



©FAO/Simon Maina

The Rigor Revolution in Impact Assessment: Implications for CGIAR



Independent
Science and
Partnership
Council

Standing Panel on Impact Assessment (SPIA)

ABOUT ISPC

The Independent Science and Partnership Council (ISPC) is a part of CGIAR, a global research partnership for a food secure future dedicated to achieving a world free of poverty, hunger and environmental degradation. The ISPC provides independent advice to CGIAR System Council which is comprised of funders and representatives of developing countries. The mission of the ISPC is to help strengthen the quality, relevance, and impact of CGIAR research by enhancing the System Council's capacity to make evidence-based decisions in support of effective agricultural research programs for sustainable development.

ABOUT SPIA

The Standing Panel on Impact Assessment (SPIA) is a sub-group of the ISPC. SPIA's mandate is to provide CGIAR with timely, objective, and credible information on the impacts of research at the system level; provide support to and complement CGIAR centers in their ex post impact assessment activities; and, provide feedback to planning, monitoring and evaluation functions in CGIAR.



Rights and Permissions

This work is available under the Creative Commons Attribution 3.0 IGO license (CC BY 3.0 IGO) <http://creativecommons.org/licenses/by/3.0/igo>. Under the Creative Commons Attribution license, you are free to copy, distribute, transmit, and adapt this work, including for commercial purposes, under the following condition:

Attribution—Please cite the work as follows:

Stevenson, J., Macours, K., & Gollin, D. 2018. *The Rigor Revolution in Impact Assessment: Implications for CGIAR*. Rome: Independent Science and Partnership Council (ISPC).

Funding Acknowledgements

This research was supported by ISPC-SPIA under the grant '[Strengthening Impact Assessment in the CGIAR \(SIAC\)](#)'

Independent Science and Partnership Council (ISPC)

<http://ispc.cgiar.org>

Cover image: ©Chris Steele-Perkins/Magnum Ph

Design and layout: Macaroni Bros

Editor: Heidi Fritschel

James Stevenson
Karen Macours
Doug Gollin

The Rigor Revolution in Impact Assessment: Implications for CGIAR

November 2018

TABLE OF CONTENTS

EXECUTIVE SUMMARY	III
FOREWORD	V
1. THE RIGOR REVOLUTION	1
1.1 Does investing in agricultural research pay?	1
1.2 A decade of rapid methodological development	3
1.3 Does it still pay to be ignorant?	7
2 . MEASUREMENT MATTERS	8
2.1 Existing data on adoption of improved varieties: Fit for purpose?	8
2.2 DNA fingerprinting for varietal adoption: Establishing proof of concept	10
2.3 Opening Pandora's box? Or a gold mine?	12
2.4 Natural resource management: Huge scope for improved data collection	13
2.5 Disadoption	15
2.6 Outcome variables: Are we measuring what we think we're measuring?	15
3. CAUSALITY AND BIAS	17
3.1. Randomized control trials: A powerful methodology to be used wisely	17
3.2. How do we evaluate promising technologies?	19
3.3 Appropriate methodology is the gold standard	21
4. UNDERSTANDING CGIAR IMPACTS ON A LARGE SCALE	25
4.1 External validity, evidential standards, and the long-run effect on aid allocation	25
4.2 Rigorously measuring outcomes at scale	26
4.3 Micro-meso-macro: Interactions across scales	26
5. CONCLUSION AND SUMMARY	28
6. REFERENCES	30

EXECUTIVE SUMMARY

This synthesis report highlights major changes in methodology and standards of evidence in impact assessment that have taken place over the past decade, taking as its starting point the publication of the landmark report *When will we ever learn? Improving lives through impact evaluation* by the Center for Global Development (CGD). The CGD report called for greater rigor in the evaluation of development projects and in claims about the impact of aid. The report heralded a new era in which donors and other aid actors have insisted on higher-quality evidence on the effectiveness of aid expenditures. This shift, which we refer to as the “rigor revolution,” is the focus of section 1 of this synthesis report. Impact assessment in CGIAR has a long and proud tradition, focused on generating estimates of the economic returns to investments in agricultural research. Summaries of this literature have consistently shown large aggregate returns, suggesting that investing in agricultural research is a good use of scarce aid dollars. Donors to the CGIAR system have consistently communicated to the Standing Panel on Impact Assessment (SPIA) that they highly value such estimates. However, we highlight three major methodological challenges to the cost-benefit analysis tradition in CGIAR that need to be addressed to obtain more rigor in impact assessment. Section 2 (“Measurement matters”) highlights our findings concerning measurement errors and data quality. Section 3 (“Causality and bias”) discusses the new emphasis on intellectual stringency in establishing causal inference between research investments and claimed outcomes. Section 4 (“Understanding CGIAR impacts on a large scale”) speaks to the challenges of evaluating impacts at large spatial scale based on models and observational data. Section 5 concludes.

Section 1 reviews how calculations of the Internal Rate of Return (IRR), used for decades, may have consistently over-estimated rates of return to investment in agricultural research. Hurley, Pardey and Rao (2016) propose instead a “Modified Internal Rate of Return” (MIRR). Re-estimating prior studies using their MIRR across 2,829 evaluations they find that the mean IRR is an implausible 59.6 percent whereas the recalibrated MIRR is a more realistic – though still high -- 14.3 percent. Impact evaluations (in which causality is established using rigorous research designs), have boomed and spread across many realms of the international aid system having started first in health, and later permeating education and social protection before reaching agriculture in significant numbers only in the past few years. This rigor revolution has changed standards of evidence for demonstrating impact, and we argue that CGIAR needs to respond positively to these changes in the wider aid sector.

Section 2 shows how challenging it is to accurately measure – at the relevant scale and at reasonable cost – all the variables we need for the purposes of impact assessment. New insights from DNA fingerprinting, by collecting the same data using “traditional” methods and comparing these data to DNA fingerprinting results, have highlighted how the data we rely on for impact assessment may be subject to significant measurement error. These are errors that are sufficiently large to change the interpretation of analysis. Traditional approaches to measurement are often found to be biased rather than just “noisy”. In a different vein, new sensors and satellites are launching all the time, changing the possibilities for remote sensing as a tool for assessing diffusion and/or impacts– especially with regard to natural resource management research.

Section 3 discusses how randomized control trials have given us new insights into specific impact pathways from research to development outcomes. In particular, the behavioral adjustments that farmers make during and following the adoption of an innovation may serve to amplify, or to dampen, the extent of the development outcomes that might have been expected absent such an adjustment. For example, when adopting an improved variety that reduces risk of crop failure (e.g. a risk-reducing trait such as flood tolerance), farmers may be more likely to apply fertilizer and invest their own labor more heavily on that crop. This can lead to increased yields and profits relative to a scenario in which the technology is appraised ignoring such adjustments. Heterogeneity in environmental conditions, and among the population of potential adopters, significantly conditions outcomes. Failing to account for such heterogeneity can lead to unrealistic expectations of adoption and impact. Given that not all innovations can be evaluated with randomized control trials, we argue that appropriate methodology based on credible assumptions should be the gold standard in impact assessment for CGIAR research.

Section 4 addresses the themes of external validity – in other words, the process of going from “it worked somewhere” to “it will work here”. Many impact evaluation studies provide information on a particular context; but often we are interested in the broader question, of whether the innovation will work outside of its original context. For external validity, individual empirical studies should have limited weight; we learn from evaluating multiple studies that provide a fuller and better rounded view of the technology. This is particularly important for agriculture, where the contextual factors that can condition heterogeneity of outcomes are varied and impossible to fully control for. However, as many recent reviews have highlighted, the average level of quality in the published impact

assessment literature is low, with many biased studies. Rigorously documenting adoption rates for large representative populations is complementary to studies identifying specific causal relationships. Furthermore, given the complex nature of some impact pathways, macro-level models are needed, but we argue that these deserve more serious and sustained scrutiny than has been the case to date.

Section 5 concludes by making two key conclusions and one further affirmation of SPIA’s evolving role in CGIAR. **The first conclusion is that we need to institutionalize detailed data collection related to CGIAR activities along the results chain from investments to outputs to outcomes.** For such an effort to be practical, we suggest focusing on a few key locations as a first step toward catching up, with carefully implemented geo-located surveys featuring DNA fingerprinting of the major crops and livestock, reliable data on farmers’ management practices, and detailed socioeconomic data, combined with information on the policy and institutional environment. **The second conclusion is that impact evaluation and efficacy studies need to focus on causal relationships for which we have the greatest uncertainty and for which information would have the highest value.** This suggests a greater focus on theory—away from searching for “what works” in the abstract and toward finding out why certain things work and others do not in particular contexts. Finally, given the wide range of activities carried out by the CGIAR, it is clear that a broad toolkit of approaches will be needed to assess impacts. This means that **standardization and simple messages are hard to come by, but SPIA is committed to its role as convener and intermediary between the CGIAR research community, external researchers, and the donors that fund the system.** In doing so, we hope to ensure that we can raise the ratio of signal to noise and help incentivize greater clarity, realism, and rigor in the thinking about impacts from investments in CGIAR.

FOREWORD

In 2006, the Center for Global Development – an influential independent think tank in Washington DC – published a watershed report for the development sector entitled: *When will we ever learn? Improving lives through impact evaluation*. The report summarized long-held concerns about the role of evidence in decision-making in development and proposed a way forward that emphasized the need for greater rigor in the design of studies of impact. In fact, so preoccupied were these authors with the idea that we needed more rigor, that the words “rigor” or “rigorous” feature a generous 72 times in that report. A little more than a decade later and this message has become quite mainstream.

In this ISPC report – *The Rigor Revolution in Impact Assessment: Implications for CGIAR* – authors James Stevenson, Karen Macours and Doug Gollin reflect on the fundamental changes that have come about, partly resulting from the influence of the Center for Global Development report. The toolkit of applied economists has changed dramatically over the past decade, with a much greater emphasis on carefully identifying causal relationships. The saying that “correlation is not causation” is just the start in determining where we should draw the line on appropriate standards of evidence when we try to compare the impacts of different kinds of aid investments. And given the sheer range of topics that CGIAR works on, and the very different modes in which it engages with those topics – from inventing new technological solutions to problems, to advising government agencies – there is much food for thought here on how we should think about the process of determining where things are working well and where they are not.

The authors argue that accurate measurement of variables is a critical feature of any concept of “rigor” we might want to employ, and yet this can be overshadowed by a preoccupation with causal identification. Clearly both are important. New technologies (or at least cheaper and more accurate versions of old technologies) in the fields of DNA fingerprinting and remote sensing have emerged that make it possible to measure things that we simply could not measure accurately in the past. This is an active area of considerable research effort.

Finally, and perhaps most challenging of all, the authors argue that our understanding of impacts change depending on the timing of the assessment, and especially the scale at which we study. Typically, CGIAR impact assessments have been too localized in their geographic focus, making it hard to draw inferences for wide, representative geographies. Pursuing collaborations with well-institutionalized surveys that are routinely carried out in large geographies, is one strategy towards building more representative evidence. However, impact pathways linking research to development outcomes are often quite complex, which suggest the need for developing models to study them at a macro level. Better farm-level evidence should help us with the design and testing of better models.

This report draws on the contributions of many researchers whose work was funded under the [Strengthening Impact Assessment in CGIAR program](#) (SIAC) which SPIA ran from 2013 – 2017. Rather than summarizing that body of work, it synthesizes some of the key insights on methodology that the SPIA team gleaned from their involvement. I recommend this report to research managers, science leaders and research scientists of all disciplinary backgrounds.

Leslie Lipper
Executive Director, ISPC Secretariat

1 THE RIGOR REVOLUTION

1.1 DOES INVESTING IN AGRICULTURAL RESEARCH PAY?

Impact assessment of investments in agricultural research has a long and proud tradition in CGIAR, aimed largely at providing answers to the question of whether it pays to invest in agricultural research. Until the mid-2000s, ex post impact assessments were dominated by the use of an economic model of demand for and supply of agricultural products in partial equilibrium. The basic ideas for this approach were sketched out by Griliches more than half a century ago (Griliches, 1957, 1958). Griliches had observed the rate of adoption of hybrid maize varieties in different states of the United States and created a simple model for linking the benefits from higher maize yields back to investments in research.

For CGIAR, the appeal of such a model is its simplicity. The first task in implementing the model is an adoption study to establish whether CGIAR innovations have been adopted at a large scale. The impact of widely adopted innovations on aggregate agricultural productivity is then modeled as an exogenous “shock” to a market that was assumed to be in partial equilibrium. In this model of research impacts, the supply curve is assumed to shift outward such that more output is produced for a given level of input use, and the magnitude of the shift is calculated using yield advantage data from agronomic trials, econometric analyses or any other sources the authors can find. All else being equal, this economic surplus is assumed to be shared out between producers and consumers according to a series of conditions that approximate the context. The partial equilibrium model gives economists a way of then estimating a “stream” of the benefits that are assumed to flow from the adoption of innovations over time, measured in dollars. This has the tremendous advantage that it can be directly compared to the total funding used in the research, either by a particular research center or program that helped to generate the innovation or by the CGIAR system as a whole, in a cost-benefit analysis.

Estimates of the economic returns to research—generated using such a partial equilibrium Griliches-style model feeding into cost-benefit analyses—have been produced periodically throughout the history of CGIAR (Alston, Norton, & Pardey, 1995; Raitzer & Kelley, 2008). Historically, donors to the CGIAR system have communicated to SPIA that they highly value such estimates, and there is continued demand for these aggregate numbers, which can play an important role in defending the allocation of scarce public funds to international agricultural research.

Reappraising the literature on rates of return

The literature on rates of return to research is open to two broad critiques: (1) the methods used to generate the estimates are heavily dependent on strong assumptions leading to numbers that are implausible; and (2) the aggregate rate of return is no longer a useful metric in the era of Millennium Development Goals (MDGs)/Sustainable Development Goals (SDGs), which is characterized by multiple objectives and concern about distributional consequences.

