ANNUAL REVIEWS

Annual Review of Resource Economics

The Rigor Revolution: New Standards of Evidence for Impact Assessment of International Agricultural Research

James R. Stevenson,[1,2] Karen Macours,[3,4] and Douglas Gollin[5]

[1]CGIAR Standing Panel on Impact Assessment, Alliance of Bioversity International and CIAT, Rome, Italy

[2]International Food Policy Research Institute, Washington, DC, USA

[3]Paris School of Economics, Paris, France; email: karen.macours@psemail.eu

[4]Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), Paris, France

[5]Department of Economics, University of Oxford, Oxford, United Kingdom

Annu. Rev. Resour. Econ. 2023. 15:495–515

First published as a Review in Advance on June 28, 2023

The *Annual Review of Resource Economics* is online at resource.annualreviews.org

https://doi.org/10.1146/annurev-resource-101722-082519

JEL codes: O12, O13, O32, O33, Q16

Keywords

agricultural research for development, impact assessment, measurement, causality, scale

Abstract

We take stock of the major changes in methodology for studying the impacts of international agricultural research, focusing on the period 2006–2020. Impact assessment of agricultural research has a long and recognized tradition. Until the mid-2000s, such assessments were dominated by a 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. We describe the implications of heightened standards of evidence for good practice in three domains of research design: causal inference, valid measurement, and statistical representativeness. We document advances in each of these domains and review recent evidence that demonstrates the lessons that can be learned from adopting these practices, emphasizing the importance of evidence at-scale, the need to consider portfolios of

innovations at a national level, and the challenges of accounting for innovations that are promoted as bundles.

1. INTRODUCTION

1.1. Does It Pay to Invest in Agricultural Research?

Impact assessment of investments in agricultural research has a long and proud tradition, aimed largely at providing answers to the question of whether research offers a good return on investment relative to other possible uses. Until the mid-2000s, impact assessment studies typically relied on models of demand for and supply of agricultural products in partial equilibrium. The basic framework for this approach was sketched out more than half a century ago (Griliches 1957, 1958). Griliches had observed the rate of adoption of hybrid maize varieties in different US states and created a simple model for linking the benefits from higher maize yields back to investments in research.

The appeal of such an approach lies in its simplicity. The first task in implementing the model is an adoption study to establish whether specific innovations originating from research activities have been adopted at a large scale. The impact of a widely adopted innovation on aggregate agricultural productivity is then modeled as a technology-induced change in the marginal cost curve for the commodity in question: Griliches' *k*-shift (Fulginiti 2010). The magnitude of this shift is calculated using yield advantage data from agronomic trials, econometric analyses, or other sources. All else being equal, the economic surplus generated is assumed to be shared between producers and consumers. Surpluses can be projected forward or backward in time to represent the dynamic flow of benefits obtained from research-derived productivity gains. This setup has the tremendous advantage that the benefits can be directly compared to the total funding used in the research in a cost-benefit analysis. Estimates of the economic returns to research using such an approach have been produced periodically (Alston et al. 1995, Raitzer & Kelley 2008), and there remains some demand for these aggregate numbers.

However, the methods used to generate such internal rate of return (IRR) calculations are heavily dependent on strong assumptions, sometimes leading to numbers that are implausibly large (Hurley et al. 2016, Rao et al. 2020). Hurley et al. (2016) note that the calculation of IRRs is clearly inconsistent with the realities of the benefit and cost streams associated with agricultural research. Rao et al. (2020) propose a modified internal rate of return; building on aggregate evidence from 2,829 evaluations in the database of the International Science and Technology Practice and Policy (InSTePP, v3.0) program, Rao et al. (2020) calculate the IRR for agricultural research to be 14.3%. This is still high, suggesting that aid investments in agricultural research pay off handsomely but at a more realistic order of magnitude than earlier estimates.

