The approximately 80 cases currently requested from the Centers represent about 60% of the hypothetical number of outputs completed. However, Centers often report cases related to single output targets that represent a smaller research effort than a multi-year output effort and may be more targeted, and easier to track and attribute. The request for outcome cases therefore is modest, and cherry picking is possible.

5. In the SC's view this hypothetical calculation above illustrates that it is unreasonable to expect outcomes, let alone impacts from each line of research or donor funded special grants, and also unreasonable to expect systematic documentation of each plausible outcome. The indicator therefore cannot be used to establish the "success rate" of all research in terms of outcomes but is more useful for establishing incentives for planning, monitoring and documenting outcomes in a strategically wise manner, which fulfils the requirement of accountability. Part of the strategy should be collection of baseline data that form the basis of outcome (and impact) studies. Use and adoption studies and other impact pathway monitoring can, furthermore, facilitate *ex post* impact documentation and positively affect impact by providing an early feed-back loop to research implementation.
6. The quality of the indicator assessment process requires assessment by a team of peers with a good grounding in the research for development subject matters and uniform understanding of how each criterion is applied. It is a time consuming activity for the peers and not easily transferable. However, removing peer assessment and basing the scoring on fully objective criteria would not be desirable due to the differences in the nature of the cases and subsequent loss of useful feed-back to Centers.
7. An optimal management of the indicator might be using a standard team of reviewers (for example three persons) sufficiently knowledgeable of research and research for development processes that would review the cases jointly thereby securing a similar understanding of the criteria and how to assess them and appropriate calibration of the scoring scale for cases of different nature reporting different areas of research.

[15] The Independent Review's suggestion that each Center present its outcomes within a results-based framework each year for the Science Council's rating would seem to require setting target dates for clearly defined outcomes and documenting those outcomes with an expectation that they are achieved.

[16] In the MTP the intended users and the expected outcomes and impact pathways are described at the output level (lines of research of about 3-5 years expected to produce significant solutions to specific problems).

Impact culture

Design

Context and challenge

For the CGIAR, the key impacts are those on poverty alleviation and food security through the generation of technologies, institutional innovations, policy research and the promotion of sustainable rural development.

In most situations impact is far down the path from the research activity. Typically, productivity related research results (e.g., germplasm improvement) are used by other researchers for adaptation to local conditions then to extension or dissemination agents, then to early adopters who experiment with results, and finally on to more widespread adoption and eventual impacts in terms of the welfare of farmers and/or consumers. In the case of environmental protection, policy research and other less technology focused research, the pathways typically are less straightforward, determined by individual circumstances, and more difficult to predict and trace out *a priori*. Good research performance does not necessarily equate to high levels of impact, since impacts depend on numerous factors outside the control of the research that have entered the impact pathways over time. Herein lies the challenge in devising a System-level impact indicator equably applicable across the different CGIAR Centers engaged in a wide variety of research activities.

Another major challenge in impact assessment relates to attribution. Measures of impact often incorporate the inputs of a great number of agents in addition to those of the researchers and research institutions involved. A performance measure for research institutes must take into account the performance (good or bad) of the other agents involved as well – something that is very difficult to do in most cases. Even where attribution is possible, it may not always be politically desirable to separate the attribution of results to CGIAR Centers from those attributed to NARS or other partners.

The intention of the WGPM was to have an indicator which measures actual impact and, preferably, in a way that allows comparisons across Centers; several donors were keen to have such an indicator. While SPIA/SC recognized the value of such an indicator and the potential for using these results for strategic allocation purposes, it emphasized the immense challenges and the associated dangers in trying to measure and compare actual impact across Center in terms of CGIAR goals. The 15 Centers' research outputs and impact pathways are highly diverse and target different direct and indirect channels of impact and types of impact, e.g., economic, social and environmental, few of which lend them selves to straightforward measurement. Also, the age of the Centers has a substantial influence on the size and extent of realized impacts. The distribution of Centre ages is wide in the CGIAR and this confounds straight impact comparisons. Finally, *ex post* impact performance measure relates to the work and the objectives that existed at the time the research was initiated, not current program performance.