Hurley et al. (2016) address the first of these two critiques by comprehensively reviewing the literature on estimates of the internal rates of return (IRRs) to investments in agricultural research. The figures for IRR reported across hundreds of studies has remained very high, at well above 40 percent, whereas funding from donor countries to agricultural research has stalled or declined over the past few decades. However, as Hurley et al. (2016) note, there has been a history of misinterpreting the IRR as being the correct metric for assessing social returns to investment in agricultural research - the assumptions implicit to the calculation of IRR are not consistent with the realities of the benefit and cost streams associated with agricultural research.¹ Rao, Hurley, & Pardey (2016) therefore propose the modified internal rate of return (MIRR). Examining more than 2,829 evaluations in the database of the International Science and Technology Practice and Policy (INSTEPP, v3.0) program, they find that the mean IRR is an implausible 59.6 percent whereas the recalibrated MIRR is 14.3 percent²—still high, suggesting that aid investments in agricultural research pay off handsomely, but at a much more realistic order of magnitude.

A profusion of models

Despite these methodological concerns, the institutional history of regularly publishing studies on the impact of investments arguably made the CGIAR of the late 1990s a leader in development aid effectiveness. Many other institutions operating in the same broad field of international development were likely investing less in impact assessment, less systematically, and less often. However, the introduction of the MDGs was a turning point in donors' expectations regarding aid effectiveness and had significant implications for impact assessment methodology.

Donor agencies became interested in a wider set of pathways from research to impact. The expected aggregate ratio of benefits to costs may remain an important criterion for guiding funding allocations, but donors increasingly look to CGIAR to deliver development outcomes that address specific societal concerns—cutting poverty, reducing food insecurity, improving nutrition, ensuring environmental sustainability. All these objectives speak directly to the SDG agenda (and the MDGs before them). In theory, the impact of agricultural research on some of these high-level objectives can be estimated using the same family of partial equilibrium models used to estimate aggregate rates of return, but would suffer the same shortcomings as the earlier IRR calculations. Hence, in practice, impact assessment must adapt to using evidence from a broader range of methods to stay relevant.

In 2015, CGIAR published a revised Strategy and Results Framework (SRF) (CGIAR, 2015). This framework of three system-level outcomes (reduced poverty, food and nutrition security, environmental sustainability) and 10 intermediate development outcomes describes a universe of potential causal pathways from investments in CGIAR (see figure 1 below). Despite this, research strategies in CGIAR still often focus specifically on increasing agricul-

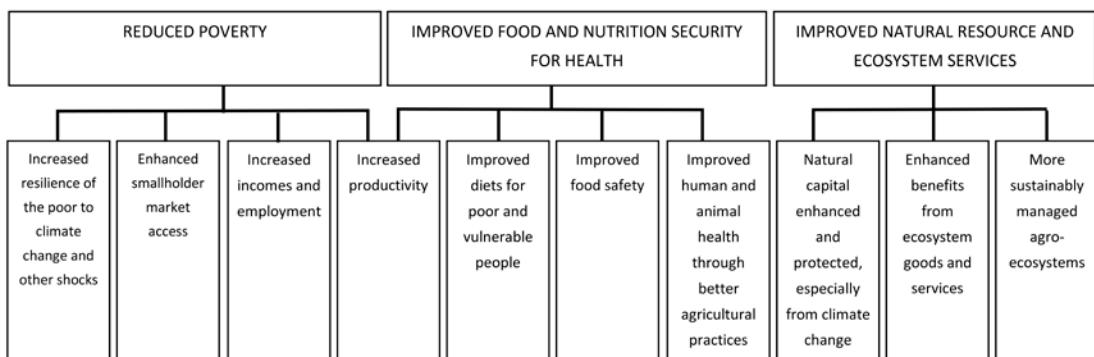
¹ As Hurley, Pardey, & Rao (2016) note, high IRR values and their implausible implications are the result of two key assumptions used in the calculation. First, the calculations assume that the beneficiaries of the investments (e.g., farmers and consumers) can reinvest their benefits at the same high rate of return. Second, the cost of the investment over time is discounted at the same high rate of return. These two assumptions inflate the reported rate of return on the investment when compared with historically more reasonable reinvestment and discount rates.

² This figure reflects an assumed research lag of 30 years and a discount rate of 5 percent.

tural productivity. Moreover, while agricultural productivity conceptually should measure a ratio of output to all inputs, agricultural productivity in practice often seems to be treated as synonymous with agricultural yields (kg/ha). And while yields may be the relevant metric to maximize when agricultural land area is limiting, in many contexts increasing farmers' income by increasing their

labor productivity (\$/hour worked), rather than their land productivity, is what is needed to have [impacts on poverty](#). Similarly focusing research efforts on yield enhancement is not necessarily the best strategy to increasing food and nutrition security, and it certainly does not automatically lead to environmental sustainability.

Figure 1. Top levels in the CGIAR Strategy and Results Framework for 2016 – 2030 (CGIAR, 2015)



Given that poverty reduction is a system-level mission, we should be troubled by patterns such as that outlined by Barrett and Upton (2013) who show that partial productivity gains in agriculture have been biased in favor of land rather than labor throughout Sub-Saharan Africa since the turn of the millennium. Between 2000 and 2009, production (in USD) per hectare increased proportionally a lot more than production per worker across the majority of countries in the region. Given that increasing labor productivity is essential for reducing poverty, and that Sub-Saharan Africa is the region of the world that received most attention from CGIAR over this period, we should ask whether the research is targeting the right problems. Certainly, identifying the kinds of research investments that will increase labor productivity may be difficult ex-ante. However, taking the task seriously is likely to be critical to any sustained effort CGIAR may want to make to reduce poverty (Gollin, Probst and Brower, 2018). This argues for an important role for impact assessment results to feedback into priority-setting.

Thus, it is a broad range of pathways that donors, SPIA, and others interested in the impacts of the

CGIAR should be interested in trying to analyze and document. Over the course of a decade, CGIAR has gone from arguably having too few mental models of how agricultural research leads to impact to having too many. CGIAR has been trying to meet these demands for evidence on an ever-broaden range of outcome metrics while also adapting to the new tools and norms of the "rigor revolution" of the past decade. It has been a struggle to keep pace with expectations. The risk for CGIAR is that without a coherent and plausible solution to the issue of results tracking and reporting, donors become more risk averse and double-down on requirements for indicators in multiple formats, with different timelines and priority geographies. Our hope is that, with this paper, we can start a conversation that leads us in a more productive direction.

1.2 A DECADE OF RAPID METHODOLOGICAL DEVELOPMENT

In a 2002 paper, Lant Pritchett developed a model addressing what he saw as the chronic under-investment in rigorous evaluation in international development at the time. Pritchett's model ex-

plores the interplay between the actions of “advocates” (program directors) and those providing resources (“voting public”).³ Advocates must secure resources for their programs—they are the entrepreneurs of the development industry. Advocates believe that they know the true effectiveness of the program they are implementing and that a rigorous evaluation will reveal this true effectiveness (Pritchett, 2002).

In Pritchett’s model, advocates can pursue one of two strategies to secure resources for their program. The first is to subject their program to rigorous impact evaluation and offer the evidence from these evaluations to donors for their consideration. The alternative strategy is to do what Pritchett calls “pilot and persuade”—that is, to implement the program in a location, to show that it is not physically impossible to do so, and then to invest in communication materials to persuade donors to give money to replicate this “success.” Pritchett shows that in many circumstances pursuing rigorous evaluation is simply not rational from the perspective of the program director/advocate.⁴ Rather, it pays to be ignorant of the true effectiveness of the program. The resources that could be spent on rigorous impact evaluation are better spent on communication, which allows advocates to get even better at persuading donors (and the public that they represent) that their programs are effective.

When will we ever learn? 10 years of the impact evaluation boom

In 2006, the Center for Global Development published its landmark report *When will we ever learn? Improving lives through impact evaluation* which laid out the extent to which the lack of rigorous evidence in international development was

a pervasive problem (Savedoff et al., 2006). The report brought to the surface some long-term issues in our understanding of aid effectiveness and led to the creation of the International Initiative for Impact Evaluation (3ie). Now a clearinghouse for rigorous evidence on aid effectiveness, 3ie coordinates funding of impact evaluations and systematic reviews of development policies and programs. 3ie received initial funding from a small number of key donors—in particular, the UK Department for International Development (DFID) and the Bill & Melinda Gates Foundation—with strong commitments to protecting the overall development aid budget and improving aid effectiveness.

In the period following the publication of *When will we ever learn?* several factors came together to generate both a greater demand for and supply of rigorous impact evaluations in the development sector. Certainly, more funding was made available for impact evaluation than ever before, supported by high-level multilateral agreements such as the Paris Declaration on Aid Effectiveness and the Accra Agenda for Action (OECD Development Assistance Committee, 2005, 2008).

The past decade has also seen several other institutional innovations that have helped program directors generate evidence and secure funds for evaluations from donors that expected rigor. The Jameel Poverty Action Lab (J-PAL) is a network of academic researchers across the globe established in 2003 at MIT, and has been running randomized control trials (RCTs) with a range of development agencies, as well as providing training for thousands of nongovernmental organization (NGO) staff. Innovations for Poverty Action (IPA) was founded in 2002, became an early partner with J-PAL, and is now an international nonprofit implementing RCTs with

³ In the context of the CGIAR, instead of individuals providing resources to charities, donor agencies represent the public and provide resources to CGIAR Research Programs.

⁴ Three parameters are modeled as determining donor funding decisions in Pritchett’s model. The first is the donor’s prior belief held in the effectiveness of the program; the second is the weighting of gains achieved through the program rather than through some other use of those public funds (i.e., opportunity cost); and the third is a measure of how persuadable (using communications materials) a given donor is. Pritchett illustrates this framework with reference to three stylized types of donors: core supporters, the middle ground, and hard-headed donors. Core supporters have prior beliefs in the program, weigh gains achieved through the program highly (as opposed to development gains achieved through other means), and are highly persuadable—they are “loyal” to the program. Hard-headed donors are the opposite and will give money only when they see rigorous evidence of impact—they have low prior belief in the program, place a low weight on gains achieved through the program rather than through some other use of those funds, and are not at all persuadable using communications materials. The middle group, as the name suggests, falls between these extremes on all measures.

researchers in countries around the world. Other academic initiatives soon followed (e.g. the Center for Effective Global Action (CEGA) at the University of California) and many donor agencies also increased their in-house capacity and requirements for rigorous impact evaluations.

In addition, advisory bodies such as GiveWell, New Philanthropy Capital, and ImpactMatters were created to foster greater effectiveness in the rapidly growing, multibillion-dollar NGO industry. These institutional innovations were underwritten by advances on the web and the data revolution that have taken place over the same period. Indeed, there is now a global movement for “effective altruism” supported through a growing online community. Their goal? To ensure that a commitment to helping others is married to an equal commitment to ensuring that such “help” does indeed help.

Impact evaluation answers a rather different question from ex post impact assessment in the tradition of CGIAR, and it denotes a different methodological toolkit. Impact evaluation seeks to answer the question: What is the impact (or causal effect) of a program on an outcome of interest? (Gertler et al., 2011). In the case of CGIAR, the causal effect of a new technology or management practice would be estimated using experimental methods (RCTs) or econometric methods that seek to emulate the conditions of a randomized experiment (quasi-experimental methods). Perhaps the most fundamental difference between impact evaluation and CGIAR tradition of ex post impact assessment is that the former requires rigorous estimation of a counterfactual,⁵ whereas the latter was more prag-

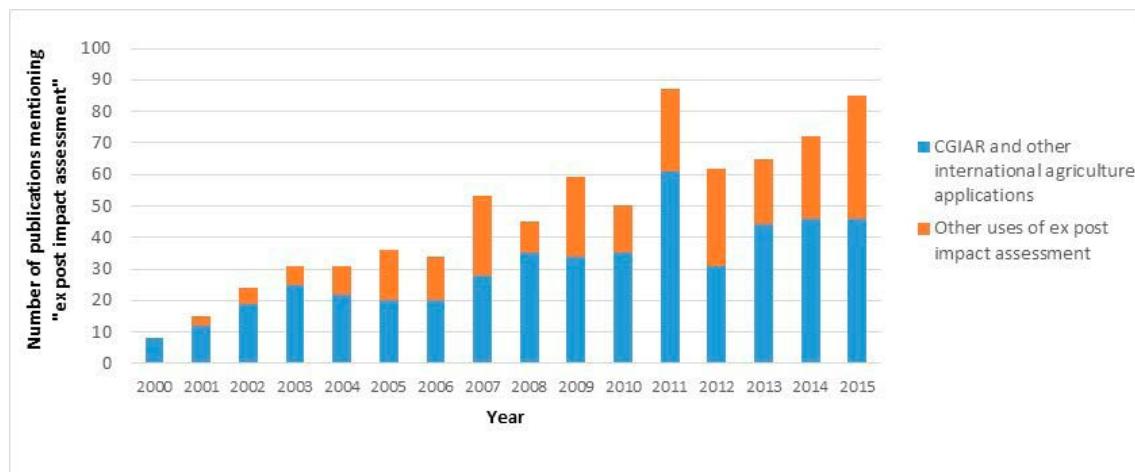
matic and willing to make stronger assumptions about a counterfactual.

As Figure 2 shows, ex post impact assessment is a methodology (and terminology) particular to CGIAR and international agricultural research. Within the sphere of international development, agriculture has lagged behind health, education, and social protection in adopting impact evaluation as a methodology (Figure 3). Of course, the share of aid allocated to agriculture research is dwarfed by the investments in health and education. Moreover the disparities between sectors in part reflect the long-established practice of medical trials with double-blind randomized treatment assignment to study the effectiveness of a new health or nutrition treatment. Those differ from the more recent surge in the use of RCTs by economists to evaluate development interventions, which often specifically aim to study and incorporate behavioral adjustments to understand their impacts. Until recently, agriculture also lagged behind on such evidence, however, possibly in part because the complexity, dynamic nature and inherent uncertainty related to farmers’ decision making make them a particularly challenging population to study. Nevertheless, as the standards of evidence from other sectors have started to percolate through donor agencies, agricultural research has been challenged to respond to this dramatic change in expectations for what constitutes rigorous evidence of impact.⁶ Recent examples show how researchers can properly account for the particularities of the agricultural sector in the design of RCTs (de Janvry, Sadoulet and Suri, 2017) and other impact evaluation studies.

