Article

Evaluation
2014, Vol. 20(1) 96–114
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1356389013517444
evi.sagepub.com
SAGE

Using performance measurement to assess research: Lessons learned from the international agricultural research centres

Sirkka Immonen
CGIAR Independent Evaluation Arrangement, Italy

Leslie L. Cooksy
Sierra Health Foundation, USA

Abstract

In the past few decades performance measurement systems have been employed in the public sector to enforce accountability and enhance the efficiency and effectiveness of operations. Performance measurement systems as part of research evaluation are a special case. In this article, the suitability of performance measurement for the evaluation of research is analysed, drawing on experiences from a group of 15 international agricultural research centres that conduct research with a development mission. The performance measurement system initiated by donors was intended to become part of a streamlined monitoring and evaluation system and to enhance transparency, accountability, learning, and decision making, including decisions about future funding. The experiences showed that: (i) there were large year-to-year fluctuations that were probably related to the selected indicators rather than actual performance, therefore undermining annual conclusions about performance; (ii) using the indicator information for resource allocation influenced performance reporting and emphasized comparison between centres; and (iii) performance measurement information was not used efficiently in other evaluations. The article examines lessons drawn from the objectives, expectations and results of the performance measurement exercise for planning monitoring systems for research in multi-partner settings.

Keywords

agricultural research, evaluation, indicators, performance measurement, research for development

Corresponding author:
Sirkka Immonen, CGIAR Independent Evaluation Arrangement, c/o FAO, NRDE, Viale delle Terme di Caracalla, 00153 Rome, Italy.
Email: Sirkka.Immonen@fao.org

Introduction

Research intended to produce solutions to the problems of developing nations, known as *research for development*, takes place in complex environments. Its outcomes depend on multiple events and actors along the impact pathway, in a context of political, institutional and biophysical constraints. Evaluation of such development-oriented research is also challenging. Several strategies have been used, including peer review, institutional evaluations, historical tracing, participatory outcome mapping, econometric analyses and experimental or quasi-experimental designs (Guena and Martin, 2003; Ruegg and Jordan, 2007). In addition, some organizations have put performance measurement (PM) systems in place to monitor various aspects of research (Guena and Martin, 2003; Organisation for Economic Co-operation and Development [OECD], n.d.). However, PM systems have been subjected to limited systematic assessment and have received little scrutiny when applied to research. This article presents an assessment of the development and use of PM in the research-for-development context, as applied in the system of international agricultural research centres (IARCs).

Overview of PM

With the overarching aim of increasing the efficiency and effectiveness of public management and policy making, PM systems are intended to track performance over time, using, among others, indicators selected to provide feedback and establish accountability (Hatry, 2006; Van Thiel and Leeuw, 2002). PM started spreading in the public sector more than three decades ago under 'new public management' and is part of the results-based management approach (Van Dooren and Thijs, 2010). By mandating performance-based quantitative measurement of programme outputs and outcomes, the United States' 1994 Government Performance and Results Act (GPRA) was an early spur to the spread of PM systems (Natsios, 2010). Hood et al. (2009) suggest three reasons for the continued use of performance indicators:

- the development of information technology that spans the worlds of consultancy, academia and government agencies;
- the appeal to some public managers of apparently 'transparent' steering processes, such as rating and ranking; and
- the appeal to politicians of seemingly 'objective' systems that provide easily understood data on performance to their constituents.

Public research organizations have also adopted PM. For example, the US Agency for International Development requires annual reporting on a set of quantitative indicators for all 'Feed the Future' funded programmes, including those focused on research (http://www.feedthefuture.gov/progress). A large research organization, the Brazilian Corporation for Agricultural Research, EMBRAPA, has had an annual PM system in place since 1996. Its PM system was designed internally, overseen by its Board, to inform management (Avila et al., 2008; Da Silva e Souza et al., 1999). Universities and university funders in many countries have used PM to assess research productivity (Guena and Martin, 2003).

Potential benefits and pitfalls of PM

The value of PM systems has been debated. Monitoring key indicators over time is considered to provide needed information for accountability, inform management decisions, and potentially serve as an early-warning system, which would prompt questions rather than provide guidance on what should be done (Hatry, 2006; Holzer and Kloby, 2005). It is seen as directing attention to the desired outcomes of a programme or organization. The value of PM data as a foundation for programme evaluation has also been noted (De Lancer Julnes, 2006). Guena and Martin (2003) indicate that PM, in the research context, both provides accountability to funders and encourages dissemination of research results. Lepori and Reale (2012) propose that 'science and technology indicators' can have their most important role in formative evaluation contributing to a socially constructed discourse rather than in a top-down process ordered by the programme funder.

Several authors have recommended caution in the use of PM systems, and many consider it a trend, the effects of which have not been well assessed (Hood et al., 2009; Jann and Jantz, 2008; Poister, 2010; Radin, 2011; Van Thiel and Leeuw, 2002; Winston, 1999). Specific concerns include resources required (including time) and such unintended consequences as goal displacement – a shift of attention and resources away from the mission to those aspects of performance that are measured. Goal displacement is exacerbated when programmes or institutions are compared with each other for their results and rewarded for higher results (Davies, 2006). Natsios (2010) suggests that the desire in aid agencies to measure success is shifting attention to programmes that are most easily measured, but are least transformational. Linking expected performance rigidly to planned targets can inhibit innovation and reduce responsiveness to emerging opportunities (a phenomenon termed ossification by Smith, 1995).

PM systems have also been characterized as reductionist and rooted in a 'top-down hierarchical "control" model' (Perrin, 1998). Specific concerns include a focus on control over organizational learning (Jann and Jantz, 2008), the implied declining trust in professionalism as a basis for accountability (Davies and Lampel, 1998) and an emphasis on competition, regulation and supervision. PM is also considered as a subjective, value-laden concept that takes place in a political context (Jann and Jantz, 2008; Thomas, 2003). Finally, there is little evidence that performance information has been used in management decisions and improvements (Ammons and Rivenbark, 2008; De Lancer Julnes, 2006; Hood et al., 2009; McDavid and Huse, 2012; Moynihan, 2008). As Poister (2010) emphasizes, programme managers need to be confident about the programme logic connecting activities to outcomes to interpret performance monitoring information appropriately.