Ever since the launching of the PMS, SPIA has been reluctant to pursue the development of an annual indicator of actual impact for all the above reasons, and has chosen rather to

focus on 'impact culture', which emphasizes commitment to measuring and documenting impacts achieved.

Specifically, the impact culture indicator measures Centers' efforts to document impact from past research (hence *ex post* impact assessment, or epIA[17]) to fulfil their accountability imperative towards CGIAR stakeholders. It also measures their efforts to institutionalize impact culture among their own researchers and partners.

It is emphasized strongly that this indicator does not measure performance of Centers' present research in terms of achieved impacts on CGIAR goals. Nevertheless, good performance with respect to this indicator, i.e., demonstrating strong commitment to documenting impacts *ex post*, is likely to be correlated with 'impact' of a research center.

Evolution

Initially, SPIA developed two impact related indicators (see Annex x): one focused on overall IA performance using four broad criteria (epIA studies; innovation in and advancement of epIA; communication /dissemination and capacity enhancement; and, impact culture) and a second one focused on evaluation of two Center epIA case studies for quality and rigor using seven explicit criteria. Based on SPIA's consultation with the Center IAFPs these impact indicators have undergone annual revisions. SPIA has sought to maintain a balance between change (for improvement) and stability (for consistency). Limiting the extent and degree of changes has been important to allow for consistent measurement and comparison over time. Nevertheless, some major changes have been deemed necessary for improving the usefulness of the indicator and relate to enhancing clarity (terms) and transparency (data requested) and reducing subjectivity (in evaluation). The changes have been: adopting a standard format with explicit scores and weights associated with various components and sub-components (2006), simplifying or revising specific components of impact culture, e.g., giving higher weights to "negative impacts", and lowering benchmarks (2006-2008) and merging the two indicators into a single indicator consisting of three components (2008).

The current composite indicator (re-labelled impact culture) includes the following measures;

1. *Ex-post* Impact Assessment (epIA) studies[18] / advancement of epIA methods (45%);
2. Building an impact assessment culture at the Center, including communication / dissemination and capacity enhancement (20%); and,
3. Quality of submission of one published epIA study during the past three years that

[17] EpIA as defined here is a specialized area of evaluation designed to identify and measure consequences resulting from earlier interventions of a program or project. Its timing is an epIA's defining characteristic: it takes place after the program's or project's investment has generated the intervention, and sufficient time has elapsed and experience has accumulated to assess the intervention's performance in terms of longer term economic, social, and environmental consequences.

[18] An epIA study refers to a published journal article, conference paper, book chapter (but not entire edited book), report or any other publication that has entered the public domain, which is not a revised version of an earlier submission, that documents empirically the impact of a Center's research output in terms of CGIAR goals. The impacts measured may be short-term, medium-term or long-term but must be linked to a clearly discernible intervention derived from research.

effectively demonstrates the impact of the Center's research on the poor or food insecure people and to the environment, as judged by peer reviewers appointed by SPIA (35%).

Implementation

The indicator assessment has evolved more towards Center self-assessment. During the last two years, SPIA has relied on Center self-evaluations for assessing (and scoring) the quality characteristics of the epIAs submitted and accepted under the first component listed above. These submissions (and scores) were verified through the CGIAR Secretariat managed auditing of a sample of submissions. SPIA exercised judgement over which submitted studies would be counted as *bona-fide* epIAs and carefully evaluated the characteristics of each of the legitimate studies based on the summary description provided (see further comment below). The second component (building an IA culture) has been evaluated based on an explicit weighting and scoring method but some subjective judgements by reviewers is required. For example, judgement is required in evaluating components related to establishing benchmark databases (for the counterfactual) and use of epIA in planning and priority setting. Component 3 above (quality of published IA studies) is scored by SPIA but relies heavily on external peer-assessment of the rigour and quality of one selected epIA per year.