A slight modification of the standard approach has been to calculate the returns to research by measuring changes in total factor productivity (TFP) in agriculture over some geography and time period, and then to use regression analysis to tease out the contributions from research investments, as distinct from other sources of variation across time and space. Macro versions of this approach are modeled on the work of Fuglie and various coauthors (summarized by Fuglie 2018). Increasingly, however, researchers have sought to measure TFP change at the micro level, linked to specific technologies or innovations (see Pingali 2012 for a set of high-quality examples). However, the nature of agricultural production—subject as it is to risk and characterized by enormous heterogeneity—makes TFP calculations tricky at best, as documented by Gollin & Udry (2021). Not least among the challenges is the measurement of all relevant inputs and outputs.

More fundamentally, the aggregate rate of return generated by higher agricultural output is at best an incomplete measure of welfare gains. Indeed, output growth may be an entirely unhelpful metric in the era of the United Nations (UN) Sustainable Development Goals (SDGs). International agricultural research now targets a wide range of objectives, including climate adaptation, environmental health, poverty reduction, and resilience, among others. Despite these methodological and conceptual concerns, the institutional history of regularly publishing studies on the impact of investments arguably made the CGIAR of the late 1990s a leader in estimating development aid effectiveness. Many other institutions operating in the field of international development were investing less in impact assessment, less systematically, and less often. The introduction of the UN Millennium Development Goals 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 agricultural research to impact and increasingly expect to see evidence of impact on development outcomes that address specific societal concerns: cutting poverty, reducing food insecurity, improving nutrition, and ensuring environmental sustainability. 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, though it would require an even stronger set of assumptions than the earlier IRR calculations. Hence, in practice, impact assessment must adapt to using evidence from a broader range of methods to stay relevant. Such a shift requires a conceptual move away from a narrow focus on land productivity, accompanied by a series of methodological shifts.

Although agricultural productivity conceptually should measure a ratio of output to all inputs, agricultural productivity has often been treated as synonymous with agricultural yields (kg/ha). Yields may be the relevant metric to maximize when agricultural land area is limiting, but in many contexts, development programs care more about increasing farmers' income by increasing their labor productivity ($US/hour worked), rather than their land productivity. In contexts where people living in poverty are primarily found in rural areas and are working in agriculture, labor productivity growth probably maps more directly into poverty reduction than does land productivity growth (Macours 2019). If the objective is for agricultural research to help reduce global poverty, we should be troubled by evidence showing 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 (Barrett & Upton 2013). Identifying the kinds of research investments that will increase labor productivity may be difficult ex ante, which is why it is important to obtain rigorous causal evidence on which innovations successfully increase labor productivity. Taking the task seriously is likely to be critical to any attempt to achieve poverty reduction through international agricultural research (Gollin et al. 2018). 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. All of this argues for an important role for impact assessment results to feed back into priority setting.

1.2. Does It Pay to Be Ignorant?

In a 2002 paper, Lant Pritchett (2002) developed a model addressing what he saw as the chronic underinvestment in rigorous evaluation in international development, exploring the interplay between advocates (program directors) and those providing resources (voting public). Advocates must secure resources for their programs—they are the entrepreneurs of the development industry. Advocates believe that they know the true effectiveness of their programs and that a rigorous evaluation will reveal this true effectiveness. Advocates can pursue one of two strategies to secure resources. The first is to subject their program to rigorous impact evaluation and offer the

evidence from these evaluations to donors. The alternative strategy is to 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. 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 that their programs are effective.

In 2006, the Center for Global Development published its landmark report "When Will We Ever Learn?" (Savedoff et al. 2006), which laid out the extent to which the lack of rigorous evidence in international development was a pervasive problem. It 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). In the period following the publication, 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 2005, 2008). The past 20 years have also seen several other institutional innovations that have helped program directors generate evidence and secure funds for evaluations from donors that expect rigor rather than pilot and persuade type communications. The Abdul Latif Jameel Poverty Action Lab (J-PAL) is a global network of academic researchers established in 2003 at MIT and has been running randomized controlled trials (RCTs) with a range of development agencies and providing training for thousands of development actors. 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 researchers in countries worldwide. Other academic initiatives soon followed [e.g., the Center for Effective Global Action (CEGA) at the University of California, Berkeley], and many donor agencies also increased their in-house capacity and requirements for rigorous impact evaluations. Today, research conducted in many institutions outside academia, including multilateral development banks (e.g., World Bank, InterAmerican Development Bank) and private sector organizations (e.g., IDinsight or Mathematica) also relies heavily on RCTs to evaluate the impact of development interventions. This also applies to IFPRI (International Food Policy Research Institute), one of the CGIAR centers, which has been strongly involved in RCT research since the early 2000s, notably in evaluating innovative social protection programs in Latin America (Skoufias 2005).