In the research context, the exploratory nature of research has been considered contradictory to the expectations of PM, which is usually based on the assumption that outputs will be regularly produced in a predictable timeframe and that causality from activities to outcomes is warranted (De Lancer Julnes, 2006). Smith (1995) and also Hatry (2006), who has generally promoted PM, consider PM unsuitable for activities whose outcomes may take a long time to occur. Cozzens (1997) considers that research performance is best captured by descriptive and sophisticated analyses appropriate for peer review panels but not a good fit with PM where indicators are partial, capturing some aspects and not others of the phenomenon of interest.

Other aspects of research not captured well in PM systems are research quality and intellectual influence (Butler, 2003, 2007; Feller, 2002; Moed, 2007; Perrin, 1998, 2002). Butler's (2003) analysis of Australian research publication performance over two decades concludes that rewarding publication quantity appeared to have altered publishing habits thus contributing to a considerable reduction in the citation impact of journal publications. Moed (2007) suggests that in research quality evaluation, publication counts and journal impact factors should probably play no role at all, but instead research evaluation should combine sophisticated citation impact indicators and transparent peer review. Feller (2002) notes that PM systems in research can affect the distribution of authority and influence within an organization, as well as the forms of evidence deemed legitimate in decision making. He cautions against aggregating performance data to higher levels of decision making when there are few opportunities to challenge the data or the related analyses. As the processes through which research outputs, outcomes and impacts are generated are complex and largely dependent on external environment, performance data are easily misinterpreted; a dysfunctionality identified by Smith (1995).

In 1999, Winston called for PM systems to be assessed across a range of programmes and organizational settings to determine success and failure factors and factors that lead to unintended outcomes from such systems. Although considerable progress has been made in the assessment of PM in general in terms of its design, its implementation and assessment have received limited attention (Van Helden et al., 2012). There continues to be little research on the effects of PM on the quality, strategic orientation and effectiveness of research.

Objectives and context of this study

This article aims to contribute to filling the gap in information about the nature and effects of PM in research organizations, with particular attention to research for development. We present a critical analysis of one PM system in the framework of potential positives and pitfalls identified in the literature, and use that analysis as the basis for recommendations for monitoring of research.

The context of the study is the Consultative Group on International Agricultural Research (CGIAR), a global partnership that supports 15 international agricultural research centres (IARCs). The mission-oriented centres engage in agricultural research on commodities, policies, systems and natural resources for addressing development problems. The CGIAR's high-level impact goals are to reduce poverty, hunger and malnutrition, and enhance the sustainability of natural resources. During the time discussed in this article, an independent advisory body, the Science Council, was responsible for monitoring the relevance and quality of IARC research and conducting evaluations. The CGIAR's specific PM arrangement analysed here ended in 2010 when the CGIAR was restructured (CGIAR, 2009).

Various strategies have been used by the CGIAR for evaluating research (CGIAR Science Council, 2005; Özgediz, 1999). Annual reviews of IARC three-year rolling Medium-Term Plans (MTP) by the Science Council were intended to monitor the relevance of planned research and provide formative feedback. Each centre was subject to a periodic comprehensive review by an external, independent panel of peers (the external programme and management review). Impact studies of research areas were organized by an independent, CGIAR-associated body (Kelley et al., 2008). In addition, centres have conducted evaluations and impact studies as part of self-assessment for internal management purposes, and have reported to and been the subject of reviews by individual donors.

The article uses the CGIAR's PM experience to illustrate the challenges of evaluating research performance through annual measurement of indicator data. Detailed descriptions of the PM system (operational in 2005–10) have been published elsewhere (CGIAR, 2003, 2004; CGIAR Science Council, 2009a; CGIAR Science Council and CGIAR Secretariat, 2009).

Design of the CGIAR PM system

Drivers and objectives of the PM system

The PM system was introduced to the CGIAR as a donor initiative. The World Bank's justification for launching a performance-based approach to fund allocation was to improve the effectiveness of its global programmes following what was seen as a global trend towards improving efficiency and demonstrating accountability for results (Iskandarani and Reifschneider, 2008). Already in 2003 the World Bank allocated 12.5 percent of its CGIAR funding on the basis of a temporary set of indicators it had negotiated with the centre directors (CGIAR Science Council, 2004). Financial problems in some centres provided another motive to have regular 'early warning' reporting directly to donors (CGIAR, 2003). It was also hoped that a common, annual PM system would decrease the multiple parallel reporting requirements by individual donors. The Science Council, although participating in the design of the PM system, particularly for research performance, took a rather critical stance of the process, acknowledging the difficulties to measure progress towards achieving research goals. The Council's chairman stated: 'attempts to focus performance measurement on what can be counted and compared across programmes or centres tend to give misleading results, because those things that can be counted may not be good indicators of progress towards achieving impact on poverty' (Pinstrup-Andersen, 2004).

A Working Group on Performance Measurement (WGPM), facilitated by the World Bank, identified four objectives for the PM system: aid funding allocations, provide benchmarks and targets, demonstrate accountability, and inform decision making and management (CGIAR, 2004).

Main features of the PM system

In designing the CGIAR PM system, the Brazilian system (EMBRAPA) was considered one model (Beintema et al., 2010). An important difference was that EMPRAPA's PM was designed as an internal management process and not used for fund allocation. Furthermore, unlike EMBRAPA, the CGIAR centres are autonomous institutions and thus the PM was imposed as an external management tool.

The expected use of the PM data, at least by the World Bank, for allocating funds across the IARCs influenced the PM design. The performance measures had to apply equally to all the centres irrespective of their different disciplinary activities, modes of operation and partnership arrangements, and be as unambiguous and fair as possible when used for categorization of centres by their performance (CGIAR Science Council, 2009a). Although it was acknowledged that the initial indicators would need improvement over time (CGIAR, 2004), the WGPM chose not to field-test a small number of measures first. Instead, the system was designed to capture the entire chain of causality from resources and inputs to outputs, outcomes and impacts. The indicators confirmed in 2006, following a pilot year, and the changes applied in 2009–10 are shown in Table 1.

The most important design issues relate to measurement of outputs, outcomes and impact, and the setting of performance targets. To emphasize the mission-orientation of the CGIAR centres, the initial output indicator was a measure of centre achievement of annual output targets. Output targets, contributing to broader objectives, specified the deliverable achievements of different kinds expected to be completed in the year of reporting. The measure was

Table 1. The CGIAR PM indicators as of 2010 and changes from the first year of full implementation (discarded indicators shown in italics).