SPIA has provided feedback to Centers each year in an effort to improve the quality of submissions in the subsequent year. This particularly included individual feedback on studies that were not considered legitimate epIAs and hence not included in the review and evaluation process. The more general research evaluation-type studies, farmer preference/ demand–type studies, adoption constraint analyses, pilot technology evaluations, and such *ex-ante* assessments do not qualify as epIA for this exercise. Only those studies that document impact (*ex-post*) of Center research have been accepted. While projections in ex-post studies are not uncommon, there has to be adoption and some *ex-post* impact to qualify. Many IAFPs have been in contact with SPIA in preparation for the annual exercise.

The evolution towards a more transparent and standard format for evaluating impact culture, should allow transfer of some components of this indicator to the Centers themselves—with appropriate auditing/verification in place, while other components could be managed or overseen by an external body (for maintaining credibility). Examples of the latter are currently evaluated by SPIA, e.g., assessing whether studies submitted constitute legitimate epIAs (component 1) and assessing the rigour and quality of one 'best' CGIAR Center epIA (component 3).

Results

The results for Impact culture (results for impact indicator A in 2006-2008 and the new composite indicator in 2009) have been published at the CGIAR PMS Web site and summarised in the SC's feed-back reports (see footnote 3). There is a significant improvement in the overall score in 2009 (2008 data) compared to the previous years. This is at least partly explained by the lowering of the benchmark for one of the components. Subsequently all but four Centers received the maximum number of points

for Criterion I.A (20 points), and all but two scored over 18 points.

The results in 2009 also include component 3 that previously was an independent indicator assessed once every three years (assessed previously only once in 2006). In 2009 this component showed the biggest variation among Centers. Many Centers show quite a bit of variability between the assessment in 2006 and 2009. Only two Centers scored consistently high (above 8 out of 10) in both periods and six Centers had consistently below average performance (below 7) in both years.

While average scores were higher in 2009 than in 2008 (7.5 vs. 6.00, respectively), this mainly reflects the lowering of the minimum requirement for number of epIAs produced annually. However, the shift to the lower minimum (one epIA per $20 m of budget) was deemed appropriate in terms of providing the right incentive for targeting large-scale, widely adopted research derived innovations on which epIAs are based. As emphasized during the SPIA-IAFP meeting in Brasilia in November 2008, fewer but higher quality epIAs are preferred over more numerous lower quality (small scale adoption/impact studies) ones. Interestingly, although the benchmark was lower in 2009, a significant number of studies submitted by the Centers—more than 40%—were not considered to be epIAs (and therefore not included in the assessment), a percentage similar to that observed in previous years.

Conclusions on impact culture indicator

1. In conclusion, SPIA believes that with the changes made to the Impact Culture indicator in 2009, especially in clarification of what constitutes an epIA and in consolidation and simplification of components the impact culture indicator is a more accurate reflection of a Center's commitment to documenting impacts.
2. The relative impact culture score received by the Centers each year has been a pretty good reflection of the seriousness of the Centers commitment to doing impact assessments and establishing an impact culture, and in that sense should factor into a donor's assessment of the performance of a Center—with respect to building an impact culture.
3. SPIA/SC recognizes that modest or even major fluctuations in annual scores for any indicator are to be expected—for a variety of reasons—and hence it has advised against placing too much weight on an individual year's score, as this may not be the best measure of 'performance'. For this reason, three-year moving averages have been recommended by the SC when monitoring performance for any indicator. Multiple years of results are also useful for Center management in tracking their own performance over time.
4. SPIA/SC has now added more clarity about what constitutes a legitimate epIA and has introduced a standard template for accurately describing the studies. This should improve the utility of the indicator, and make its implementation easier.
5. SPIA/SC recognizes that achieving development impact is a complex and long term process that involves many actors and participants other than simply research organizations. Nevertheless, every CGIAR Center needs to remain accountable, not for achieving impact from every research initiative, but for measuring impacts from a select number of successful programs, in terms of CGIAR goals or indicators closely

associated with them. Documenting impact is considered essential in demonstrating the efficacy of agricultural research as an effective means of addressing poverty and food security, thus underpinning donor confidence in the CGIAR and agricultural research *per se*.