In addition, advisory bodies emerged, such as GiveWell, New Philanthropy Capital, and ImpactMatters, created to foster greater effectiveness in the multibillion-dollar nongovernmental organization industry and social impact investing firms. In recent years, this trend has tipped over to a global movement and online community for effective altruism. Effective altruism's goal is to ensure that a commitment to helping others is married to an equal commitment to ensuring that such help does indeed help. The philosophy of William MacAskill—effective altruism's figurehead and best-selling author (MacAskill 2015)—would simply not work had there not been an explosion in the kind of impact evidence that allows the virtuous to dedicate themselves to allocating their capital to effective altruistic interventions.

1.3. Defining the Rigor Revolution

The move from a Griliches-inspired ex post impact assessment toolkit toward a broader range of empirical methodology has taken its time to be realized in agricultural research. Concerns about

outdated methodology were raised by a review of social science practice in CGIAR many years ago (Barrett et al. 2009). Over the past 15 years, the international agricultural research system has been through a series of reforms with different articulations of its impact pathways. The status quo can now be characterized as an ever-broader range of outcome metrics being of interest to donors. The funders of development interventions, including research, have new and high expectations of the quality of evidence, particularly in relation to causality and measurement, reflecting the advances of the past two decades. While the empirical and conceptual challenges for rigorous impact assessment are large, the need to embrace the rigor revolution is real. In the absence of coherent and plausible strategies for tracking results and measuring impacts, there are concerns that research funders will become more risk averse and set their own idiosyncratic standards for indicators, formats, time lines, and priorities. Our hope is that this review points in a more productive direction.

We describe a rigor revolution that has taken place in evidential standards regarding the impact of agricultural research. We see this as an expansion of the frontier of good empirical practice in three dimensions:

- Measurement: toward accurate and valid measurement of treatment (i.e., research-derived agricultural technology use) and outcomes (e.g., productivity, poverty, nutrition, environmental health);
- Causality: toward a research design that allows for an unbiased causal relationship between treatment and outcomes; and
- Scale: toward estimates that are representative of a policy-relevant scale.

We review all three dimensions in the next sections, with a specific focus on evaluating innovations that originate from agricultural research. Two recent and highly relevant reviews complement ours. Dillon et al. (2020) tackle the relationship between measurement and causality, and Carletto et al. (2021) address strategies for reconciling measurement concerns with a desire to reach large scale.

2. MEASUREMENT MATTERS

2.1. Sowing Doubt

Genetic technologies, in the form of improved varieties of major food crops, lie at the heart of the history of agricultural research's contribution to international development. Reliable data on the adoption of improved varieties therefore have long been recognized as the cornerstone of any assessment of the impact of investments in plant breeding research (Walker & Crissman 1996, Walker et al. 2008). 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, allowing for reliable estimates of adoption rates through experts' or farmers' report. While the initial improved varieties for rice and wheat were easy to identify in the field, improved varieties from other crops were often less visibly distinct. Moreover, subsequent improvements on secondary target traits for breeding in cereals have 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. The diversification of breeding effort across crops and across traits hence poses a deep challenge to measuring the adoption of new varieties in farmers' fields, particularly in smallholder contexts characterized by complex, informal seed systems. Adoption data have always been scarce, and yet obtaining such data has in some ways become even harder over the past few decades.

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 two decades that the technology has beaten Moore's law.[1] Between 2007 and 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, and the use of genotyping can be considered a core part of the methodological toolkit for impact evaluation of genetic technologies in agriculture. Establishing proof of concept for the use of DNA fingerprinting in generating variety adoption estimates started with small-scale pilots beginning in 2012 (Maredia et al. 2016) and eventually reached nationally representative scale in Nigeria (Wossen et al. 2017) and Ethiopia (Jaleta et al. 2020, Kosmowski et al. 2020).