I. Indicators of Results	
Element 1: Outputs	• Indicator 1: Composite measure of centre research publications • Indicator 2: Percentage of scientific papers published with developing country partners • *% Medium-Term Plan output targets achieved*
Element 2: Outcome	• Indicator 3: Science Council assessment of Centre outcome reports
Element 3: Impact culture	• Indicator 4: Composite indicator of centre impact assessment culture
Impacts	• *Science Council rating of Overall Impact Assessment Performance* • *Science Council rating of two Centre impact studies done in the period 2003–05 for rigor*
II. Indicators of Potential to Perform	
Element 4: Quality and relevance of current research	• *Number of peer-reviewed publications per scientist* • *Number of peer-reviewed publications per scientist that are published in journals listed in Thomson Scientific/ISI* • *Percentage of scientific papers per scientist that are published with developing country partners*
Element 5: Institutional health	• Indicator 5A: Summary score on governance check list • Indicator 5B: Assessment of Board statements • Indicator 5C: Summary score on culture of learning and change checklist • Indicator 5D: Percentage of women in management • Indicator 5E: IRS Nationality concentration • *26 separate indicators under 5A* • *11 separate indicators under 5C* • *Gender diversity goals: Does your Centre have Board approved gender diversity goals?* • *Diversity in recency of PhDs*
Element 6: Financial health	• Indicator 6A: Long-term financial stability (adequacy of reserves) • Indicator 6B: Cash management on restricted operations • *Short term solvency (liquidity)* • *Efficiency of Operations (indirect cost ratio)*
III. Stakeholder Perceptions (survey every three years)	

intended to both respect the diversity of the IARCs' missions while being equally applicable to all centres irrespective of their research orientation. However, the very high achievement rates and an implicit expectation of 100 percent success in research revealed faults in this indicator and, following the CGIAR Science Council recommendation (2008a), it was dropped. Subsequently, falling back on traditional measures of scientific outputs, publications were reclassified as outputs. While an important product from research, they had not been considered to reflect sufficiently the research-for-development orientation of the CGIAR. Other measures of outputs common to all centres, such as capacity building and management of data as international public goods, were discussed but these activities did not lend themselves to designing credible measures that would have set appropriate incentives and captured the essential elements of high performance.

The indicator for outcomes, defined as external use, adoption or influence of IARC's outputs, was derived from peer review and scoring of a set number of outcome cases clearly attributable to a centre's research. Acknowledging that not all successful research leads to outcomes, the indicator was intended to integrate a measure of the centres' diligence in documenting outcomes across their research portfolio and a measure of significant outcomes derived from a part of IARC activities.

It was acknowledged that an annual indicator for impact would not be able to measure impact due to various factors, such as the long time lag from research to development impacts, challenges of attribution back to research, and the lack of research organizations' control over impact (CGIAR Science Council, 2009a). A composite indicator that focused on impact culture evolved to combine various parameters of impact culture and rigour of impact assessment with weighted elements of quantitative and qualitative data. The intention was to encourage IARCs' efforts to document impact from past research and to institutionalize impact culture among their researchers and partners.

The PM system was heavily skewed towards institutional indicators grouped as 'Potential to perform'. The 'Institutional health' element originally included numerous independent sub-indicators that were purposefully not weighted and most indicators initially had no benchmarks or targets. Interpretation of the results was left open to individual donors, which left the IARCs vulnerable to funding shifts with a high level of ambiguity. Therefore, in 2009 benchmarks and summary scores were adopted for some groups of indicators.

PM results and their use

Trends in PM data over years

Selected results across several years are presented here to illustrate the utility of the PM system. Table 2 shows annual averages for the main indicators across the centres for the years 2006–10 (for data of the previous year).

The five-year trend analysis reveals little change for the key indicator results averaged across the 15 IARCs. The 'governance' and 'culture of learning' checklist summary scores increased in the final year likely reflecting adoption of certain desirable policies.

At the individual centre level, considerable fluctuations occurred in some indicators, resulting in large shifts in IARC's annual ranking on the various measures. Figure 1 illustrates fluctuation of the five-year results for five centres with the most variable trends. For the others, little steady change in performance was seen except for the financial indicators where centres tended to stay within or get closer to the targets defined. As several modifications had taken place it cannot be determined to what extent the trends represent changes in centre reporting and score calculation rather than in performance. Notwithstanding the fluctuations and the acknowledged need to learn, and adjust the system, the results were presented to the donors annually for their use at face value.

The results for the initial output indicator illustrate a particular challenge: where to set the success expectation of research that is inherently risky. Expectation of 100 percent success clearly carries perverse incentives. Evidence of that in the CGIAR PM system is shown in Figure 2. The '% output targets achieved' reported by the IARCs decreased when the measure was no longer included among indicators for fund allocation, although the data were still collected (years 2009–10). The level of 'full achievement' dropped for 13 of the 15 centres.

Table 2. Average PM indicator results over five years across IARCs.

Indicator	Year of Reporting (of previous year's performance)				
	2006	2007	2008	2009	2010
Results					
Publications (composite indicator)	–	–	5.9[a]	6.2	6.1
Peer-reviewed publications (non-ISI journals)	1.3	1.0	1.2	1.2	1.1
Peer-reviewed publications (ISI journals)	0.8	1.0	1.1	1.1	1.2
Co-publishing with developing country partners (%)	46.4	42.4	45.4	46.6	47.1
% achievement of output targets	*87.1*	*88.1*	*89.2*	*(63.9)*	*(75)*
Outcome (adjusted to 10 scale)[b]	8.1	7.6	6.2	6.8	7.6
Impact culture (adjusted to 10 scale)	6.0	6.4	5.8	7.5	6.8
Potential to perform					
Governance checklist score (%)	–	–	–	85.5	88.0
Peer assessment of Board statements (adjusted to 4.5 scale)[c]	3.2	3.1	3.5	3.0	3.0
Culture of learning checklist score (%)	–	–	–	50.6	54.3
Women in management (%)	19.1	26.6	27.8	24.7	24.8
Nationality prevalence (1st most common)	17.3	18.0	15.7	15.7	16.3
Nationality prevalence (2nd most common)	10.6	10.4	11.4	10.2	10.7
Long-term financial stability (75–90 days)[d]	120.6	115.3	105.5	117.6	118.0
Cash management on restricted operations (benchmark <1.0)	0.9	0.6	0.5	0.5	0.2

[a] Piloted in 2008
[b] Change in scale (3 in 2006, 2 in 2007, 10 in 2008–10)
[c] Change in scale (9 in 2006–07, 4.5 in 2008–10)
[d] Excludes one centre that maintained very high reserves

The lessons from the PM system are discussed below according to the WGPM's four intended uses of the PM data.