4. Experience on the utility of the research-based indicators

The SC considered the PMS as one component in an integrated M&E System for the CGIAR Centers that would complement annual planning and the 5-year external program and management reviews (EPMRs). It simultaneously facilitates recording of data and information for the purposes of internal monitoring and external reviews. Observations on the utility of the indicators can therefore be made by looking at how the EPMRs have used the PMS results in their analyses; indications of improvements at Centers in processes (such as planning and record keeping) related to these indicators; and feed-back from Center and donors.

Use of PMS in EPMRs

Since 2006 when the first PMS results would have been available, 11 Centers have had their EPMRs. All EPMR reports mentioned the CGIAR's PMS process and included the PMS results in an annex among the list of materials reviewed.

All EPMRs use publications as a standard metric for monitoring and evaluating research quality and productivity. Six EPMRs referred to PMS publications results in order to compare the Center under review with other Centers. However, EPMRs used other sources of data for program specific publications metrics and a more detailed analysis of the Center publication performance and often included assessment of content, relevance and citation records of publications in their analysis at the Center and program level. In one EPMR the Panel cautioned against splitting valuable research results into multiple publications thereby reducing their potential impact. This would be an unintended result from aiming at large quantity of publications.

Output target achievement was noted in two EPMR reports and the result, in both cases >95% achievement, was taken at face value. Only in three EPMRs were the outcome indicator results mentioned and this was purely for comparison with other Centers. Even where outcome assessment was conducted by the Panel, the PMS results were not linked to that analysis. The 2008 Independent Review considered the accumulated outcome cases an increasingly important source of information for external reviews but the EPMRs have not used them as such despite major emphasis on outcomes in most EPMRs. The impact culture results were mentioned in six EPMRs; two EPMRs discussed the findings in more detail using them not only to compare Centers but to complement their own analysis of impact assessment. In one case this analysis led to a more elaborate discussion about the EPMR results regarding the importance of impact assessment and there is indication that the Center has increased its attention to epIA.

It can be concluded that the use in EPMRs of the PMS results for research has been at a rather superficial level, often reporting comparative numbers because they were

available. Despite the fact that data over 2-3 years were available only to some EPMRs, it is surprising that the EPMRs, albeit only few, used results of a single year to point out differences between Centers as a legitimate comparison.

Some EPMRs raised more generic issues about performance measurement. In the World Agroforestry Center EPMR (conducted in early 2006, when only pilot results were available) the Panel constructed its own performance management agenda of key indicators of institutional effectiveness and problems that it considered would provide timely warning on areas that need managerial attention and intervention. The Panel recognized that each item would require systematic data-gathering, monitoring, and analytical activities. The CIP EPMR panel encouraged Center management to rigorously assess and discuss the research-related indicators as one way of making more of the PMS in management.

The SC agrees that factors that would help make a PMS more effective for improving performance are (i) a high level of ownership of management and proper analysis of what constitutes high performance, coupled with (ii) an analysis of what management interventions are required in response to the noted performance results. With a PMS established externally, there may be a risk of dissociating the need to supply the indicator data from the conduct of analytical activities and management interventions that would be needed for adjustments. A response (funding adjustment) made solely on the basis of the indicator value is completely divorced from the necessary analysis.