⁵ In impact evaluation, the counterfactual is what the outcome would have been for program participants if they had not participated in the program. It is a concept that is central to any attempt at causal inference.

⁶ Referred to by Angrist & Pischke (2010) as the “credibility revolution” in economics and by Pearl and Mackenzie (2018) as the “causal revolution” for the field of causal inference in computer science and machine learning.

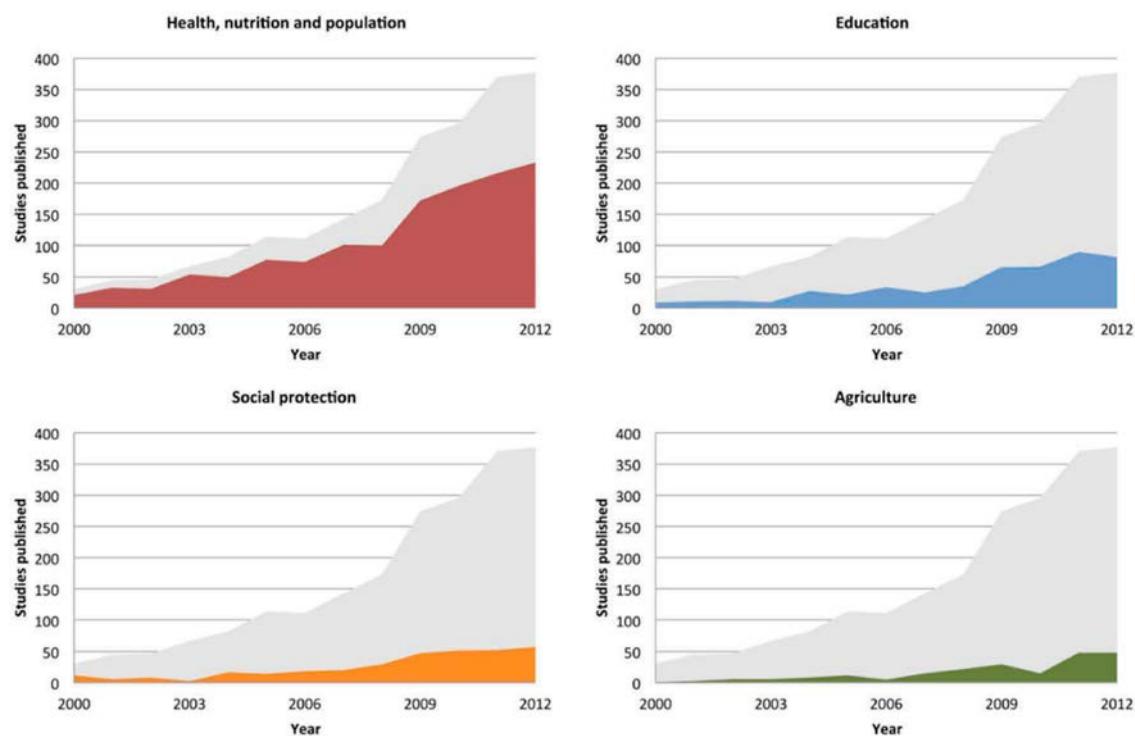
Figure 2. Use of ex post impact assessment in international agricultural research compared with other disciplines, 2000 - 2015



Source: Authors

Note: Figure shows results for Google Scholar search (carried out in Dec 2016) for numbers of publications containing the phrase "ex post impact assessment" by year. Publications were screened for their content to determine whether they related to international agricultural research or not.

Figure 3. Number of new impact evaluation publications by sector, 2000–2012



Source: Cameron, Mishra, & Brown (2016).

Note: Grey segment shows annual total of impact evaluations across all four sectors, with color band in each figure showing the relative contributions from each sector.

1.3 DOES IT STILL PAY TO BE IGNORANT?

Between January 2013 and July 2017, the CGIAR Standing Panel on Impact Assessment implemented a program called Strengthening Impact Assessment in the CGIAR (SIAC). This synthesis report summarizes how the program has grappled with the dramatic changes in impact evaluation in the aid sector and shows how SIAC has attempted to change how impact assessment is carried out in CGIAR. In this report we have built on the initial revolution in impact evaluation, and have taken the lessons in areas important for CGIAR to improve the quality of the evidence base it uses to make the case for its impact.

We consider the rigor of impact estimates to be a function of three properties of the studies from which they are drawn:

- accurate and valid measurement of treatment (i.e., agricultural technology use) and outcomes (e.g., productivity, poverty, nutrition);
- a research design that allows for an unbiased causal relationship between treatment and outcomes; and
- the extent to which the estimates are statistically representative at scale

The work from the SIAC program, as well as selected external papers, that illuminate these properties and their importance, are the focus of sections 2 (measurement), 3 (causality), and 4 (statistical representativeness). Section 5 concludes.

2 MEASUREMENT MATTERS

"As has often been remarked, probably no two individuals are identically the same. . . [W]hen the eye is well practiced, the shepherd knows each sheep, and man can distinguish a fellowman out of millions on millions of other men."

—Charles Darwin, *The Variation in Animals and Plants under Domestication* (Darwin, 1868, 1:361)

As noted in section 1.1, CGIAR has now oriented its activities toward a Strategy and Results Framework (SRF), comprising 3 high-level goals (reducing poverty, improving food and nutrition security, and ensuring environmental sustainability) with 10 development outcomes nested immediately underneath. To be fully integrated into the CGIAR portfolio, all research programs must have causal pathways that are theorized to flow through at least one of these development outcomes and then contribute toward one of the high-level goals. Testing whether such processes are indeed happening in reality requires a combination of valid measurement of "treatment" (outlined in sections 2.1–2.5); valid measurement of development outcomes (section 2.6); and a research design that allows us to rigorously uncover causal relationships by controlling for the myriad confounding factors (section 3). These are big challenges that are essential for CGIAR to meet.

2.1 EXISTING DATA ON ADOPTION OF IMPROVED VARIETIES: FIT FOR PURPOSE?

Genetic technologies, in the form of improved varieties of major food crops, lie at the heart of the history of CGIAR, as well as its ongoing comparative advantage in the global market for agricultural research. The early period of the Green Revolution in Asia in the 1960s and 1970s was underwritten by a huge turnover of genetic material in farmers' fields. Semi-dwarf varieties of wheat and rice, bred by scientists working for the nascent International Maize and Wheat Improvement Center (CIMMYT) and International Rice Research Institute (IRRI), spread rapidly through the irrigated wheat and rice production systems of several Asian countries (Dalrymple, 1978). The adoption of these improved varieties represented a significant shift, from the traditional tall-standing varieties that put much of their energy into vertical growth (and therefore relatively less into the production of grain) to shorter plants that had a much higher yield of grain per unit of area. The improved varieties were immediately noticeable to the naked eye—they looked different.

While the initial improved varieties from CGIAR were easy to identify in the field,

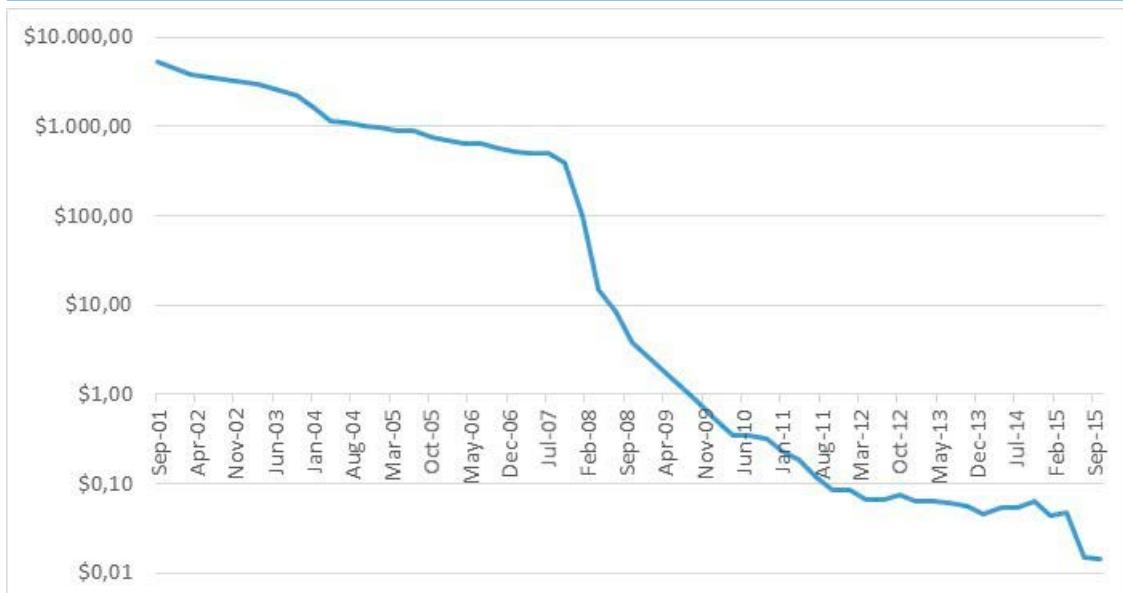
improved varieties from other crops were often less visibly distinct. Perhaps more important, further improvements on secondary target traits for breeding built on the original high-yielding varieties. This has led to a situation where it is difficult to identify varieties in the field—not only new generations of improved varieties compared with older generations of improved varieties, but also improved varieties compared with landraces.

This diversification of breeding effort across crops and across traits poses a deep challenge to understanding the adoption of new varieties in farmers' fields. Reliable data on adoption of improved varieties have long been recognized as the cornerstone of any assessment of the impact of investments in plant breeding (Walker & Crissman, 1996; Walker et al., 2008). Adoption data have always been scarce, and yet obtaining such data has, as we have outlined, in some ways become even harder over the course of the past few decades. Darwin may

have been considerably less optimistic about man's ability to distinguish variation within a population had he studied maize in farmers' fields in sub-Saharan Africa.

Fortunately, disruptive technological change, with the development and commercialization of "next generation sequencers," has pushed down the cost of sequencing DNA so sharply over the past decade that the technology has beaten Moore's law (see Figure 4).⁷ In just two years—2007 to 2009—the cost to sequence a million base pairs fell from the high hundreds of dollars to less than one dollar. We are now at a point where the laboratory costs are manageable for all but the smallest research project, and the use of genotyping can be considered a core part of the methodological toolkit for impact assessment. SPIA has invested a significant amount of effort in understanding how to use this tool, and a number of experiments have been commissioned under the SIAC program.

Figure 4. Cost of sequencing one million bases of DNA, 2001–2015



Source: National Human Genome Research Institute (2016).

Note: Cost is given in constant US dollars on a logarithmic scale.

⁷ DNA sequencing is the process of determining the specific order of nucleotides – the bases, or “building blocks”, namely adenine (A), guanine (G), cytosine (C) and thymine (T) – in a strand of DNA. DNA fingerprinting is the process of matching samples of genetic material collected from individuals to known reference profiles. In our case, this means that DNA extracted from samples of plant tissue from farmers' fields is compared to the DNA extracted from reference samples representing the universe of varieties that could be present in a specific country (or as close to this ideal as possible).

2.2 DNA FINGERPRINTING FOR VARIETAL ADOPTION: ESTABLISHING PROOF OF CONCEPT

There is a growing literature on measurement experiments in agriculture (e.g., Beegle, Carletto, & Himelein, 2012; Carletto et al., 2016; Kilic & Sohnesen, 2015)—empirical studies that collect the same data in multiple ways to test for consistency across methods. When one data collection method in a study is a clear benchmark for accuracy (as is the case for DNA fingerprinting), it is then possible to test the extent to which other methods for collecting data are just noisy (i.e., measurement error is classical) or biased (i.e., measurement errors are biased).

Under the SIAC program, SPIA commissioned a series of tests in different contexts to compare DNA fingerprinting data against expert opinion elicitation data (as commonly used in ex post impact assessment) as well as against a range of survey-based methods for eliciting adoption information (as commonly used in impact evaluation). For each experimental test, the aim was to collect varietal adoption data about the same crop in the same geographies in multiple ways and compare them using the DNA fingerprinting estimate as the benchmark or “gold standard.”

The designs of the eight data experiments commis-

sioned under the SIAC program are summarized in Table 1. There are four cassava case studies, which partly reflects the fact that fingerprinting is a simpler operation for clonally propagated crops, so these are good candidates for testing. Sweet potato follows a similar logic, and beans and rice are both largely self-pollinating, whereas maize is a particularly complex case that requires careful handling (e.g., there are many hybrids with similar parentage; out-crossing occurs in the field).⁸

The full results for these experiments are at various stages in the peer-review process so we cannot provide full details here (but see Maredia et al, 2016, for results on cassava in Ghana and beans in Zambia; and Kosmowski et al, 2018 for results on sweet potato in Ethiopia). However, we highlight two key messages. First, we find extensive mismatching between the DNA fingerprinting results and both expert opinion estimates and farmer self-reported data. Errors occur in both directions—in some cases farmers and experts significantly overestimate the aggregate level of adoption of improved varieties, and in others they underestimate. To date we have seen no clear pattern that would allow using a simple deflation or inflation of these other kinds of data to be good proxies. Second, as Figure 5 shows, accurately identifying individual varieties (often the data needed to document adoption and diffusion and to understand impact pathways) may require DNA fingerprinting in many settings.