Determining funding allocations

Application of the indicator results for annual funding decisions was the most controversial use of the PM data. Although the system was intended to increase transparency about decision making and funding (CGIAR, 2003; Iskandarani and Reifschneider, 2008), the IARCs consistently argued against using PM data for fund-allocation (Alliance Deputy Executive, 2007, 2008). Also the Science Council advised against using the PM data for direct and mechanistic funding decisions (CGIAR Science Council, 2009b).

The proportion of the World Bank funding allocated on the basis of the PM results rose from 25 percent to 50 percent in the first four years as the PM system was considered to be

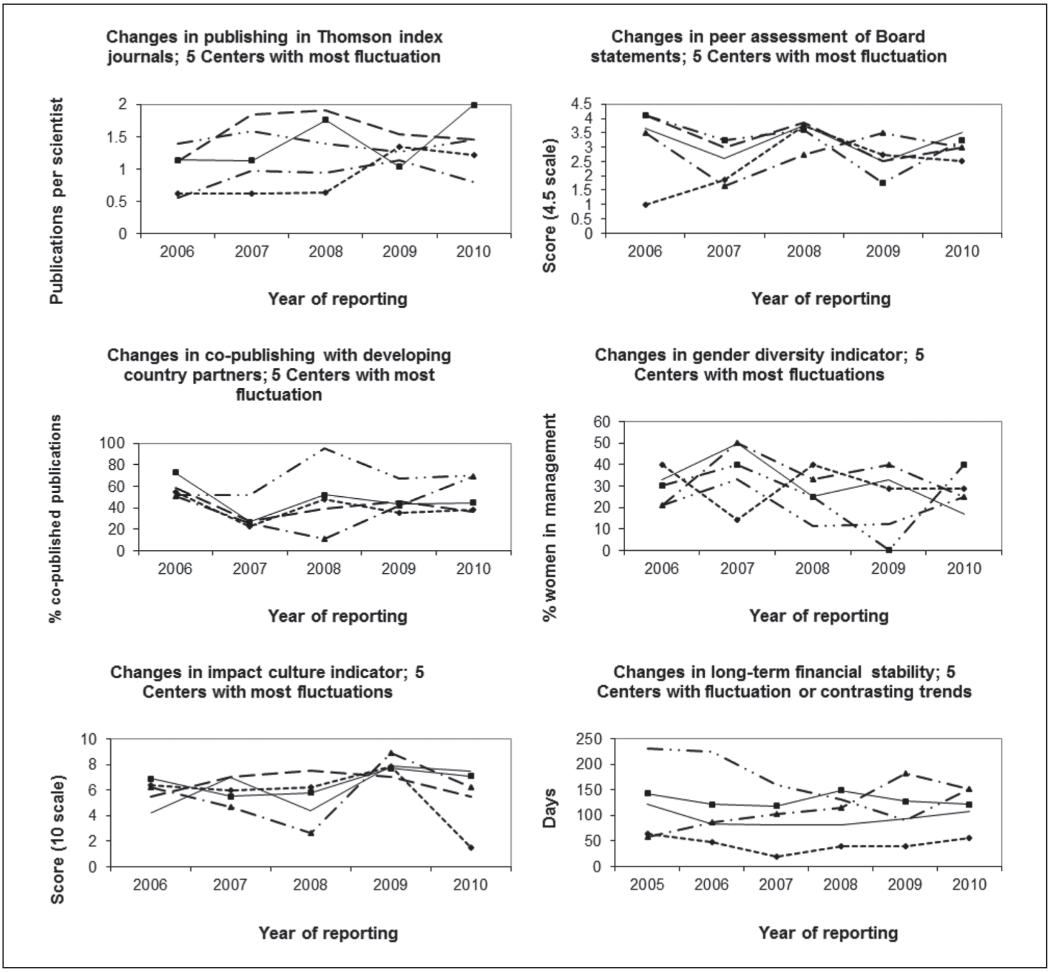

Figure 1. Trends for five centres with most variability in the results of six selected PM indicators over 5 years.

Pilot year data reported in 2005 are included for 'long-term financial stability' indicator.

maturing with robust data. The weights assigned to each PM criterion were not consensus weights, a decision that was criticized by the CGIAR Independent Review Panel (2008). Centres, in addition to being grouped into three funding brackets by size, were grouped into three categories by performance: satisfactory, superior and outstanding, which determined the level of funding (Iskandarani and Reifschneider, 2008). Due to annual variability centre placement in a performance category varied.

Germany initially used data from six PM indicators, including impact culture and some institutional sub-indicators for allocating about 7.5 percent of its total annual funding to the CGIAR. It stopped this use because of dissatisfaction with the impact culture measure as a proxy for impact and a sense that the other indicators were not clearly linked to performance (Marlene Diekman, 2012 [personal communication]). There is anecdotal evidence that some other donors adjusted their funding on the basis of the PM results, particularly for impact culture, when it was initially interpreted as a measure of actual impact.

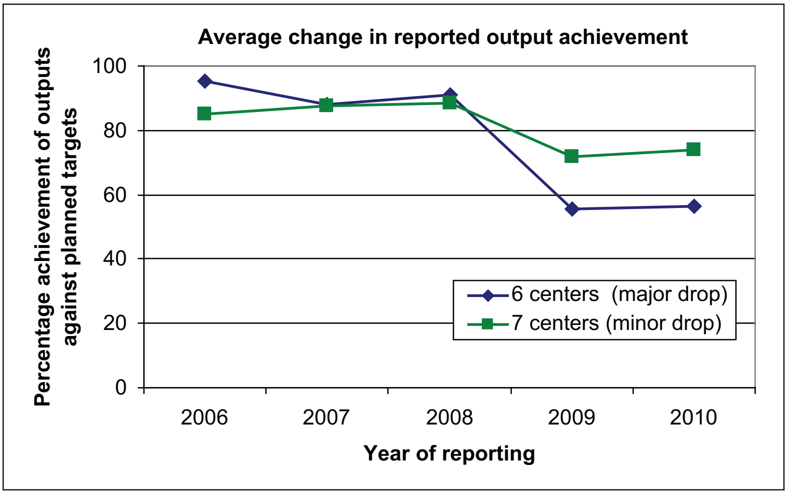

Figure 2. Effect of indicator status to output reporting; results on output achievement before and after status change.
'Major drop' = difference between average for 2006-08 and 2009-10 ranging from 28 to 61%; 'Minor drop' = difference ranging from 8 to 24%.