One EPMR (of IITA) cautioned against the CGIAR indicators of performance becoming ends in themselves: "Publications may be developed simply to meet performance measures. The opportunity cost to the conduct of research by this type of activity can be high. It is unfortunate that CGIAR stakeholders may misinterpret the real value of indicators and the Panel urges the CGIAR to ensure that information (and investor education if necessary) is made available to ensure that Centers are not penalized inappropriately nor diverted from practical, comprehensive studies of adoption and impact." This echoes the SC's concern of the risks which are associated with performance indicator results having direct influence on funding if the results are also not analyzed in context or their intention and limitations are not properly understood. Such behavior can result in subsequent goal displacement (i.e. a focus on the measure instead of the mission).

The PMS, due to its very character can have only limited utility in an EPMR. This is because the indicators have needed to a) be relatively easy to manage; b) apply equally and fairly to different Centers and c) be by and large acceptable to the Centers considering that they are used for funding decisions. For these reasons only a few aspects of research performance have been covered. The EPMRs that area are focused in the program area particularly on quality, relevance, impact and quality of research management need to go to a much more detailed level, including assessment at the program level that the PMS does not reach as well as to consider a five year time-frame.

Influence of PMS on internal processes

An implicit expectation from the PMS was that it would encourage Centers to record the performance data, such as publications data, more systematically therefore making data retrieval for any purpose more efficient. In many cases this may have happened. However, there are also indications that in some Centers the request for performance information is viewed as a recurrent, externally imposed burden. For example, some Centers do not appear to have regularly updated publications databases that would serve multiple internal and external needs including the PMS. The fact that verification has repeatedly revealed a very high level of misreporting for publications, suggests that the indicator has not led to as careful and systematic recording of publications as could have been hoped for. Furthermore, there are indications that the outcome cases prepared for the PMS are not always known within the Center and therefore not used for multiple purposes.

Another shortcoming of the PMS, particularly regarding publications, outcome cases and impact studies, is that these data, despite being centrally collected, have not become centrally available so that Centers themselves, partners and other stakeholders could see what the CGIAR produces and therefore are likely to go unnoticed. A spill-over benefit from the PMS publications data is that the CGIAR's ICT-KM unit has been able to conduct analysis of access and availability that without the comprehensive lists of peer-reviewed publications from all Centers would not have been easy. However, for anyone to access these PMS data, there is a need for an authorised individual to cut and paste or re-write the data scattered under individual years and Centers in the PMS system, which is highly inefficient.

Regarding research planning, there are indications that a number of Centers have internalised the impact pathway thinking and monitor and record outputs and outcomes systematically. The feed-back on the SC's peer-assessment of outcomes and impact culture and interaction with the ADE and Center IAFPs may have stimulated some Centers to more carefully monitor, identify and document diffusion of their research results, outcomes and impact and establish processes that enhance impact culture. The better the performance measures are linked to the institute's organizational mission the higher the credibility in the eyes of those whose performance is being measured and the motivation to perform toward the mission. Despite the aim at comparable measures, the extent to which the common PMS has sufficiently reflected the missions that the Centers within their mandates pursue is not clear.

SPIA believes that the PMS exercise has been helpful in encouraging the Centers to do more large scale impact assessments and in providing the right kind of incentives for moving further down the impact pathways in their studies.

External use of indicator information

Collectively the best outcome and impact cases represent success stories from the CGIAR SPIA/SC in collaboration with the Centers has published the best impact cases that have been much appreciated also by the donors. The SC intends to do the same for the best outcomes cases in 2009. As mentioned above, results related data are not readily available or in a useful form for external use.

It is not known whether other donors than the World Bank, Germany and Italy have used the PMS results directly for decisions related to funding.[19] There is anecdotal evidence that some donors have misinterpreted the indicators on impact culture as reflecting actual impact and adjusted their funding on the basis of this interpretation of the results.