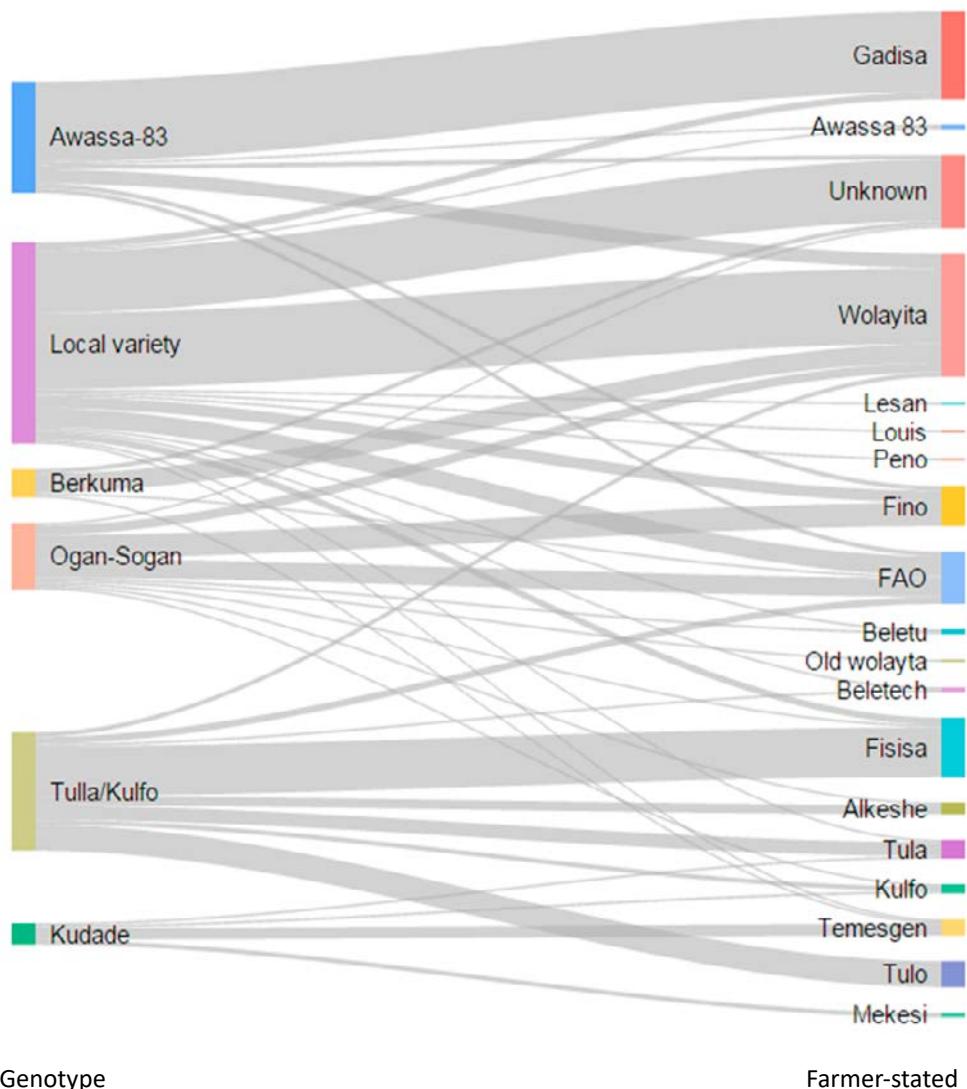
⁸ In clonally propagated crops, plants are produced using material from a single parent and as such there is no exchange of genetic material so plants are essentially identical to the parent. In self-pollinated crops, pollen from the male parts of the plant (anthers) fertilize the female (ovum, via the stigma) of the same plant. In cross-pollinated crops such as maize, the anthers of one plant release pollen that is blown by wind (or pollinating insects) to the female of other individual plants. Hybrids are the product of cross-pollinating two compatible in-bred (successively self-pollinated) lines of the crop to make seed that, when grown out, will display uniformity for most traits in the first season after planting, but for which uniformity will be lost in each successive season for which the seed is recycled (thus farmers are advised to purchase new hybrid seed each year).

Table 1. Experimental design for a series of eight new measurement experiments

CROP	COUNTRY	SAMPLE SIZE	SAMPLE DRAWN FROM	DNA FINGERPRINTING BASED ON SAMPLES OF: FINGERPRINTING BASED ON SAMPLES OF:	A: EXPERT OPINION ESTIMATION	SURVEY-BASED METHODS USING FARMERS' SELF-REPORTED DATA		
						B: ASK THE FARMER "IS THIS AN IMPROVED VARIETY OR A LOCAL/TRADITIONAL VARIETY?"	C: ASK THE FARMER "WHAT IS THE NAME OF THIS VARIETY?" AND THEN ATTEMPT TO MAP LOCAL NAMES TO A KNOWN IMPROVED VARIETY.	D: ASK THE FARMER ABOUT PHENOTYPIC CHARACTERISTICS OF THE VARIETY (E.G., LEAF SHAPE; GROWTH HABIT) AND THEN MAP TO SET OF "REFERENCE RESPONSES" PROVIDED BY BREEDERS OF THESE VARIETIES.
Maize	Uganda	550	Iganga and Mayuge; random	Grains	Yes	Yes	Yes	Yes
Sweet potato	Ethiopia	231	Wolayita zone; snowball	Leaf		Yes	Yes	Yes
Cassava	Malawi	1,200	National; random	Leaf		Yes	Yes	Yes
Beans	Zambia	855	Northern and Muchinga provinces; random	Seed		Yes	Yes	
Cassava	Nigeria	2,500	National; random	Leaf		Yes		
Cassava	Vietnam	1,570	National; random	Leaf	Yes	Yes	Yes	
Rice	Indonesia	798	Lampung province; random	Seed	Yes		Yes	
Cassava	Ghana	914	Brong Ahafo, Ashanti, and Eastern regions; random	Leaf		Yes	Yes	Yes

Note: Cells shaded grey denote that this data collection method was not part of the experiment.

Figure 5 – Correspondence between farmer-elicited data on varieties and genotype as established through DNA fingerprinting for sweet potato varieties in Ethiopia



Source: Kosmowski et al. (2016).

2.3 OPENING PANDORA'S BOX? OR A GOLD MINE?

The findings of these studies raise questions about the accuracy of the accumulated stock of knowledge about varietal diffusion to date. While it has simply not been possible to carry out this kind of empirical check on our methods before, methodological advances now allow to update the knowledge base. Hence, rather than being too pessimistic about what these results mean, we highlight two examples of how integrating DNA fingerprint-

ing into impact studies can expand the kinds of questions we can hope to answer.

Self-reported data can be biased against the technology

Wossen et al. (2017) carried out a nationally representative sample of 2,500 cassava producers across Nigeria, asking for farmers' self-reported data about varieties in use, as well as collecting samples of leaf material for DNA sequencing. Those authors find a significant proportion of errors in the

self-reported data: 28 percent of responses were false negatives (farmers thought they were growing local varieties when they were actually growing improved varieties), and 13 percent were false positives (vice versa). Furthermore, Wossen et al. show that the likelihood that farmers misreport varietal status is not independent of observable household characteristics. That is, farmers who are better educated and have access to more sources of information about varieties are more likely to provide accurate data. This has significant implications for analysis of the impacts of improved cassava varieties, as it reveals an additional source of endogeneity in studies that attempt to uncover the impact of agricultural technologies independent of the characteristics of the farmers that choose to use them. When the same model linking productivity to varietal status is estimated twice—once using self-reported data on varietal status and a second time using DNA-fingerprinted varietal status—the productivity advantage of improved varieties over landraces goes up 18 percentage points (from 42 to 60 percent).

The kind of measurement error uncovered by Wossen et al (2017) is referred to as non-classical measurement error (NCME). The data are not just noisy, but exactly because measurement error is correlated to other characteristics may lead to biased empirical estimates. A recent paper by Abay et al (2018) takes this a step further and looks at correlated NCME when measurement error is present in more than one of the key variables. In such a scenario, when we correct for NCME in one of the variables (e.g. by introducing a more reliable data collection technique such as DNA fingerprinting instead self-reported data on crop varieties), we may still have NCME present in other variables. If NCMEs in the original data for both variables were correlated with each other, only correcting for one of them, we may unintentionally bias the outcome of research even further. Abay et al (2018) use the example of agricultural yields and correcting for plot size NCME (using GPS or compass and rope) but not for harvests (still using self-reported data). We should hence not be overconfident in our inferences from advances in measurement that only apply to a subset of the variables we are interested

in. Whenever possible, data quality should therefore be tackled systematically on all fronts simultaneously, by doing fewer surveys and doing them better (as previously argued by Doss 2006).

Adoption as a continuous rather than a discrete outcome?

When seed purchased by farmers is impure (with multiple varieties inadvertently or deliberately mixed together in the seed system), or when farmers choose to cultivate multiple varieties in a single plot (mixing the seed from multiple sources or varieties together before planting), it challenges the concept of “varietal adoption” as a discrete binary decision that can be neatly analyzed in an econometric model. Indeed, an emerging literature on input quality in African agriculture examines how the same issues may apply not just to seeds but also fertilizers (Bold et al., 2017; Fairbairn et al., 2016) and herbicides (Ashour et al., 2016).

Given that seed impurity and/or farmer mixing is likely to be the reality for many agricultural plots in sub-Saharan Africa, this argues for a more quantitative and rigorous approach to measuring adoption and diffusion of new genetic technologies used by farmers. DNA fingerprinting has the potential to open up productive new directions for the study of how genetic, environmental, and management factors interact with farmers’ behaviors and decision-making to determine agricultural outcomes. There is much work to be done to take full advantage of these opportunities.

2.4 NATURAL RESOURCE MANAGEMENT: HUGE SCOPE FOR IMPROVED DATA COLLECTION

An under-evaluated research portfolio

Research on natural resource management (NRM) technologies and practices now represents a significant proportion of total investment in CGIAR. However, there have been few efforts to track adoption of NRM technologies or practices at large scale (Erenstein & Laxmi, 2008). As previous re-

views have highlighted (Renkow & Byerlee, 2010), one possible explanation for this lack of attention to tracking adoption has been a lack of clear methodology. For example, the Food and Agriculture Organization of the United Nations (FAO) compiles global estimates of the area under conservation agriculture, but these data are often based on the opinion of a single expert in each country in a manner even more ad hoc than has been the norm for varietal adoption data. The spread of information technologies—such as the improved global coverage of satellites equipped with sensors of ever greater acuity and the continuous increases in saturation of cell phones in every country around the world—suggests that it should be possible to use these technologies to help gather better data on adoption of NRM technologies.

Certainly, there is no single gold-standard method for measuring adoption of NRM technologies that could serve as a reference, and many NRM technologies are complex bundles of practices, including some that are directly observable at a single point in time (e.g., whether a field has been plowed or not) combined with others that are dynamic practices (e.g., crop rotation). There is, however, significant potential to harness the dynamic technological change taking place in data science in the service of assessing the technological change taking place in agricultural development. Indeed, the CGIAR platform on Big Data was conceived with exactly this goal in mind.

Remote sensing and related information technologies are in their infancy with regard to their application to adoption and impact studies, so more time and piloting are required. Yet this message may be too conservative and cautious about an area of dynamic innovation. We are increasingly unconstrained in our ability to source useful remotely sensed data. For example, the European Space Agency is making vast quantities of data from its Sentinel 2 missions—satellites fitted with sensors capable of monitoring a range of agricultural practices from space—freely available in the public domain. The limiting factor is often our capacity to handle, process, and make sense of the petabytes of data from a sophisticated range of

sensors, but as competition continues to emerge in this field, the costs of private sector services are likely to drop over time.

A first step: A series of pilots

Remote sensing has been used to identify candidate geographies for NRM interventions for many years. However, such *ex ante* assessments offer only limited guidance to the challenges that arise in applying remote sensing to the real-time tracking, or *ex post* impact evaluation, of the adoption of specific technologies. To learn more about new measurement approaches to tracking adoption of specific practices, we reviewed all the adoption claims made in every CGIAR center annual report for the 10-year period from 2004 to 2013 (the start of the SIAC program). From this mass of information about potential research successes, we focused on five NRM technologies that have been the subject of sustained research funding and attention over the years:

- Conservation agriculture (primarily in maize-based systems; CIMMYT/International Crops Research Institute for the Semi-Arid Tropics [ICRISAT])
- Fertilizer trees (as a specific agroforestry technology; World Agroforestry Centre [ICRAF])
- Fertilizer micro-dosing (primarily in maize-based systems; ICRISAT)
- Alternate wetting and drying (in rice-based systems; IRRI/International Water Management Institute [IWMI])
- Integrated soil fertility management (primarily in maize-based systems; International Institute of Tropical Agriculture [IITA])

Measuring the adoption of each of these practices poses specific challenges. For example, the ability to monitor the extent of tree cover on farms is quite well established (Zomer et al., 2016). This information does not, however, tell us how the farmer manages the trees—which would build confidence that this farmer had adopted the practice of agroforestry—nor about the species or the kinds of benefits that accrue to the farmer. Alternate wetting and drying in rice relates to a change in the irrigation regime, so measuring adoption requires

capturing changes in periodicity of flooding and implies a choice to change from a more intensive use of irrigation water. Similarly, a farmer's conscious choice to practice fertilizer micro-dosing must be distinguished from another farmer's low average fertilizer application rates by paying close attention to the spatial concentration of the limited fertilizer available. Conservation agriculture and integrated soil fertility management are complex technologies with multiple components operating in tandem.

The empirical results of the adoption studies outlined above, and more reflection on the methodological lessons learned (as debated in a recent workshop⁹), are reported in detail in a subsequent synthesis report (Stevenson and Vlek, 2018) specifically on NRM research outcomes.

2.5 DISADOPTION

Measuring disadoption is arguably much more important than most agricultural researchers acknowledge. As social scientists have begun to explore more carefully the processes by which farmers learn about technologies, there is much to be gained from studying disadoption patterns more systematically and rigorously. Such patterns reveal where new technologies and practices failed to deliver returns for the farmers who tried them. CGIAR should therefore place much more emphasis on understanding the reasons for such failures. In many ways, disadoption is more challenging for impact assessment as it is harder to establish a good counterfactual through randomization. Moreover, the institutional incentives for individual centers to pursue such an endeavor may be unclear, so again this argues for SPIA (as a system-wide entity) to facilitate the kinds of longitudinal country-level analyses that can help to understand these dynamics.

2.6 OUTCOME VARIABLES: ARE WE MEASURING WHAT WE THINK WE'RE MEASURING?

Validity of indicators

As Jerven (2013) exposes in his ethnographic study *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*, while there is considerable heterogeneity across the continent, on average African governmental statistics agencies suffer from low political independence, misaligned staff incentives for rigorous work, and inadequate financial and human resources. This first-order concern about the veracity of official statistics at aggregate levels is troubling but is endogenous to the politics of the continent and unlikely to change in the medium term. And yet more fundamentally, even in cases where statistics are more reliable, we need to be cautious in accepting specific indicators as valid measures of the concepts we are interested in.