Apart from these examples, it is unclear whether the PM system influenced donor decision making. Traditionally, donors have targeted funding to particular centres or types of research based on their own priorities. Raitzer and Kelley (2008) found that funding decisions are made in a complex political environment in which information about results is only one component.

Providing benchmarks and targets

According to Iskandarani and Reifschneider (2008), donors using PM data to drive funding were setting 'clear incentives for good performance and in particular for organizational change in terms of governance, culture of learning and change as well as diversity'. At centres, prospect of financial gain likely motivated relatively simple changes. Specifically, the PM governance checklist appears to have increased the updating of certain policies; for example, the use of staff satisfaction surveys (CGIAR Independent Review Panel, 2008). On finance performance, centres tended to move towards targets that were given. With 'impact culture' and 'outcomes' the criteria for peer assessment were known and could be considered a benchmark. However, most indicators did not have clear targets. The rationale was that individual donors would be able to interpret the results and give weights to the indicators following their own emphasis and policies (Iskandarani and Reifschneider, 2008). Regarding incentives for improving performance, this practice sent an unclear message about what represented desirable performance.

Demonstrating accountability

The accountability purpose is complicated when there are multiple accountabilities – in the CGIAR to diverse beneficiaries, partners, donors and tax payers – and multiple interpretations of the intended accountability. Each centre, the CGIAR's administrative level and a group of very heterogeneous donors all likely had their own interpretations. As the IARC's research is funded primarily from development – rather than science – funding streams, the accountability expectations set by national constituencies and policy-makers focus on 'aid efficiency'. Yet, accountability for results was problematic: despite perverse incentives it was nonetheless very difficult to remove the initial indicator for outputs – *% outputs achieved* – precisely because of the perceived accountability imperative (CGIAR Secretariat, 2008).

Serving as an early warning system, particularly for administrative and financial health, was another accountability-related purpose of the PM system. CIAT (International Centre for Topical Agriculture) became a test case as it fell into financial and institutional crisis while the PM was operational (Field and Özgediz, 2008). However, only CIAT's 2007 external review brought the situation to the broader attention of the CGIAR donor community (CGIAR Science Council, 2008b) and forced action that eventually resulted in major changes in management and governance. The indicators on Board performance did not reveal an approaching crisis and although financial health indicators could have been informative, their critical thresholds were not well communicated and they were not acted upon. Field and Özgediz (2008) claim that the World Bank used funding effectively to reinforce accountability for performance, but CIAT's case does not support such a conclusion.

Use for management decisions

There is little systematic information about how centre management interpreted or acted on the PM results for their own centre, if at all. The PM system may have triggered more attention to outcome planning evident in some centre MTPs (for instance, CIAT, 2009; CIP, 2009), although reporting outcome cases in the PM relied more on anecdotal evidence than systematic documentation (Immonen and Fischer, 2012). There is also anecdotal evidence that the emphasis on impact culture may have stimulated more regular and rigorous impact documentation. In many cases centre management may not have known which direction the change should go or agreed with the direction implied. In the research managers' view, the PM designers' intentions of what to achieve and measure remained unclear (Alliance Deputy Executive, 2007). Importantly, the CGIAR was encouraging linkages across centres for integrated research, a goal that would have needed to be balanced against the motive to perform competitively on the PM measures. Perrin (2002) argues that PM systems that aggregated information disguise sub-unit differences where management decisions might be most needed. A search of external programme and management review reports completed while PM data were available showed that evaluation panels did not validate or confirm the PM results and hardly used the data. Due to these issues, it was likely difficult for centres to use the PM system to develop a systematic strategy to improve their performance.

Conclusions on the use of PM in the CGIAR

Challenges in measuring research outputs

Attempts to measure research output in the CGIAR PM system illustrate several of the pitfalls of PM in research, including risk avoidance, goal displacement and difficulty in setting targets. Measuring output performance against set targets failed because in the absence of an agreed threshold for satisfactory achievement (in any case difficult to determine), the performance goal was an unrealistic 100 percent and levels of achievement close to that were reported. The causes or significance of variability across centres or across years at that high end of the achievement spectrum were impossible to judge.

Given the uncertainty involved in research, the high output achievement likely reflected lack of ambition in research planning or very generically defined targets. Deliberately low targets undermine the value of the planning process intended to establish meaningful goals, and potentially stifle the risk-taking behaviour essential to the scientific endeavour. The risk in any target setting is that the achievement is fixed below the unit's full potential.

The CGIAR Challenge Programmes, a parallel research implementation vehicle used in the CGIAR, provide a contrasting case. Although implemented by IARCs, these programmes were not included in the PM system. In at least one of them, purposefully ambitious planning was used as a strategy in contracting partners and keeping them focused (CGIAR Science Council, 2006). Under a PM system, this programme would have been penalized for its ambition. Perrin (1998, 2002) observes that evaluations that assess the achievement of predetermined objectives may penalize programmes with ambitious objectives and favour mediocre ones. Such concerns may underlie EMBRAPA's decision to remove the 'effectiveness criterion' (ratio between planned and actually performed) from its PM, instead making this process a matter of negotiation with supervisor (Avila et al., 2008).

Using publishing rate as an indicator for outputs was questioned by the centres themselves because of the large variation in the topics addressed, the publication venues available and the speed with which publications are issued (Alliance Deputy Executive, 2007). Furthermore, underscoring the power of indicators to drive behaviour, this indicator may have drawn centre attention unduly to the volume of publishing. One external review panel observed that potentially high impact research results were split into multiple publications, some in inappropriate journals, which, in the panel's view, had severely reduced the scientific impact of the research (CGIAR Science Council, 2008c).

Challenges in designing indicators of research outcomes and impacts

The challenge of monitoring outcomes and impacts, mediated by many events and actors, is well-known as research organizations have little or no control over the longer-term results of their research. However, aid donors who may also fund research expect evidence of development benefits, often in the short term, as 'value for money' of their investment. As an annual measure, only indicators reflecting the *culture* of outcome monitoring and impact assessment could be included in the CGIAR's PM. However, it would be more appropriate to negotiate with the donors a level of result that is a satisfactory proxy for the desired development change, such as broad-scale adoption of technologies and knowledge generated by research. Such results can be documented periodically but are not amenable to annual measurement. Also the CGIAR Independent Review Panel (2008) concluded that for outcomes and impacts annual comparisons were not suited. The indicators that were eventually agreed upon for results had limited face validity as indicators of the extent to which centres were developing research-based solutions to agricultural development problems. Fundamentally, the PM system was unable to identify measures that corresponded accurately to the real world of valuing public good research.