5. General comment on the objectives and use of performance indicators

The SC has emphasised that the PMS (currently applied to the Centers), composed of a set of performance indicators, should be one of the components in a broader set of M&E tools. [20] It was recognised that Centers were increasingly under a burden to respond to different reporting and review requirements deriving from internal, the CGIAR's, and individual donors' needs, and therefore the SC emphasised that the different M&E components should complement each other in an integrated way. The Centers have provided very critical comments emphasising the aspects of burden, clarity in the interpretation of the indicators and how they are used by donors, and the need for comparing a Center performance against its own earlier performance for monitoring improvement.

To the extent that indicators used in the PMS are able to reflect actual performance, this information can have value for decision making, in particular for Centers' managers. Measuring within Center performance, over time, and benchmarking against what can be expected from high-performing organisations is an essential of good practice in the management of a Research Center provided that the indicators reflect the most important aspects of performance. The principle of rewarding good performance with increased allocations of resources, as practiced in some measure by several key donors in the CGIAR is essentially a sound one as it is a powerful and effective incentive for encouraging good performance across Centers; requiring, however, that there is common understanding of what constitutes good performance and there is a consistent interpretation of the indicators used.

However, as with any instrument, there are always risks of unintended and undesirable effects from donors' inappropriate use of performance indicators. It was acknowledged at the start that development of an established set of performance indicators could take a long time and that there would be need for revisions. During the five years of the PMS, the indicators have been revised from time to time on the basis of experience to provide for clearer incentives for good performance. Despite these revisions, there are still questions about how effective the indicators have been in terms of generating the incentives they were intended to generate at management level, how comprehensively they have covered important aspects of results-related performance, whether they have

[19] The World Bank has used all the indicators, and increased over the years the weight given to the results-based indicators. Germany chose six indicators that in 2008 included the two impact culture related indicators, but was going to re-consider its use of the PMS as basis of funding decisions.

[20] Until now the other tools have included assessment of Center plans (Medium-Term Plans) for relevance of research, EPMRs and Center Commissioned External Reviews.

effectively complemented other M&E and how closely the specific indicator results have reflected the performance that they are proxies for.

Given experiences in other settings, a particular risk that needs to be acknowledged in the context of the CGIAR is the establishment and use of a link between indicator-based performance and resource allocation. It must be emphasized that this should not be made in a mechanical way. There are three reasons why care should be exercised in making a direct or overly simple link between reported performance and resource allocation:

a. Without adequate validation procedures in place to ensure accurate reporting, there will be a tendency to over-state performance/impact in order to get more resources, which in turn could distort Center activities;
b. Some Centers having performance problems may actually require temporary support to bring them to a higher level of performance. Penalizing them by reducing financial support could in fact exacerbate the underlying performance problems. Hence, provision should be made under some cases for allowing poorly performing Centers to submit an explanation of, and an action plan for dealing with, its performance deficiencies prior to decisions taken about resource allocation. Furthermore, the PMS does not allow differentiation between high performing and poorly performing units where targeted measures or support would be needed to rectify the situation.
c. Aiming for high values in the indicators included may reduce attention to achieving high levels of performance in areas that do not easily lend themselves to being measured through annual indicators, such as relevance and long term actual impact.

Annual performance indicators should therefore inform but not constitute the sole basis for funding decisions by donors; they do not provide a comprehensive assessment of all performance. They may complement others that provide information on Center/program performance over time, particularly on relevance, quality and quantity of results of different kinds and their potential for impact. These should also factor into the overall assessment, especially given the heterogeneity of research domains and challenges faced by the fifteen Centers. This heterogeneity will not be reduced in Mega-programs. Furthermore, considerations like temporary needs, current research proposals, emerging challenges (climate change, increased food prices, etc) should also be considered in funding decisions.