Consider the objective of increasing agricultural productivity – as previously discussed, an integral part of the strategy of CGIAR to date. Given the myriad factors other than agricultural research that influence productivity, it is safe to assume that CGIAR will need a careful approach to measuring productivity in order to detect small marginal improvements over a baseline level. Data experiments can help show how much the approach to data collection influences the validity of indicators, as for the DNA fingerprinting cases described in section 2.2. In the case of productivity, we are concerned with measuring output and input not only accurately but also cost-effectively. Gourlay, Kilic, & Lobell (2017) compare self-reported maize yields to crop-cut estimates and remote-sensing productivity estimates for the same plots in Uganda. The remote-sensing data are more closely related to the data from the crop-cuts than to the self-reported data. Examining the distribution of the self-reported production data shows that significant rounding errors are associated with those data. When yields are comparatively small, these rounding errors

⁹ "Managing Natural Resources for Sustainable Production Systems: A Research Agenda at the Crossroads?" Organized by CGIAR Standing Panel on Impact Assessment (SPIA) and CGIAR Research Program on Policies, Institutions and Markets (PIM), 26th – 28th February, 2018, IFPRI, Washington DC

matter a lot.

The literature on accurate plot area measurement [is summarized in a guidebook](#) by the Living Standards Measurement Study (LSMS) team at the World Bank (Carletto et al., 2016). In the 1970s the gold standard for area measurement was to use a compass to measure every angle around the perimeter of a plot and rope to measure the length of each side. With this process, it takes an average of one hour to collect the data from a single plot. At the other extreme, simply asking the farmer to state the area of the plot is easy and fast—it takes only a few seconds to ask the question and record the response. A compromise approach is to take GPS coordinates by walking the perimeter of the plot. The data quality of this approach is high—the correlation coefficient with the compass and rope is almost 1 but importantly, walking the plot with GPS takes much less time (≈ 15 mins) than compass and rope, and may be less susceptible to enumerator error. GPS also has the tremendous advantage of being modern, which may make it more attractive to forward-looking ambitious senior officials in statistics agencies.

While clear empirical standards for best practices are emerging for some variables, many outcomes described in the CGIAR Strategy and Results Framework (Figure 1) are either poorly defined and / or are conditioned by context. Resilience, market access, food security, natural capital, ecosystem goods and services, agricultural practices – all of these are examples of outcomes for which we could reach for any number of definitions and accompanying measurement approaches. Therefore, within CGIAR (and the broader literature) these concepts get operationalized in a wide range of ways, some

of them appealing, but others much less so. Arguably, CGIAR has a comparative advantage in advancing theoretically-grounded, precise measurement or estimation of some of these concepts. Unfortunately, as attested by external reviews by high-level panelists (e.g. Barrett et al, 2009; CGIAR Independent Science and Partnership Council, 2012), CGIAR has largely yielded this challenge to others, including FAO, donor agencies and consultants.

Nutrition outcomes are one area of research outcomes with a long tradition of testing the validity of the metrics used. A standard approach to measurement can and has been established, partly because social context is not necessarily seen as important in shaping the relationship between dietary intake and nutrition outcomes. As Steel (2007) has outlined, biological causal mechanisms are predictably “transportable” from one setting to another – at least more than is the case for those that determine social outcomes. But metrics for social or complex ecological phenomena are unlikely candidates for any future universal standard for data collection and some will always need to be defined in context specific ways, instead of aiming for standardized metrics. That said, there is ongoing work to identify better ways to measure certain social outcomes such as women’s empowerment in ways that can be compared across contexts, and to reduce the cost of using those best-practice metrics that have been tested and confirmed. IFPRI researchers have been instrumental in the development of the [Women’s Empowerment in Agriculture](#) (WEAI) index, for example. The impact assessment community in CGIAR should be aware of these and the related opportunities for improving practice in CGIAR.

3 CAUSALITY AND BIAS

3.1 RANDOMIZED CONTROL TRIALS: A POWERFUL METHODOLOGY TO BE USED WISELY

Adoption is a choice

Establishing whether a farmer's outcomes improve because she adopted a new technology is fundamentally difficult because in almost all cases the farmer self-selected into using this technology. She presumably had a good reason to do so, and her decision-making likely included consideration of a great many factors, such as the availability of alternative technologies; complementarity with her soil; her land and labor endowment; her access to other inputs, credit, or insurance; her access to output markets; trade-offs between higher yields and more risks; or food security considerations. She is likely to have imperfect information about many of these aspects; she needs to account for uncertainty related to weather, pests, prices, or health shocks; and she must factor in potential dynamic gains from learning. She will likely draw on her past experiences to make inferences about some of these uncertainties, and she may make mistakes in the process. The probability of making mistakes may depend on her skills and experience. These and many other factors are being considered by all farmers potentially exposed to a new technology. In any given season, some of them end up adopting whereas others do not. Comparing outcomes for farmers who adopt with those who do not adopt will lead to a fundamentally biased estimation of the gains from adoption. This follows simply because those who decided to use a new technology did so because they expected it to be beneficial for their particular case and in their particular circumstances. Given the sheer magnitude of factors that enter the decision-making process, it is almost always impossible for the empirical researcher to take these factors into account ex post.

Some quantitative empirical methods are built on the assumption that we can observe, and therefore control for, the major factors that condition farmers' decision to adopt. Such "selection on observables" methods, such as propensity score matching (PSM), are hence particularly ill-suited to shed light on the impacts of agricultural technologies. This was clearly argued by de Janvry, Dustan, & Sadoulet (2011) in a paper commissioned by SPIA. De Janvry, Dustan, & Sadoulet argued for microeconomic impact analysis with explicit research designs based on either natural or randomized experiments. In some cases, institutional knowledge about the rollout of a new technology may provide a researcher with natural temporal variation that can be exploited to identify impacts, when verification of

the plausibility of the underlying assumptions is feasible. In other cases, geographical discontinuities or external factors driving technology availability in ways unrelated to potential impacts can help establish counterfactuals. In short, impact evaluations should seek exogenous sources of variation in access to technologies and need to be able to document the origins of variation in order to support the assumptions underlying the empirical estimates.

A particularly powerful way to solve the self-selection problem is for researchers to work with development partners (NGOs, government agencies, agro-dealers) to create an experimental variation in which some farmers or villages have access to a new technology and others do not—i.e., setting up a randomized control trial. If such manipulation is random or quasi-random and properly accounts for potential spillovers (i.e., people getting access to, or benefiting from, the technological choices of others), this can provide the researcher with a credible counterfactual. If enough farmers with access to the technology subsequently decide to try the new technology, its impacts can then be estimated. The latter is a non-trivial condition, however, as take-up rates are often low, and temporarily subsidized (or free) provision may be needed to obtain sufficiently high take-up rates.

No causes in, no causes out

Randomization of an intervention offers the possibility that it is a statistically independent (orthogonal) factor and that it is the only variable that is different between the treatment and control group. All methods that aim to draw causal conclusions require causal identification assumptions (or, as philosopher Nancy Cartwright succinctly puts it: “No causes in; no causes out”) (Deaton and Cartwright, 2017). In RCTs, the treatment assignment is specifically manipulated by the researchers so that

the orthogonality assumption holds in expectation, which is the central advantage of the method of RCTs.¹⁰ However, randomization in and of itself does not guarantee orthogonality, but instead only buys balance between treatment and control *in expectation*. Hence with limited sample sizes, there may be some imbalance on observable or unobservable factors that remain different between the two groups after randomization. Unhappy randomizations can happen, particularly when sample sizes are relatively small. Deaton and Cartwright (2017) therefore argue that the orthogonality assumption must be defended on a case by case basis, and RCTs studies often include checks for balance on observables. Focusing specifically on RCTs in agriculture in developing countries, Barrett and Carter (2010) also point to the importance of correctly characterizing environmental and structural conditions in which RCTs are conducted, crucial for drawing the appropriate inference from the often highly-stylized experimental designs.

Another key set of assumptions (that collectively are referred to as the Stable Unit Variable Assumption or SUTVA) needed to derive appropriate inference from RCTs (and any other identification method) are spelled out by Cook (2018) in his deconstruction of the conditions under which RCTs warrant the label “gold standard” that they often attract: “... *the control group does not include any dimensions that are meant to be unique to the treatment, for this will reduce the size of the planned treatment contrast; ... there is not “compensatory rivalry”, as when the control group responds to not getting the treatment by trying harder than it would otherwise have done, also called a “John Henry” effect; there is not “compensatory equalization”, as when an administrator observes the unequal distribution of resources that an [RCT] requires and tries to stifle any anticipated resulting discord by providing extra resources to the control group...; ...the control group does not become demoralized through learn-*

¹⁰ This point is noted in the commentary on RCTs by Oakes (2018) worthy of quoting at length as follows: “*RCTs require researchers to do something. The actual experimental manipulation of a policy, drug, or environment is greatly underestimated. I maintain that too much of our collective research portfolio is devoted to observational studies with weak identification strategies and regrettable analytic approaches, including p-hacking. The amount of correlational health and social research is overwhelming. And a vast proportion of it involves what I call hypothetical interventions, such as changing poverty rate, eliminating discrimination, establishing healthcare universal, or making higher education free. While laudable, such analyses are often red-herrings. I would prefer more real life manipulations, more scientific transparency, more data sharing, and more meta-analyses for key questions.*” Oakes (2018, p.2)

ing they have not been favored with the intervention – a process that entails a causal direction from the control group to the outcome rather than from the treatment to the outcome, as intended.” (Cook 2018, p.3).

As such, violation of SUTVA may seem inevitable in RCTs that relate to agricultural technology. As Imbens (2018) argues, however, such violations may sometimes simply need to be taken into account, or in other contexts may actually be the main focus of the analysis. Saturation designs that experimentally vary the density with which access to a new specific technology is offered, for instance, aimed to do exactly that, and can lead to important insights regarding social learning and diffusion (Baird et al., 2012; Glennerster and Suri, 2015).

The recent chapter by de Janvry, Sadoulet, and Suri (2017) in the Handbook of Field Experiments discusses in more detail these and other considerations for applying RCTs in agriculture and provides specific guidelines of how to avoid common pitfalls and maximize lessons learned.

Behavioral adjustments matter

The goal of a good impact evaluation is to establish not only whether a specific technology improved outcomes, but also how and for whom. Given the complexity of farmers’ decision-making as already outlined, behavioral responses to new technologies can be at least as, or even more, important in determining development outcomes as the improvement embedded in a technology per se. For example, Bulte et al. (2014) show that households adjust labor efforts when they are knowingly testing new technologies, but not otherwise. Given that we are interested in impacts in real life, we do not want to “switch off” such adjustments.¹¹ Rather, impact evaluations ought to be designed to measure the different potential behavioral adjustments implicit in the process of adoption (de Janvry, Sadoulet, & Suri, 2017). Behavioral adjustments by active economic agents with access to a new agricultural technology can be anticipated

(e.g. the treatment group, being unblinded, allocate more labor or other inputs than the control group) and should be measured to ensure that the most important adjustments or strategies do not go unobserved. Ultimately, it is the combination of the intended treatment and the behavioral response by those adopting that we are interested in knowing about from a policy perspective.

Emerick, Janvry, & Sadoulet (2016) provide a powerful case in point in showing that farmers who adopted the Swarna-Sub1 rice variety in India also adopted a more labor-intensive planting method and had greater cultivated area, fertilizer usage, and credit demand. It is through these behavioral responses that the returns to the new technology are substantially increased, and without a randomized control trial it would have been impossible to observe these adjustments and learn from them. Managing the quality of the design and implementation of randomized control trials is essential to avoid the pitfalls and to assure that relevant lessons can be learned from impact assessment. This includes careful consideration of the nature of the treatment that is being applied, and the population it is applied to, so that it can inform about impacts and the possible causal pathways in which it can affect outcomes beyond the specific case of the experiment.

3.2 HOW DO WE EVALUATE PROMISING TECHNOLOGIES?

Selection processes, conflicting objectives, and farmers’ desire to please

New technologies originating from CGIAR are developed with the objective of increasing food security, productivity, or resilience, often in combination with an environmental objective. Once these technologies leave the lab or the experimental station, it is often assumed they do exactly that. History provides some powerful examples where improved varieties played a major role in increas-

¹¹ This is sometimes done through double-blind designs, in which, for example, farmers are given seeds that they are told could be a new improved variety or could be a placebo of the same variety they were using previously—as shown in section 2, it is difficult to tell the difference.

ing yields and reducing the cost of production for smallholder farmers, most notably during the Green Revolution in Asia. Yet this type of large-scale and first-order impact has proven hard to replicate for other technologies and in other settings. A better understanding of whether, how, and under which conditions new technologies improve farmers' outcomes in real-life conditions is therefore key to help guide the technology generation process.

Identifying the target population and assuring that a new technology is suited for this target population and can contribute to the desired development outcomes are criteria that should enter the technology production process much more systematically. The methods used to select populations on whom new technologies are initially tested, and how the testing is carried out, might well limit the potential payoffs of the development of new technologies. New technologies are typically evaluated based on their potential for yield increases in controlled settings. While technologies are often subsequently tested in farmers' fields, agronomic trials conducted with farmers are often still highly controlled by the researchers, in order to maximize the agronomic insights. Yield gains obtained in such trials are typically compared with the costs of inputs to determine whether a certain technology holds promise. And the conclusions of such calculations guide not only dissemination efforts, but also further research efforts.

Yet there are multiple reasons to believe that yields gains obtained in typical agronomic trials are not very representative of yield gains the average farmer could obtain in real-world settings. Researchers might select certain types of plots and/or certain types of farmers for such trials, they may establish the agronomical practices to be applied, and they may even supply the specific inputs (seeds, fertilizer, pesticides) to be used. Moreover, farmers themselves might adjust their practices on the trial plots (for example, by being more careful about weeding), and hence applying more effort. Farmers may also learn from the process of participating in the trials, potentially increasing yields. And while trials are typically designed to maximize yields, farmers'

private decisions may be directed toward maximizing profitability, food security, or some other outcome, making the maximum yield gains potentially irrelevant.