Challenges in developing PM for research creating positive incentives

PM systems inevitably create incentives or dis-incentives. The CGIAR PM system did not and could not cover all elements of performance important for research and its mission-oriented objectives. Notably it missed quality, relevance and actual results of research and thus provided no incentives for improvements in those areas. It may have diverted attention from those aspects. Any incentives for learning would have been weakened by the fact that the centres saw themselves in competition, which was enforced by the ranking of centres in the performance reporting and classification by the World Bank. Despite the critique from the centres that the PM system promoted single centre mindsets and competition (Alliance Deputy Executive, 2008), individual centres highlighted the performance status granted annually by the World Bank when it was favourable (‘outstanding’).[1] Such conclusions can override information based on more comprehensive and in-depth evaluation of research quality, relevance, and results.

One research director characterised the PM system as having been a massive data collection effort involving many staff and numerous iterations, and encompassing subjective or meaningless indicators or scoring approaches where the scores were mostly a reflection of the effort placed on the submission (Achim Dobermann, 2012 [personal communication]). This resonates with others’ observations about burdensome, irrelevant and dysfunctional aspects of PM and the cynicism it generates in research context (Feller, 2002; Perrin, 1998). Similarly Radin (2011) concludes that PM systems have evoked a compliance mentality and cynicism among the individuals in the bureaucracy who are expected to change. Feller (2002) suggests that the research context transforms the limited but reasonable strategy of PM into a vacuous but effort-demanding undertaking. Feller also postulates that PM may have contributed to the bureaucratization of decision making, and to a relative elevation of the influence of administrative staff at the expense of the influence of peers. Had the practice of direct and mechanistic rewarding of centres on the basis of annual indicator results continued or become more widespread among donors, this could have had unpredictable, but likely negative, consequences for research prioritization and conduct.

Challenges serving multiple purposes and audiences

The CGIAR Independent Review Panel (2008) concluded that the system was not well positioned for learning because it was difficult for one tool to play divergent roles: accountability, resource allocation and learning. It is apparent that for learning, the PM system would have needed to be a management tool without the underpinning awareness of the direct rewards and sanctions.

The PM system was intended as an integral part of the CGIAR’s M&E system. However, two trends caused deviation from this goal. First, the use of the PM information by evaluations was limited and unsystematic. This is likely because of the lack of face validity of the PM results as indicators of centre progress and performance in areas central to both formative and summative evaluation. Second, the PM was explicitly used for funding decisions while the comprehensive external reviews that combined qualitative and quantitative information on institutional and research performance were not. These two approaches risked becoming competing, rather than complementary, influences on centres and donors. The PM system, with its focus on the micro level, risked atomizing the centres' work rather than helping scale it up (Alliance Deputy Executive, 2008).

Conclusions on the suitability of PM to research

It can be concluded that the CGIAR's PM experiment failed against all the intended purposes. There were inherent difficulties in developing a set of annual indicators with high validity in reflecting the kind of performance that research institutions are expected to demonstrate, on outputs, outcomes and impacts. The system therefore was dominated by simpler observations related to quantitative records and institutional issues with unclear connections to performance of research organizations.

These experiences apply generally to PM systems, particularly when applied to research where results are uncertain and impact pathways from research to societal benefits are complex and protracted. Indicators intended to simplify complexity can be a poor match to the activity or variable they intend to measure and result in ambiguous interpretations. Perrin (1998) concludes that use of metrics that poorly represent a complex reality reduces the appropriateness of the actions subsequently taken. Focusing on short-term, easily measured performance objectives risks the loss of attention on complex impact pathways and long-term objectives.

When a single system is created to meet the needs of multiple audiences, as in the case of the CGIAR, it is likely that some purposes will have to be sacrificed in order to achieve others (Feller, 2002; McDavid and Huse, 2012). Reporting that is simplified for donors may not serve either the purpose of research management or the donor accountability. Scientists should regularly monitor their progress, and for learning emphasis should also be placed on delays and unexpected results or 'failures'. A narrow focus on measurement is inconsistent with a focus on change and improvement that requires constant questioning about what else can be done or done better (Perrin, 1998). Rewarding mechanisms presumed to enforce accountability easily become a primary driver for behaviour changes reducing the incentives for learning. De Lancer Julnes (2006) reports that while the call for accountability has stimulated development of PM systems, it has also been perceived as a deterrent due to fears of repercussions from non-compliance and inability to meet set targets.

As discussed by Lepori and Reale (2012), research programmes are a highly differentiated domain regarding their objectives, modes of selection and funding, and the contractual relationships between donors and research groups. Certain characteristics of research programmes, identified by them, are also typical of the CGIAR: there are multiple funding schemes that can address partially conflicting goals, and the programmes are jointly designed and implemented with their beneficiaries. That means that complex interactions arise among largely autonomous and strategic actors belonging to 'policy', 'society' and 'science' spheres (Lepori and Reale, 2012). As Gray (2008) states, new models of doing research create stakeholder groups that didn't exist before and that also need to be served by research monitoring and evaluation. In the CGIAR, all research involves national and international partners, and negotiation, trust and communication among partners is essential. Thus, indicators need to be context-specific and debatable (Lepori and Reale, 2012). PM systems with rigid reward and punitive mechanisms are likely to divide, rather than unify research partnerships.

In the context of monitoring research performance, our analysis reinforces the importance of being clear about the purpose of monitoring systems and designing them accordingly;

placing emphasis on appropriate incentives for improving performance and learning; presenting data and conclusions in ways that facilitate the appropriate use of the monitoring information; and considering short-term monitoring as one component of a comprehensive evaluation and impact assessment plan.

In the following we discuss the implications of these lessons for three aspects of performance monitoring in agricultural research for development: annual performance monitoring focusing on research progress and operations; periodic monitoring of outcomes and impacts; and linking monitoring and evaluation.

Performance data collected annually needs to serve research managers who need to monitor current conditions (e.g. efficiency in operations, deviations in plans resulting from risk in research, achievement of milestones) as well as anticipate future issues (e.g. budget shortfalls). Management should rely on data that already are or can be routinely collected. In an organizational culture of transparency, management shares these performance data with research staff and uses them to stimulate discussion of how to work most effectively. Short-term monitoring can optimally serve learning and improvement and generate incentives for risk and ambition in breaking new ground in research.