To some extent the experiences with the PMS demonstrate the incompatibility of mixing goals. A primarily internal process initiated by the Centers might have allowed the Centers to more easily learn from each other, identify the areas where more accountability and effectiveness could be achieved and it might have been more highly valued and thoughtfully pursued by all Centers. Such a goal seems to have been

antagonistic to a process that directly influences access to resources; the rewards to rigor and self-scrutiny became less clear.[21]

6. Application of PMS in new CGIAR

The SC has already presented some thoughts of how performance measurement may be organised in the new CGIAR.[22] A very important consideration is that performance measures stimulate high performance and do not stifle ambitious and novel research that can have high impact if successful. Some of the current results-based indicators may be more suitable for Center performance management and could be more effectively geared to provide a specific management tool towards improving processes, practices and rigor of self-assessment at Centers.

The SC has advocated for the use of integrated M&E approaches. For example, indicators that themselves involve peer assessment could contribute to assessment of planning and internal monitoring of different aspects of performance and be complemented by external reviews. The combined information from these different tools would be used to improve performance through incentives and corrective measures, to steer Center or program activities and provide rewards. While the information from several sources cannot be integrated into a simple metric it can be informative to management and used by an independent science council to advise donors.

The current PMS was not designed for use by programs. The design of the monitoring and evaluation of Mega-programs should not be restricted by an aim to have universally applicable indicators (as the Center PM indicators were). The SC has argued[23] that program performance indicators should be tailor-made for each program and adjusted for the phase of the program. The indicators should reflect the specific characteristics and expectations regarding research processes and results. Indicators for key performance aspects should be part of an integrated M&E system to allow early intervention for enhanced performance. Monitoring may need to include incremental research advances and evidence of plausibility in addition to concrete deliverable research results. The SC's view is that peer assessment for qualitative assessment of certain indicators will be needed (for example for outcomes) rather than a "mechanistic" scoring that reduces the opportunity for analysis of the performance results and may encourage gaming.

Some of the current set of indicators (used or planned as discussed below) may be better suited for Centers and others to mega-programs. Particularly output performance may need to include many other aspects than publications to be considered for the Mega-programs individually. As concluded by Independent Review, conceptualization of other

[21] This experience is also echoed in the 2008 Independent Review that concluded that the PMS was not well positioned as a learning tool for the Centers because it was difficult for the one instrument to play three divergent roles—accountability, resource allocation, and learning.

[22] http://www.sciencecouncil.cgiar.org/fileadmin/user_upload/sciencecouncil/Highlights/ME_20in_20the_20new_20CGIAR_20revised_20June_201_202009.pdf

[23] Performance Measurement System for the Challenge Programs, SC Secretariat 2008, internal document.

significant outputs is needed and performance monitoring should give strong incentives to those who implement research to make their results available and useful for development.

Publications are an important measure of scientific productivity at both levels and can provide assurance of certain quality standard to staff and managers alike. The Centers need to be able to link basic research to development applications and compete and partner with other research institutions outside the CGIAR. In the new CGIAR the Centers need to maintain credibility and visibility for their research results that comes through peer-reviewed publishing. On the other hand, the Mega-programs need to generate research-based advice to policy makers and research-based technologies and solutions for development problems. In large part, publishing research results in peer-reviewed publications is the most effective way of disseminating them broadly. For these reasons and for documenting outputs tracking publications and maintaining incentives for reasonable level of high quality publishing remains important. As publishing is commonly monitored both at program and staff levels it seems sensible to monitor publication both at Center and at mega-program level.

It is likely that capacity building is going to feature explicitly in Mega-program implementation and may be reflected in performance contracts. Capacity building through partnerships (not just training) is likely to be an integral part of research implementation and appropriate measures for monitoring results will be needed. Such measures will need to be tailor-made to the nature of the mega-program and to its state in a normal cycle of early innovation, implementation, extrapolation, etc. In the SC view new indicators of effective partnerships and of capacity strengthening need to be developed for the mega-programs.