Indeed, the yields obtained in such trials are hardly ever replicated in uncontrolled, larger-scale settings in farmers' fields. This was one of the main drivers behind the former Participatory Research and Gender Analysis (PRGA) program of the CGIAR in the early 2000s. Understanding the reasons for these yield gaps is key to understanding the impact of CGIAR technologies, because they can help explain potential reasons for lack of diffusion. Moreover, if trials can be designed and analyzed based on insights from the different farmer selection processes and farmers' behavioral responses, they may lead to different conclusions, and as such to point to different directions in the future research agenda.

Learning from on-farm trials

Focusing on these problems with agronomic trials, Laajaj et al (2018) study mechanisms through which the returns estimated from on-farm trials might not necessarily provide good estimates of gains from adoption in real-world circumstances in Western Kenya. They focus on the role of farmer and plot selection, but also on measurement questions, and the role of effort and complementary technical advice. Initial results from this study show large adjustments in yield and yield increment calculations when these different factors are taken into account. These results in turn help us understand the dynamic learning and adoption patterns by different types of farmers following the trials (Laajaj and Macours, 2016). Similar collaborations between CGIAR scientists and economists in the design and analysis of future trials will be useful to draw broader lessons and analyze different trade-offs and selection concerns. This may ultimately result in the development of new guidelines for on-farm trials that account for different selection and behavioral adjustments, so that broader lessons can be learned regarding the returns to different technologies for heterogeneous populations and conditions.

On-farm trials, apart from being important steps in the R&D process, are separately also often used as mechanisms to promote technology diffusion. Several extension approaches (such as “mother-baby” trials, volunteer farmer trainer programs, or farmer field schools) are designed to provide on-farm demonstration of new technologies by identifying and working with “lead farmers” in target communities. The theory has it that by carrying out trials of new technologies on their land, such lead farmers influence the behavior of others in the community and make it more likely that they will adopt the technologies. However, until recently the empirical evidence supporting such approaches was weak. A systematic review of farmer field schools, for instance, concludes, *“Farmer field schools can have beneficial effects for participating farmers, in pilot programmes in the short term. The impacts on agricultural outcomes may be of substantial importance to farmers, in the region of a 10 per cent increase in yields and 20 per cent increase in profits (net revenues). There is little evidence of diffusion of improved practices or outcomes from FFS participants to non-participating neighbour farmers... [T]here is no evidence that any diffusion of practices is sustained over time, nor any evidence for adoption of more complex IPM [integrated pest management] practices via diffusion”* (Waddington et al., 2014, 18–19). Unanticipated or unobserved heterogeneity between lead farmers and the rest of the community (i.e., the very factors that make them “leaders”) is likely a sufficient explanation for this situation.

Recent advances in the role of social networks for the diffusion of new technologies also points in that direction, but also provides more promising experimental evidence on the potential of diffusion when appropriate local agents of change can be identified on whose farm to conduct the trials, and that learning from more than one person can be important (Beaman et al, 2015). Along the same lines, demonstration through field days to observe and hear about experiences and outcomes with a new rice variety also can increase adoption (Emerick and Dar 2017). But social learning can lead to slower adoption than direct exposure if networks are segregated or small (Beaman and Dillon, forthcoming) and trial farmers that more closely resemble farmers targeted may be more effective (BenYishay and Mobarak, 2014 and Tjernstrom, 2017).

The primary objective of impact evaluation studies and subsequent systematic reviews should be to help CGIAR reach its goals in orienting new research toward areas with potentially high payoffs rather than areas too slanted toward the short-term priorities of donor agencies. Such a shift also requires putting in place feedback mechanisms that allow scientists to learn from the evaluations and adapt to their findings. In turn, the design of new impact evaluations should explicitly account for scientists’ own questions and concerns about the trade-offs implied by certain technologies.

Thus, establishing credible causal evidence requires a further shift toward planning that anticipates smart evaluation designs: once such designs are in place, they allow researchers to study both expected and unexpected behavioral responses and to understand the pathways to impacts as well as the underlying reasons for potential lack of impact. Credibly documenting and learning from such “zero” results is arguably even more important than establishing success stories, and hence deserves attention and recognition in CGIAR’s impact assessment portfolio:

3.3 APPROPRIATE METHODOLOGY IS THE GOLD STANDARD

The case for methodological pluralism

Well-designed and implemented randomized control trials provide more rigorous causal identification than a reliance on observational data in econometrics. Because the researcher directly manipulates the treatment assignment and hence by design can assure treatment assignment is not correlated with possible confounders, causal inference requires a more limited set of assumptions than alternative micro-econometric methods. Therefore, when the objective is to learn the impact of a newly developed technology at the micro-level and before it is widely diffused, RCTs can provide a level of rigor higher than that attainable with oth-

er methods. To assure that meaningful lessons can be learned from the RCT, it must be preceded by good diagnostics through preliminary qualitative work and piloting. Sometimes, however, the key questions will be about impact and effectiveness at scale across broad agricultural landscapes and over long periods of time. In these circumstances, other methods of impact evaluation may provide more relevant information, albeit at the cost of confidence in causality. Randomized control trials may be ill-suited, for instance, to document impacts of technologies that are already widely diffused. Measuring impacts of policy influence work or of institutional changes can also be challenging with a RCT, except where these have localized effects. Reflecting this reality, in 2012 DFID commissioned a group of evaluators to reflect on the practice of impact evaluation because it was widely believed that researchers had focused their methodologies too narrowly on randomized control trials that “*DFID has found are only applicable to a small proportion of their current programme portfolio*” (Stern et al., 2012, S4, i).

The DFID-commissioned study concluded that most of the development interventions it funds are “con contributory causes.” This language reflects a shift to a programmatic approach across the aid sector, away from simple provision of “treatments” and toward more complex designs that aim to address multiple development challenges. Evaluators use the term “causal packages”¹² to describe the confluence of causes that come together to produce outcomes in such cases. As the authors of the DFID report note, “*A reality that often has to be faced in [impact evaluation] is that there is a trade off between the scope of a programme and strength of causal inference. It is easier to make strong causal claims for narrowly defined interventions and more difficult to do so for broadly defined programmes. The temptation to break programmes down into sub-parts is therefore strong, however this risks failing to evaluate synergies between programme parts and basing claims of success or failure on incomplete analysis*” (Stern et al., 2012, S15, ii).

Research focused on policies and institutions (an important part of the CGIAR portfolio) typically does not have large populations of potential users/adopters as is the case for farmer-managed technologies. For these research investments, theory-based approaches—defined by White (2009) as “examining the assumptions underlying the causal chain from inputs to outcomes and impact”—offer much potential. In applying these evaluation approaches, and in judging the rigor of their causal claims, different standards of evidence should apply than in the case of interventions or technologies that can be randomly assigned or subjected to quasi-experimental econometric analysis (Rogers, 2009).

Two key challenges apply broadly across NRM, policy and institutions research areas, namely that there is rarely both: i) sufficient observations to identify effects rigorously (e.g. nation states adopting a new policy; watersheds taking on a specific approach to NRM), and ii) homogeneity in the “treatment” since contextual adaptation of institutions, NRM practices and policies is the norm, not the exception. This is a space where methodological advances with / by CGIAR can make a major contribution to the scientific community more broadly. Given the scale of CGIAR investment in policy, institutions and NRM, the dearth of careful impact assessment in this area is alarming. What is clear is that the methods that apply where these two conditions (small N, and varying or context-dependent treatments) apply will be quite different from situations where they do not. Defining what it means to do rigorous impact assessment in such cases is an urgent and important task.

Systematic reviews: Throwing out the bathwater while keeping the baby?

The challenges of methodological pluralism come into sharp relief in the process of systematically reviewing the state of knowledge on specific topics. Systematic reviews—which “synthesize the best available research evidence on a specific question” (www.3ie.org)—are increasingly influential

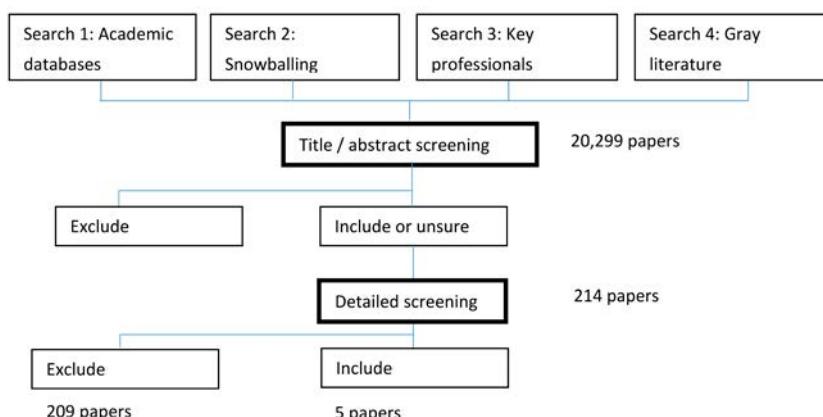
¹² The dominance of a quantitative, counterfactual-oriented view of causality is illustrated by the fact that when one enters the term “causal packages” into Google, the entirety of the first page of results is for econometric estimators to run in different statistical software programs.

in evidence-based policy in medicine (through the Cochrane Collaboration), social programs (Campbell Collaboration), and the international development sector writ large (International Initiative on Impact Evaluation, 3ie). By comprehensively and systematically searching the universe of published evidence on a specific question, researchers can systematically filter out studies that are of limited relevance and then filter out studies based on quality criteria. The devil is in the details of these quality criteria—too loose and we increase the risk that low-quality or biased studies will influence the overall result; too strict and we might throw the baby out with the bathwater.

Take the example of a study carried out by Loevinsohn et al. (2013). Their systematic review, commissioned by DFID, set out to answer the question

“Under what circumstances and conditions does adoption of technology result in increased agricultural productivity?” The authors screened 20,299 papers at the first stage, passing a healthy 214 through to the second stage (Figure 6). However, only 5 of these papers passed the second-stage screening and were candidates for in-depth review. This second-stage screening eliminated the vast majority of studies on the basis of a single criterion: whether a functional definition of adoption could be determined from the paper. Although this criterion is undoubtedly an important aspect of designing a good impact assessment of agricultural technology, it is not obvious that rejecting 202 out of 214 papers on the basis of that criterion alone is a constructive or efficient strategy (the remaining seven rejected papers being omitted based on other criteria).

Figure 6 – Screening process carried out by Loevinsohn et al. (2013)



Source: Authors based on Loevinsohn et al. (2013).

Herdt and Mine (2017) reviewed the Loevinsohn et al. report as part of the SIAC program and found that much useful evidence was filtered out unnecessarily. Using a slightly less restrictive set of criteria, Herdt and Mine retained 30 studies—still a considerable cull of 184 papers that were relevant but simply not of sufficient quality for inclusion. Of the 30 studies retained by Herdt and Mine, 23 had data related to agricultural yields, and of these, 21 demonstrated a positive relationship between technology use and productivity (the other 2 showed no difference). Of 26 studies with data on incomes, 24 showed a positive relationship (again,

2 showed no difference). These results are subject to two significant caveats: First, the technologies subjected to impact assessment are typically cherry-picked in the first place. Second, before we even observe this population of published studies, they are subject to the “file-drawer” problem (self-censoring by researchers) and publication bias (reflecting a preference for positive or attention-grabbing findings by journals). Together these factors certainly bias the distribution of published results upward relative to a strategy of selecting, evaluating, and publishing assessments from a random sample of candidate technologies.

Stewart et al. (2015) carried out a systematic review of the impacts of training, innovations, and new technologies for African smallholder farmers. From a very large screening, they ultimately retained only 19 studies owing to a “lack of rigorous research evidence” (Stewart et al. 2015, p.6). Of the 19 studies retained, 5 studies on orange-fleshed sweet potato show a consistent pattern of positive impacts on nutritional indicators. Given the methodological diversity of the retained studies—a mix of RCTs and econometric analysis from observational data—an excellent feature of the Stewart et al. systematic review is the process through which the authors score the studies for risk of bias arising from confounding, selection problems, departures from the intended intervention, missing data, measurement problems, and selective reporting of results. The authors provide summary judgments for the risk of bias: low (as for a well-implemented RCT—9 out of 19); moderate (sound for a non-experimental study—2 out of 19); serious (has some important problems—6 out of 19); and critical (too problematic to provide useful information—2 out of 19).

Garbero, Marion, & Brailovskaya (2016) take a further step. In reviewing the impact evaluation literature on improved varieties, the authors first score the studies for risk of bias and then regress the effect sizes from the studies on these bias scores. Their overall result from a meta-analysis of results from 20 relevant studies assessing outcomes related to poverty, income, or expenditure show statistically significant impacts on the order of 6 to 32 percent relative to comparison farmers. When these effect sizes are regressed on the risk of bias scores for the studies, those examining poverty outcomes show a positive correlation, suggesting that biased impact assessment design for poverty studies could be inflating the effect size. The relationship for income- or expenditure-focused studies is ambiguous.

There is much work to be done on the specific approaches to statistical meta-analyses that are ap-

propriate for those interventions for which a critical mass of rigorous studies have been carried out. For example, Meager (2017) examines a controversy in meta-analysis, looking at studies of the impact of vitamin A supplementation on child mortality – a group of RCTs characterized by high levels of heterogeneity in context and reported effect sizes. Meager shows that the specific way in which meta-analysis is carried out matters a lot in determining the major messages from such a mixed evidence base. She compares fixed effects models, in which it is assumed that there is a single, common effect size across all studies (and thus all differences in estimated treatment effect sizes can be attributable to sample or other error in the original studies), with a random effects model (in which the true effect size might differ from study to study). As the rate of Vitamin A deficiency in the population is likely to have major impact on the observed effect of supplementation there are strong theoretical grounds for thinking that treatment effects are indeed likely to be heterogeneous. By allowing the true effect to be different across studies, Meager’s framework and result allow to demonstrate the importance of using methods which can distinguish estimate precision from estimate generalizability when the literature contains heterogeneous treatment effects.