Propper and Wilson (2003) note that in PM systems intended for management use, the information is often not published externally. However, several funders who currently fund development oriented research require annual reporting. Reporting requirements should therefore be negotiated so that they enforce good management, including programme design and implementation, rather than distract from it. This kind of negotiation could also build funders' confidence on management's commitment to using short term performance information appropriately.

Given the limitations of annual reporting systems as a source of information on research progress and results, researchers – and funders – can turn to periodic studies as a complementary form of monitoring. They are used to document and analyse evidence and lessons on uptake and adoption of research results, their influence on policy-making, and the longer term benefits generated in society and the environment. In research targeting development, periodic monitoring is needed for testing assumptions underlying programmes' theories of change and providing feedback on research design, including from negative outcomes. The studies also respond to donors' requirement of documented evidence of development outcomes. However, in many areas of agricultural research more work is needed to decide on measurements that give most informative, credible information on causal linkages, progress and change, particularly where biological and social parameters are integrated. In addition, dedicated resources are needed for this kind of periodic monitoring.

A particular challenge is to make performance measurement and monitoring useful for programme evaluation. Of the complementarities identified by Nielsen and Hunter (2013), sequential and informational have the clearest relevance in the research-for-development context. Not all performance data will be relevant, especially when causality is the focus. However, regularly collected performance information can contribute data on trends over time to programme evaluations, providing insight on management and programmatic directions over a number of years. Evaluation can combine quantitative (from performance measures and other sources) and qualitative data to provide a full picture of causal chain and inform decisions about programme changes if appropriate (Poister, 2010). To make performance management (including measurement) useful both for managerial and evaluative purposes, Van Helden et al. (2012) propose an entire 'life-cycle' analysis from design and implementation to use and impacts. The reformed CGIAR is moving to this strategy, with more emphasis placed on programme design, including the development of theories of change. Both monitoring and evaluation should support such life-cycle analysis in an iterative fashion where learning and evidence of achievement contribute to enforcing or adapting the theory of change.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Note

1. Quote from IWMI Web site: ‘In 2008, the World Bank gave IWMI an “outstanding” rating for performance. IWMI is one of four centres in the CGIAR to receive this rating which is the highest of three performance categories.’ http://www.iwmi.cgiar.org/about/iwmi-performance/.
   Quote from ICRISAT Web site: ‘The “Outstanding” rating recognizes ICRISAT’s good science, great impacts, institutional health and financial health. It places ICRISAT’s performance on top of the 15 international agricultural research institutes that are members of the CGIAR.’ http://www.icrisat.org/newsroom/news-releases/icrisat-pr-2008-media12.htm

References

Alliance Deputy Executive (2007) The CGIAR Monitoring and evaluation processes and the CGIAR performance measurement system – an Alliance perspective. *CGIAR Performance Measurement Workshop*, FAO, Rome, Italy, 24 August. Rome: Office of the Alliance of the CGIAR Centers.

Alliance Deputy Executive (2008) Mid-year summary. *Performance Measurement Workshop*, Washington, DC, USA, 17–18 July. Rome: Office of the Alliance of the CGIAR Centers.

Ammons DN and Rivenbark WC (2008) Factors influencing the use of performance data to improve municipal services: evidence from the North Carolina Benchmarking Project. *Public Administration Review* 86(2): 304–18.

Avila AFD, Gomes EG, Da Silva e Souza G and Yeganiantz L (2008) Performance evaluation of Embrapa’s research centres: experience and learning process. In: *Methodological Innovations in Impact Assessment of Agricultural Research Workshop*, Brasilia, Brazil, 12–14 November.

Beintema N, Avila F and Fachini C (2010) *Brazil: New Developments in the Organization and Funding of Research*. ASTI Country Note. Washington, DC, and Brasilia: International Food Policy Research Institute and Brazilian Agricultural Research Corporation.

Butler L (2003) Explaining Australia’s increased share of ISI publications – the effects of a funding formula based on publication counts. *Research Policy* 32(1): 143–55.

Butler L (2007) Assessing university research: a plea for a balanced approach. *Science and Public Policy* 34(8): 565–74.

CGIAR (2003) *Towards Designing a Performance Measurement System for the CGIAR*. Washington, DC: CGIAR Secretariat.

CGIAR (2004) *Recommended Indicators for a Performance Measurement System for the CGIAR*. Washington, DC: CGIAR Secretariat.

CGIAR (2009) *Voices for Change. The New CGIAR*. Washington, DC: CGIAR Fund Office.

CGIAR Independent Review Panel (2008) *Bringing Together the Best of Science and the Best of Development*. Report for the Executive Council. Washington, DC: CGIAR.

CGIAR Science Council (2004) *End of Meeting Report*. Inaugural Meeting of the Science Council. ICARDA, Ted Hadya, Aleppo, Syria, 12–14 May. Rome: Science Council Secretariat.

CGIAR Science Council (2005) *Monitoring and Evaluation System for the CGIAR Centres*. Rome: Science Council Secretariat.

CGIAR Science Council (2006) *SC Commentary and Centre Responses on CGIAR Centre and Challenge Program Medium-Term Plans 2007–2009*. Rome: Science Council Secretariat.

CGIAR Science Council (2008a) *SC Suggestions to Strengthen the CGIAR Performance Measurement System (PMS)*. Rome: Science Council Secretariat.

CGIAR Science Council (2008b) *Report of the Sixth External Program and Management Review of the Centro Internacional de Agricultura Tropical (CIAT)*. Rome: Science Council Secretariat.

CGIAR Science Council (2008c) *Report of the Sixth External Program and Management Review of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)*. Rome: Science Council Secretariat.

CGIAR Science Council (2009a) *Experiences from Five Years of Performance Measurement System: Research-related Indicators*. Rome: Science Council Secretariat.

CGIAR Science Council (2009b) *Annex I. SC Guidelines for the Use of the Research-based PMS Indicators*. Science Council Report to ExCo members, 18 May 2009. Rome: Science Council Secretariat.

CGIAR Science Council and CGIAR Secretariat (2009) *Instructions for the Reporting of Performance Indicators for CGIAR Centres (2009 data)*. Washington, DC: CGIAR Secretariat.

CGIAR Secretariat (2008) *Optimizing the Utility of the Performance Measurement System: A Note on the Value of the Outputs Indicator*. Washington, DC: CGIAR Secretariat.

CIAT (2009) *Medium-Term Plan 2010–2011*. Cali: CIAT.

CIP (2009) *Medium-Term-Plan 2010–2011*. Lima: CIP.