Data and its public access can be considered one of the CGIAR's most valuable outputs of a true international public good nature provided that it is managed appropriately, shared and exploited fully. Collectively the Centers have large and important data bases in all areas of research. In a stock-taking report by the Alliance in 2008[24], it is concluded that although data sharing has become a central part of global research systems, in many cases the Centers have not adapted adequately to this way of working. This has resulted in a failure to realise much of the value from research data. The problem can be traced to a failure of the system to appreciate the intrinsic value of data, reward data sharing and the high quality management that it requires, and to change out-dated attitudes to ownership and intellectual property. The main responsibility for data may reside at the Centers that accumulate and store long-term data sets. The incentives and accountability for proper data management and exploitation must be aligned with System level policies and may best be established at Center level.

In Mega-programs the move from grant-based funding to longer-term program planning and funding puts more emphasis on explicit *ex ante* planning for outcomes that become a

[24] Improving Research Data Management and Sharing in the Alliance of CGIAR Centres. A Working Paper for Consideration of the CGIAR Alliance, October 2008.

collective responsibility of the Centers and the partners.[25] Performance contracts are therefore likely to include outcomes as one element. The experiences from the outcome indicator in the current PMS can be useful. The incentives should favour novel and even risky research that, if successful, has high impact expectations. Therefore the indicator(s) for measuring outcomes should be adjusted for each program (and, if necessary, for program components in large programs) to enhance incentives for good performance. The SC considers that appropriate outcome indicators are useful for monitoring of relevance of research as proven by its uptake among the intended users. Performance regarding results, outcomes in particular, should be measured at intervals of a few years rather than annually.

The SC feels that impact culture is an important aspect of performance that should be maintained, possibly both at Center and Mega-program levels. *Ex post* impact assessment of sufficiently large research interventions sufficiently long time after research has been completed remains essential. The issue of measuring actual impacts through an indicator may remain pertinent also in the new CGIAR. However, demonstrating long-term *ex post* impact will unlikely be the mega-programs' responsibility. A recent consensus view among SPIA and the IAFPs, following their consideration of alternative ways of creating an indicator for actual impacts of Centers, was that they would be excessively data and time intensive and require significant resources for execution. Also, given the heterogeneity in the type of research done by different Centers, coming up with a single method or consistent set of indicators of actual impact applicable to all Centers did not seem feasible. However, the alternatives discussed could have merit in the new CGIAR: (i) requesting a more systematic conduct of periodic meta-analyses of the Centers' cumulative success stories in an attempt to compile and quantify the size and nature of the economic and non-economic impacts over time; (ii) periodic impact assessment of Centers based on in-depth analysis of "impact claims" (a case study approach); and (iii) System-level meta-analysis on an annual basis. For (i) and (ii) methods to value and possibly compare these impacts among Centres would need to be developed. [26]

There are important areas of performance related to good planning and monitoring for impact that transcend the current PMS but should be explicitly emphasised in program management and monitoring. The path from research to impacts is seldom linear and researchers need to be informed by and, as appropriate, engaged in the iterative processes, for example through participatory research. Research planning for impact includes: engaging in sound processes that assess the extent and scope of the problem; undertaking an analysis of alternative suppliers and demand for research derived information; in impact pathway design considering constraining factors, capacity building needs, particular effects and constraints for women, and opportunities for

[25] Elaborate impact pathways and partnership arrangements will need to be thought through and specified.

[26] With reference to the Independent Review's conclusion that actual impacts need to be assessed and that a measure of actual impact therefore is needed as part of the PMS, SPIA/SC emphasises that impact assessment is currently an important part of every Center's and the CGIAR's M&E. Development of an indicator that would systematically quantify impacts at fixed periods is a different matter; one that could be pursued through these periodic meta-analyses.

participatory research; partnering with the most relevant collaborators both for research and development. These are means to embed research in an environment where the likelihood of impacts increases. These kinds of aspects are not easily quantified, but should be included in a sophisticated monitoring model that could apply to the Mega-programs.