The revolution in standards of evidence for causal identification has transformed the toolkit of applied economists, but it has also brought some second-order methodological problems to the fore that were less important in the past. The issue of standards of evidence when comparing across methodologies remains a challenge, but the risk-of-bias scoring carried out by Garbero, Marion, & Brailovskaya (2016) and Stewart et al. (2015), and Meager’s sophisticated examination of the correct statistical specification for meta-analysis, show us productive ways forward. Certainly, while the internal validity of the existing evidence base was still in question, it made little sense to worry too much about external validity. This situation is now changing and is the subject of the next section.

4 UNDERSTANDING CGIAR IMPACTS ON A LARGE SCALE

4.1 EXTERNAL VALIDITY, EVIDENTIAL STANDARDS, AND THE LONG-RUN EFFECT ON AID ALLOCATION

Nancy Cartwright and Jeremy Hardie's book summarizes a literature from the philosophy of science on the process of going from studies telling you that "it worked somewhere" to the conclusion that policy-makers and aid donors really want, which is "it will work here" (Cartwright & Hardie, 2012). As they show, the road from ex post impact evaluation results of a specific intervention in a specific place to the ex ante planning process for the same intervention in a different place is fraught (and paved with good intentions).

Pritchett & Sandefur (2013) characterize four features of the challenge of balancing evidential standards:

"(i) evidence rankings that ignore external validity, (ii) meta-analysis of the average effect of a vaguely-specified "intervention" which likely varies enormously across contexts, (iii) clustering evaluation resources in a few expensive studies in locations chosen for researchers' convenience, and (iv) the irresistible urge to formulate global policy recommendations" (Pritchett & Sandefur, 2013, 164). These authors are not arguing against randomization as a methodological tool for impact assessment but rather in favor of attention to context and heterogeneity via "*orders of magnitude more use of randomization, but with far fewer grand claims to external validity*" (ibid.). Any single empirical result should have far less weight, and there should be many more of them to draw from. This is particularly important for agriculture, where the contextual factors that can condition heterogeneity of outcomes are so varied and impossible to fully control for: soils, climate, water, social institutions, government policy, markets, transaction costs, etc. As Pritchett & Sandefur (2013) show, it can sometimes be preferable to take non-experimental evidence from the right context than experimental evidence from another context.

Failing to heed the warnings of Pritchett & Sandefur (2013) as well as those of Barrett & Carter (2010) and Deaton & Cartwright (2017)—by, for example, making impact assessment synonymous with randomized control trials alone—could significantly skew the focus of the agricultural research and development aid portfolio away from areas that are difficult or impossible to evaluate using randomized designs. This would be an unnecessary and unfortunate outcome and would reflect a methodological fundamentalism that is at odds with the pragmatic tradition in impact assessment in CGIAR. However, as the reviews by Garbero, Marion, & Brailovskaya (2016), Herdt & Mine (2017), Loevinsohn et al. (2013), and Stewart et al. (2015) showed, the average level of quality in the published

impact assessment literature is low, with many biased studies. This is the case even for technologies such as improved varieties, which are highly suited to being studied using randomized control trials. Under the SIAC program, SPIA has launched a campaign to increase the rigor of impact assessment central to our work program. Looking ahead, there is hence a need to simultaneously push ahead on improving the credibility of individual studies on the one hand and on broadening coverage across the CGIAR research portfolio and achieving cost-effectiveness on the other.

4.2 RIGOROUSLY MEASURING OUTCOMES AT SCALE

Empirical studies establishing causal evidence on farmers' behavioral responses to the availability of new technologies are a key part of establishing credible evidence of CGIAR impacts. However, the low take-up of new technologies that is often observed in real-life settings contains equally important information about the potential profitability of new technologies that is all too often ignored. If farmers decide not to adopt, they likely have good reasons not to do so. Rigorously documenting adoption rates for large representative populations is hence complementary to studies identifying causal relationships. While many CGIAR centers conduct "adoption studies," they often involve small, non-representative samples, and short time-frames (Doss, 2006). They are consequently of limited value. The question should not be whether some selected farmers adopted a particular technology once, but rather whether a large share of farmers representative of a population targeted by the technology decide to adopt and continue to use a new technology in the seasons after the initial adoption. This suggests the need to move from many small-scale, one-shot surveys to fewer, well-designed, and representative longitudinal surveys. Crucially, these surveys need to be institutionalized so that they have a life of their own, independent of short-term donor assistance. Such a vision is best achieved through partnerships with institutions that have a comparative advantage in surveys in countries of highest priority to CGIAR—

such as the World Bank and FAO. These partnerships will need to be oriented toward meeting the needs of countries monitoring their progress on the SDGs.

4.3 MICRO-MESO-MACRO: INTERACTIONS ACROSS SCALES

Some impact pathways are complex—particularly those mediated by markets and over borders—suggesting the need for models at a macro-level. The micro-foundations of macro models—particularly off-the-shelf models—need close and sustained scrutiny. Detailed microeconomic analyses are required to help answer questions related to, for example, the modeling of labor demand in processes of technological change.

Most of the macro models that are used for agricultural impact assessment are based on some combination of partial equilibrium analysis and general equilibrium modeling. In all cases, one "primitive" of the model is typically an initial level and/or a growth rate in total factor productivity (TFP). Impact assessment at the macro scale is then carried out by seeing how the model responds to changes in the TFP level or growth rate, such as might be generated by a new technology or innovation. In this sense, macro models may be highly complementary to detailed micro analysis of productivity growth. A careful micro estimate of TFP increases in a particular crop could be inserted into a macro model, which could then be used to generate estimates of the economy-wide impacts of the research impact. A challenge, however, is that it can be quite difficult even with detailed micro analysis to separate the TFP impact of research from the impacts of other kinds of TFP shifters (such as improvements in institutions or even weather-related factors). The models also require estimates of average TFP changes over broad geographies, rather than location-specific estimates that emerge from narrowly focused studies at the micro level. Thus, a challenge remains in finding appropriate micro evidence to feed into the macro models.

A further problem is that all the macro models inevitably build in strong assumptions about functional

forms as well as model closure assumptions. These assumptions are difficult to test or to assess through standard sensitivity analysis, but they can often matter a great deal for the outcomes of interest. For instance, models must make assumptions about production functions, such as the elasticities of substitution between capital, land, and labour; or about the functional forms used to produce output from intermediate inputs. These assumptions are not innocuous, in the sense that they can have quantitatively significant effects on outcomes of interest.

For instance, most models make assumptions about how consumers perceive domestic goods in relation to imported substitutes. This is particularly important in models that allow for agricultural trade. Is domestic rice a perfect substitute for imported rice? If so, then consumers will dramatically switch between the two goods depending on which is less expensive. Most models instead assume a different relationship between imports and domestically produced goods, using some version of an “Armington aggregator” that converts domestic goods and imports into a single composite good that is consumed. But the specific form of the aggregator will matter: a Cobb-Douglas aggregator will imply that consumers will always devote a constant fraction of their expenditure on rice to imports and a constant fraction to domestic production. Alternatively, a Leontief aggregator will imply that, regardless of prices, consumers will always consume the two goods in specific proportions. These may seem like technical details, but they have quite different implications for a model’s predictions. In a similar vein, models will be sensitive to assumptions that are built into a model about the substitutability of different crops and commodities for different categories of use (e.g., consumption, processing, animal feed).

How should we understand rigor in the context of models? The generally accepted best practice is to validate models by testing them against data other than those that were used to calibrate them. Since these models are typically calibrated such that they duplicate historical data, it is not always obvious how they can be validated in this way. Practitioners

sometimes object that there is no feasible way to run counterfactuals or to test the sensitivity of the model structure. They will normally offer some alternative scenarios for certain key parameters (e.g., low, medium, or high population growth; or two scenarios for productivity growth). But the deeper structures of the models are very seldom tested.

One feasible way to validate the model and to test these deeper structures is to engage in a kind of historical forecasting. One might, for instance, take a model like that used by Laborde et al (2017) and instead of calibrating it to the data from 2000-2015, calibrate it instead to data from 1985-2000 and see how well the model then predicts the period from 2000 onward. In other words, the point would be to show how well the model predicts out of the sample to which it is calibrated. One could equally take the model, as calibrated to data from 2000-2015 and feed it with base year data from 1985 to see how well it matches observations from 1985-2000. Any of these exercises would allow for some (qualitative) evaluation of the model against data other than those to which it was originally calibrated. If the model performs well out of sample, in this fashion, then we can trust it more for forecasting.

Another (more limited) way to test the model is to calibrate to one set of variables and to see how well the model then matches the data on a different set of variables. For instance, the calibration could involve feeding in data on agricultural inputs and output, with the validation based on seeing how well the calibrated model performs in matching variables such as service-sector productivity or non-agricultural employment. This approach is less satisfactory, in that there are often underlying arithmetic or algebraic links that imply certain relationships will hold among the variables in the model, so that the two sets of variables are not in fact independent. But to the extent that a calibration to one set of variables can generate a good fit for other variables, and to the extent that these other variables can be claimed to be plausibly unrelated, this may be an acceptable way of validating the model.

5

CONCLUSION AND SUMMARY

The CGIAR system represents a body of agricultural research and development activity of close to US\$1 billion annually. Cataloguing and tracking the outputs from this system, and determining whether and how these outputs lead to development outcomes, is an enterprise worthy of significant investment. The individual CGIAR research centers have a high degree of autonomy and are incentivized to advocate for their own effectiveness. In the absence of strong and consistent demand for rigor from donors, one can hypothesize that those centers with the most able communications teams will be those that are best able to capture a larger share of the total funding to the system.

Pressure is building for a change in culture toward more rigorous evidence, with evidential standards from other sectors influencing donor attitudes toward CGIAR. For example, the Bill & Melinda Gates Foundation published its inaugural *Goalkeepers* report on progress toward the SDGs in 2017, with a commitment to continue doing so annually until 2030. The report picks 18 of the 232 SDG indicators for monitoring on a global scale. Time series for indicators such as stunting, global HIV deaths, and maternal deaths show progress from 1990 to the present day and the possible trajectories to 2030. The aim of the report is to ensure that momentum for continued progress is not lost. However, data collection for agriculture (as well as gender and education) is shown to be inadequate and the page is filled with an empty chart stating “Insufficient data”. The message from the Gates Foundation is clear—get your act together.

For the many reasons outlined in this report, the situation is a little more complicated than that, but certainly the rigor revolution demands that we do better, in the following ways. First, we need to institutionalize detailed data collection related to CGIAR activities along the results chain from investments to outputs to outcomes. For this effort to be practical, we need to focus on a few key locations as a first step toward catching up after years of neglect. This would allow CGIAR to reconnect with its historical track record of collecting longitudinal data, best illustrated perhaps by the large body of literature resulting from the longitudinal ICRISAT villages datasets. In a new generation of longitudinal studies, we need carefully implemented geo-located surveys featuring DNA fingerprinting of the major crops and livestock, reliable data on farmers’ management practices, and detailed socioeconomic data, combined with information on the policy and institutional environment. Data quality should be of the highest priority. If we do this right, we can calibrate and take full advantage of the vast data output from the latest wave of remote sensors to interpolate certain indicators between survey waves, and possibly make out-of-sample predictions for other geographic areas.

Second, impact evaluation and efficacy studies need to focus on causal relationships for which we have the greatest uncertainty and for which information would have the highest value. This suggests a greater focus on theory—away from searching for “what works” in the abstract and toward finding out why certain things work and others do not in particular contexts. Farmers’ behavioral responses should be factored in as an important component of management, and accurately measuring different technologies through best-practice methods should be a priority. The integration across data types offers tremendous potential for new insights. It is less obvious how to make methodological breakthroughs on tracing policy influence or meas-

uring the outcomes from capacity-building efforts, though the principle of independent theory-based evaluation should be prominent.

Given the wide range of activities carried out by CGIAR, it is clear that a broad toolkit of approaches will be needed to assess impacts. This makes standardization and simple messages hard to come by, but SPIA is committed to its role as convener and intermediary between the CGIAR research community, external researchers, and the donors that fund the system. In doing so, we hope to ensure that we can raise the ratio of signal to noise and help incentivize greater clarity, realism, and rigor in the thinking about impacts from investments in CGIAR.

6 REFERENCES

- Abay, K. A., Abate, G. T., Barrett, C. B. & Bernard, T. (2018) Correlated non-classical measurement errors, ‘second best’ policy inference and the inverse size-productivity relationship in agriculture. IFPRI Discussion Series No. 1710, February 2018.
- Alston, J. M., Norton, G., & Pardey, P. G. (1995). *Science under scarcity: Principles and practice for agricultural research evaluation and priority setting*. Ithaca, NY: Cornell University Press (republished in 1998 by CAB International, Wallingford, UK).
- Angrist, J., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.
- Ashour, M., Billings, L., Gilligan, D., Hoel, J. B., & Karachiwalla, N. (2016). *Do beliefs about agricultural inputs counterfeiting correspond with actual rates of counterfeiting? Evidence from Uganda*. IFPRI Discussion Series No. 1552. Washington, DC: International Food Policy Research Institute. <http://ebrary.ifpri.org/cdm/singleitem/collection/p15738coll2/id/130598>
- Baird, Sarah, Aislinn Bohren, Craig McIntosh, Berk Özler. 2012. “Designing Experiments to Measure Spillover and Threshold Effects.” Policy Research Working Paper Series 6824, The World Bank
- Barrett, C. B., Agrawal, A., Coomes, O. T. & Plateau, J.-P. (2009) Stripe review of social sciences in the CGIAR. Rome: CGIAR Science Council.
- Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economics Perspectives and Policy*, 32, 515–548.
- Barrett, C. B., & Upton, J. B. (2013) “Food Security and Sociopolitical Stability in Sub-Saharan Africa”, Chapter 13 in Barrett, C. B. (editor) *Food Security and Sociopolitical Stability*. Oxford: Oxford University Press.
- Beaman, L. & Dillon, A. (forthcoming) Diffusion of Agricultural Information within Social Networks: Evidence on Gender Inequalities from Mali. *Journal of Development Economics*.
- Beaman, L., BenYishay, A., Magruder J. & Mobarak, M. (2015). “Can Network Theory-based Targeting Increase Technology Adoption?”, mimeo, Northwestern University.
- Beegle, K., Carletto, C., & Himelein, K. (2012). Reliability of recall in agricultural data. *Journal of Development Economics*, 98(1), 34–41. <https://doi.org/10.1016/j.jdeveco.2011.09.005>
- BenYishay, A. & Mobarak, M. (2018) Social Learning and Incentives for Experimentation and Communication.” The Review of Economic Studies. <https://doi.org/10.1093/restud/rdy039>
- Bill & Melinda Gates Foundation. (2017). *Goalkeepers: The stories behind the data 2017*. <http://www.globalgoals.org/goalkeepers/datreport/>
- Bold, T., Kaizzi, K. C., Svensson, J., & Yanagizawa-Drott, D. (2017). Lemon technologies and adoption: Measurement, theory and evidence from agricultural markets in Uganda. *Quarterly Journal of*