Cozzens SE (1997) The knowledge pool: measurement challenges in evaluating fundamental research programs. *Evaluation and Program Planning* 20(1): 77–89.

Da Silva e Souza G, Alves E and Avila AFD (1999) Technical efficiency of production in agricultural research. *Scientometrics* 46(1): 141–60.

Davies H (2006) *Measuring and Reporting the Quality of Health Care. Issues and Evidence from the International Research Literature*. Edinburgh: NHS Quality Improvement Scotland.

Davies HTO and Lampel J (1998) Trust in performance indicators? *Quality in Health Care* 7(3): 159–62.

De Lancer Julnes P (2006) Performance measurement: an effective tool for government accountability? The debate goes on. *Evaluation* 12(2): 219–35.

Feller I (2002) Performance measurement redux. *American Journal of Evaluation* 23(4): 435–52.

Field L and Özgediz S (2008) *CIAT's Institutional Crisis: Lessons Learned*. Washington, DC: CGIAR Secretariat.

Gray DO (2008) Making team science better: applying improvement-oriented evaluation principles to evaluation of cooperative research centers. In: Coryn CLS and Scriven M (eds), *Reforming the Evaluation of Research*. New Directions No. 118. Wiley periodicals, Inc. 73–87.

Guena A and Martin BR (2003) University research evaluation and funding: an international comparison. *Minerva* 41(4): 277–304.

Hatry HP (2006) *Performance Measurement: Getting Results*, 2nd edn. Washington, DC: Urban Institute.

Holzer M and Kloby K (2005) Public performance measurement: an assessment of the state-of-the-art and models for citizen participation. *International Journal of Productivity and Performance Management* 54(7): 517–32.

Hood C, Dixon R and Wilson D (2009) *'Managing by Numbers': The Way to Make Public Services Better?* Oxford: Economic and Social Research Council, Public Services Programme.

Immonen S and Fischer K (2012) *Experiences of Outcomes Monitoring in the CGIAR*. Rome: CGIAR Independent Science and Partnership Council Secretariat.

Iskandarani M and Reifschneider FJB (2008) Performance measurement in a global program: motivation, new concepts and early lessons from a new system. *Science and Public Policy* 35(10): 745–55.

Jann W and Jantz B (2008) A better performance of performance management? In: KPMG, CAPAM, IPAA, IPAC, *Holy Grail or Achievable Quest. International Perspectives on Public Sector Performance Management*. Zürich: KPMG International, 11–25.

Kelley T, Ryan J and Gregersen H (2008) Enhancing *ex post* impact assessment of agricultural research: the CGIAR experience. *Research Evaluation* 17(3): 201–12.

Lepori B and Reale E (2012) S&T indicators as a tool for formative evaluation of research programs. *Evaluation* 18(4): 421–65.

McDavid J and Huse I (2012) Legislator uses of public performance reports: findings from a five-year study. *American Journal of Evaluation* 33(1): 7–25.

Moed HF (2007) The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Policy* 34(8): 575–83.

Moynihan DP (2008) *The Dynamics of Performance Management: Constructing Information and Reform*. Washington, DC: Georgetown University Press.

Natsios A (2010) *The Clash of the Counter-bureaucracy and Development*. Washington, DC: Centre for Global Development.

Nielsen SB and Hunter DEK (2013) Challenges to and forms of complementarity between performance management and evaluation. In Nielsen SB and Hunter DEK (eds), *Performance Management and Evaluation. New Directions for Evaluation, 137*. Wiley periodicals, Inc. 115–23.

OEDC (n.d.) *Introduction STI Review No. 27. New Science and Technology Indicators for the Knowledge-based Economy: Opportunities and Challenges*. URL (consulted 22 October 2012): http://www.oecd.org/document/17/0,3746,en_2649_34451_2669841_1_1_1_1,00.html.

Özgediz S (1999) Evaluating research institutions: lessons from the CGIAR. *Knowledge, Technology & Policy* 11(4): 97–113.

Perrin B (1998) Effective use and misuse of performance measurement. *American Journal of Evaluation* 19(3): 367–79.

Perrin B (2002) How to – and how not to – evaluate innovation. *Evaluation* 8(1): 13–28.

Pinstrup-Andersen P (2004) From Science Council Chairman. *CGIAR News, September 2004*. Washington, DC: CGIAR Secretariat, 11.

Poister TH (2010) Performance measurement: monitoring program outcomes. In: Wholey JS, Hatry HP and Newcomer KE (eds), *Handbook of Practical Program Evaluation*, 3rd edn. San Francisco, CA: Jossey-Bass, 100–24.

Propper C and Wilson D (2003) The use and usefulness of performance measures in the public sector. *CMPO Working Paper Series No. 03/073*.

Radin BA (2011) Does performance measurement actually improve accountability? In: Dubnick MJ and Frederickson HG (eds), *Accountable Governance: Promises and Problems*. Armonk, NY: M.E. Sharpe, 98–110.

Raitzer DA and Kelley TK (2008) Assessing the contribution of impact assessment to donor decisions for international agricultural research. *Research Evaluation* 17(3): 187–99.

Ruegg R and Jordan G (2007) *Overview of Evaluation Methods for R&D Programs*. Washington, DC: US Department of Energy.

Smith P (1995) On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18(2–3): 277–310.

Thomas PG (2003) Performance measurement, reporting and accountability. *Public Policy Paper No. 23*. Regina: Saskatchewan Institute of Public Policy.

Van Dooren W and Thijs N (2010) Paradoxes of improving performance management (Systems) in public administration. *EIPASCOPE Bulletin* 2010(2): 13–18.

Van Helden GJ, Johnsen Å and Vakkuri J (2012) The life-cycle approach to performance management: implications for public management and evaluation. *Evaluation* 18(2): 159–75.

Van Thiel S and Leeuw FL (2002) Public sector paradox. *Public Performance & Management Review* 25(3): 267–81.

Winston JA (1999) Performance indicators – promises unmet: a response to Perrin. *American Journal of Evaluation* 20(1): 95–9.

Sirkka Immonen is Senior Evaluation Officer in the CGIAR’s Independent Evaluation Arrangement. Previously she coordinated institutional and programme evaluation of the International Agricultural Research Centres working for the CGIAR’s Science Council.

Leslie L. Cooksy is Evaluation Director of Sierra Foundation. Past career includes: American Evaluation Association, President (2010–11); University of Delaware, Associate Professor; US Government Accountability Office's Programme Evaluation and Methodology Division.