- Economics*, 132(3), 1055–1100.
- Bulte, E., Beekman, G., Di Falco, S., Hella, J., & Lei, P. (2014). Behavioral responses and the impact of new agricultural technologies: Evidence from a double-blind field experiment in Tanzania. *American Journal of Agricultural Economics*, 96(3), 813–830. <https://doi.org/10.1093/ajae/aau015>
- Cameron, D. B., Mishra, A., & Brown, A. N. (2016). The growth of impact evaluation for international development: How much have we learned? *Journal of Development Effectiveness*, 8(1), 1–21. <https://doi.org/10.1080/19439342.2015.1034156>
- Carletto, C., Gourlay, S., Murray, S., & Zezza, A. (2016). *Land area measurement in household surveys: Empirical evidence and practical guidance for effective data collection*. Living Standards Measurement Study (LSMS) Guidebook. Washington, DC: World Bank.
- Cartwright, N., & Hardie, J. (2012). *Evidence-base policy: A practice guide to doing it better*. Oxford: Oxford University Press.
- CGIAR. (2015). CGIAR strategy and results framework 2016–2030. Montpellier, France. <https://cgspage.cgiar.org/bitstream/handle/10947/3865/CGIAR%20Strategy%20and%20Results%20Framework.pdf>
- CGIAR Independent Science and Partnership Council (2012) *A Stripe Review of Natural Resources Management Research in the CGIAR*. Rome, Italy: CGIAR Independent Science and Partnership Council Secretariat.
- Cook, T.D. (2018) *Social Science & Medicine*, <https://doi.org/10.1016/j.socscimed.2018.04.031>
- Dalrymple, D. (1978). *The development and spread of the high-yielding varieties of wheat and rice among less-developed nations*. Foreign Agricultural Economic Report No. 95, 6th edition. Washington, DC: US Agency for International Development.
- Darwin, C. (1868). *The variation of animals and plants under domestication*, Volume 1. London: John Murray.
- Deaton, A. & Cartwright, N. (2017) *Understanding and misunderstanding randomized controlled trials*. National Bureau of Economic Research Working Paper 22595 <http://www.nber.org/papers/w22595>
- de Janvry, A., Dustan, A., & Sadoulet, E. (2011). *Recent advances in impact analysis methods for ex-post impact assessments of agricultural technology: Options for the CGIAR*. Report prepared for the workshop “Increasing the rigor of ex-post impact assessment of agricultural research: A discussion on estimating treatment effects,” organized by the CGIAR Standing Panel on Impact Assessment (SPIA), October 2, 2010, Berkeley, CA, USA. Rome: Independent Science and Partnership Council Secretariat.
- de Janvry, A., Sadoulet, E., & Suri, T. (2017). Field experiments in developing country agriculture. In *Handbook of economic field experiments*, edited by Abhijit Vinayak Banerjee and Esther Duflo, volume 2, 427–466. Amsterdam: North-Holland. <https://doi.org/10.1016/bs.hefe.2016.08.002>
- Doss, C. R. (2006). Analyzing technology adoption using microstudies: Limitations, challenges, and opportunities for improvement. *Agricultural Economics*, 34(3), 207–219. <https://doi.org/10.1111/j.1574-0864.2006.00119.x>
- Emerick, K., de Janvry, A., Sadoulet, E. & Dar, M. (2016). Optimizing social learning about agricultural technology: Experiments in India and Bangladesh. FERDI Policy Brief 158 (September 2016). Clermont-Ferrand: Fondation pour les Etudes et Recherches sur le Développement International.
- Emerick, K. & Dar, M. (2017) Enhancing the diffusion of information about agricultural technology. Mimeo. Agricultural Technology Adoption Initiative.
- Erenstein, O., & Laxmi, V. (2008). Zero tillage impacts in India’s rice–wheat systems: A review. *Soil*

- and Tillage Research, 100(1–2), 1–14. <https://doi.org/10.1016/j.still.2008.05.001>
- Fairbairn, A., Michelson, H., Ellison, B., & Manyong, V. (2016). *Mineral fertilizer quality: Implications for markets and small farmers*. Selected paper prepared for presentation for the 2016 Agricultural and Applied Economics Association, Boston, MA, July 31 – August 2.
- Foure, J., Benassy-Que, A., & Fontagne, L. (2012). *The great shift: Macroeconomic projections for the world economy at the 2050 horizon*. CEPII Working Paper No. 2012-3. Paris: Centre d'Etudes Prospectives et d'Informations Internationales.
- Garbero, A., Marion, P., & Brailovskaya, V. (2016). *Meta-analysis: The impact of agricultural research on poverty*. Working paper. Rome: International Fund for Agricultural Development.
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact evaluation in practice*. Washington, DC: World Bank.
- Glennerster, R. & Suri, T. (2015) *Measuring the Effects of NERICA, Short Duration Rice, on Harvest Prices*. ATAI Project, MIT.
- Gourlay, S., Kilic, T., & Lobell, D. (2017). *Could the debate be over? Errors in farmer-reported production and their implications for the inverse scale-productivity relationship in Uganda*. Policy Research Working Paper 8192. Washington, DC: World Bank.
- Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4), 501–522.
- Griliches, Z. (1958). Research costs and social returns: Hybrid corn and related innovations. *Journal of Political Economy*, 66(5), 419–431.
- Herdt, R. W., & Mine, S. (2017). *Does modern technology increase agricultural productivity? Revisiting the evidence from Loevinsohn et al.* White paper. Rome: CGIAR Independent Science and Partnership Council.
- Hurley, T. M., Pardey, P. G., Rao, X., & Andrade, R. S. (2016). *Returns to food and agricultural R&D investments worldwide, 1958–2015*. InStEP Brief. St Paul, MN: International Science & Technology Policy & Practice center, August 2016.
- Imbens, G. (2018) Comments on understanding and misunderstanding randomized controlled trials: A commentary on Cartwright and Deaton. *Social Science & Medicine*. <https://doi.org/10.1016/j.socscimed.2018.04.028>
- Ioannidis, J. P. A (2018) Randomized controlled trials: Often flawed, mostly useless, clearly indispensable: A commentary on Deaton and Cartwright. *Social Science & Medicine*. <https://doi.org/10.1016/j.socscimed.2018.04.029>
- Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Ithaca, NY: Cornell University Press.
- Kilic, T., & Sohnesen, T. (2015). *Same question but different answer: Experimental evidence on questionnaire design's impact on poverty measured by proxies*. Policy Research Working Paper 7182. Washington, DC: World Bank. <https://ideas.repec.org/p/wbk/wbrwps/7182.html>
- Kosmowski, F., Aragaw, A., Kilian, A., Ambel, A., Ilukor, J., Yigezu, B., & Stevenson, J. (2018). Varietal identification in household surveys: results from three household-based methods against the benchmark of DNA fingerprinting in southern Ethiopia. *Experimental Agriculture*, 1-15. <https://doi.org/10.1017/S0014479718000030>
- Laajaj, R., & Macours, K. (2016). Learning-by-doing and learning-from-others: Evidence from agronomical trials in Kenya. Policy brief prepared for the workshop “Learning for Adopting,” June 1–2, 2016, Clermont-Ferrand, France.
- Laajaj, R., Macours, K., Masso, C., Thuita, M. & Vanlauwe, B. (2018). “*Yield gap or field gap? Reevaluating the yield gap bringing together agronomic and economic insights*”, mimeo, Paris School of Economics

- Lobell, D. B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143, 56–64. <https://doi.org/10.1016/j.fcr.2012.08.008>
- Loevinsohn, M., Sumberg, J., Diagne, A., & Whitfield, S. (2013). *Under what circumstances and conditions does adoption of technology result in increased agricultural productivity?* Brighton, UK: Institute of Development Studies.
- Maredia, M., Reyes, B., Manu-Aduening, J., Dankyi, A., Hamazakaza, P., Muimui, K., Rabbi, I., Kulakow, P., Parkes, E., Abdoulaye, T., Katungi, E. and Raatz, B. (2016) *Testing alternative methods of varietal identification using DNA fingerprinting: Results of pilot studies in Ghana and Zambia*. MSU International Development Working Paper No. 149 October 2016. East Lansing, MI: Michigan State University.
- Meager, R. (2017) Vitamin A Supplements Often Substantially Reduce Child Mortality, But Their Impact Is Heterogeneous: Resolving A Controversy In Meta-Analysis. Mimeo, London School of Economics.
- National Human Genome Research Institute. (2016). The cost of sequencing a human genome. <https://www.genome.gov/sequencingcosts/>
- Oakes, J. M. (2018) The tribulations of trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*. <https://doi.org/10.1016/j.socscimed.2018.04.026>
- OECD (Organisation for Economic Co-operation and Development) Development Assistance Committee. (2005). *The Paris Declaration on Aid Effectiveness*. Paris. <http://dx.doi.org/10.1787/9789264098084-en>
- OECD (Organisation for Economic Co-operation and Development) Development Assistance Committee. (2008). *The Accra Agenda for Action*. Paris. <http://dx.doi.org/10.1787/9789264098107-en>
- Pearl, J. & Mackenzie, D. (2018) *The book of why: The new science of cause and effect*. New York: Basic Books.
- Pritchett, L. (2002). It pays to be ignorant: A simple political economy of rigorous program evaluation. *Journal of Policy Reform*, 5, 251–269.
- Pritchett, L., & Sandefur, J. (2013). Context matters for size: Why external validity claims and development practice don't mix. *Journal of Global Development*, 4(2), 161–197. <https://doi.org/10.2139/ssrn.2364580>
- Raitzer, D. A., & Kelley, T. G. (2008). Assessing the contribution of impact assessment to donor decisions for international agricultural research. *Research Evaluation*, 17 (September), 187–199. <https://doi.org/10.3152/095820208X331702>
- Rao, X., Hurley, T. M., & Pardey, P. G. (2016). *Recalibrating the reported returns to agricultural R&D: What if we all heeded Griliches?* Paper prepared for the annual meeting of the Agricultural and Applied Economics Association, July 30–August 1, Chicago.
- Renkow, M., & Byerlee, D. (2010). The impacts of CGIAR research: A review of recent evidence. *Food Policy*, 35(5), 391–402. <https://doi.org/10.1016/j.foodpol.2010.04.006>
- Rogers, P. (2009). Matching impact evaluation design to the nature of the intervention and the purpose of the evaluation. *Journal of Development Effectiveness*, 1, 217–226.
- Savedoff, W. D., Levine, R. & Birdsall, N. (2006). *When will we ever learn? Improving lives through impact evaluation*. Washington, DC: Center for Global Development.
- Steel, D. P. (2007) Across the boundaries: Extrapolation in biology and social science. Oxford: Oxford University Press. xi+241 pages.
- Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations*. Working Paper No. 38. London: UK Department for International Development.

- Stevenson, J. & Vlek, P. (forthcoming) Assessing the adoption and diffusion of sustainable agricultural practices: Synthesis of a new set of empirical studies. CGIAR Standing Panel on Impact Assessment synthesis report.
- Stewart, R., Langer, L., Rebelo, N., Silva, D., Muchiri, E., Zaranyika, H., et al. (2015). *The effects of training, innovation and new technology on African smallholder farmers' economic outcomes and food security: A systematic review*. Campbell Systematic Reviews 2015: 16. Oslo: Campbell Collaboration. <https://doi.org/10.4073/csr.2015.16>
- Tjernstrom, E. (2017) *Learning for Adopting - Signals, Similarity and Seeds: Social Learning in the Presence of Imperfect Information and Heterogeneity*. In Technology Adoption in Developing Country Agriculture. Policy Briefs from the Workshop Organized by FERDI and SPIA, June 1st and 2nd 2016, FERDI, Clermont-Ferrand.
- Waddington, H., Snistveit, B., Hombrados, J., Vojtкова, M., Phillips, D., & Davies, P. (2014). Farmer field schools for improving farming practices and farmer outcomes in low- and middle-income countries. Campbell Systematic Reviews 2014: 6. Oslo: Campbell Collaboration. <https://doi.org/10.4073/CSR.2014.6>
- Walker, T. S., & Crissman, C. C. (1996). *Case studies of the economic impact of CIP related technologies*. Lima, Peru: International Potato Center (CIP).
- Walker, T. S., Maredia, M. K., Kelley, T. G., Rovere, R. La, Templeton, D., Thiele, G., & Douthwaite, B. (2008). *Strategic guidance for ex post impact assessment of agricultural research*. Report prepared for the Standing Panel on Impact Assessment, CGIAR Science Council. Rome: Science Council Secretariat.
- White, H. (2009). Theory-based impact evaluation: Principles and practice. *Journal of Development Effectiveness*, 1(3), 271–284. <https://doi.org/10.1080/19439340903114628>
- Wossen, T., Girma, G., Abdoulaye, T., Rabbi, I., Olanrewaju, A., Alene, A., et al. (2017). *The cassava monitoring survey in Nigeria*. Ibadan, Nigeria: International Institute of Tropical Agriculture.
- Zomer, R. J., Neufeldt, H., Xu, J., Ahrends, A., Bossio, D., Trabucco, A., et al. (2016). Global tree cover and biomass carbon on agricultural land: The contribution of agroforestry to global and national carbon budgets. *Scientific Reports*, 6, Article 29987. <https://doi.org/10.1038/srep29987>



Independent
Science and
Partnership
Council

CGIAR Independent Science & Partnership Council (ISPC) Secretariat
c/o FAO, Viale delle Terme di Caracalla 00153 Rome, Italy
t: +39 06 570 52103
<http://ispc.cgiar.org>