This work is part of a growing literature on measurement improvement in agriculture (Beegle et al. 2012; Carletto et al. 2016, 2021; Kilic & Sohnesen 2015; De Weerdt et al. 2020; Laajaj & Macours 2021). The literature employs measurement experiments, in which the same data are collected in multiple ways to test for consistency across methods. When one data collection method offers a clear benchmark for accuracy (as is the case for DNA fingerprinting), it becomes possible to test the extent and nature of the measurement error in other methods. Stevenson et al. (2023) collected data sets from author teams that had published validation studies of farmer-reported survey data on adoption of improved varieties, in which the DNA fingerprinting approach was used as the benchmark. On average, across a wide variety of crops and regions, farmers correctly identify a variety as improved only 71.2% of the time, with both underestimates and overestimates being very common. Moreover, with breeding targeting multiple traits (stress resistance, nutritional content, etc.) we need measurement methods that allow us to quantify adoption of specific varieties and traits, rather than simply identifying varieties as improved or traditional (as was common in earlier studies). The same data show that only 24.1% of farmer responses about the varietal name match the DNA fingerprinting results. Thus, even if farmers may have a functional knowledge of the variety and its traits (which itself is not a given), the preponderance of local names for varieties make it often impossible to get accurate varietal-specific data out of household surveys using survey questions alone.

The findings of these studies raise questions about the accuracy of the accumulated stock of knowledge about varietal diffusion to date. Although it has simply not been possible to carry out this kind of empirical check on our methods before, methodological advances now allow us to update the knowledge base. Hence, rather than being too pessimistic about what these results mean, we highlight several examples of how integrating DNA fingerprinting into impact studies can expand the kinds of questions we can hope to answer.

---

[1] 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. The DNA extracted from samples of plant tissue from farmers' fields is compared to the DNA extracted from reference samples representing (ideally) the universe of varieties that could be present in a specific country.

Wossen et al. (2017) compare farmer-reported crop variety data to varietal identification through DNA sequencing from cassava leaf material. The proportion of errors in the self-reported data is substantial, and the likelihood that farmers misreport varietal status is correlated with education and access to information. 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%). Abay et al. (2023) parse out the broader behavioral and inferential implications of misreporting (recorded mismeasured data are different from the respondents' true beliefs, which are accurate) versus misperceptions (recorded mismeasured data reflect the respondents' mistaken beliefs). Misperceptions regarding crop varietal identity can affect farmers' decision making on complementary inputs and hence also affect productivity through that channel. Using maize varietal data from Ethiopia (as reported by Kosmowski et al. 2020), Mallia (2022) leverages administrative data on the roll-out of a program of seed market reform to generate an instrument for correct classification. Comparing maize farmers that know they are cultivating an improved variety (true positives) to those that do not (false negatives) shows that the former use more fertilizers and hired labor at harvest than those who do not realize they are cultivating an improved variety.

Through the application of DNA fingerprinting,[2] we are learning a lot about the true nature of farmer management of planting material, including about the level of seed impurity and farmers mixing multiple varieties in single plots, as is often the reality in Sub-Saharan Africa. This challenges the concept of varietal adoption as a discrete binary decision that can be neatly analyzed in an econometric model and calls for further advances on measuring adoption and diffusion of new genetic technologies at scale.

2.2. New Approaches to Old Measurement Challenges

Research on natural resource management (NRM) innovations, including those that are designed to support adaptation to climate change, now represents a significant proportion of total investment in the CGIAR. However, there have been few efforts to track adoption of NRM technologies or practices at large scale (Barrett 2003, Erenstein & Laxmi 2008, Barrett et al. 2009, Stevenson & Vlek 2018, Stevenson et al. 2019). As a previous review has 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.

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,

[2]Different specific DNA fingerprinting assays can be deployed for addressing different research questions/contexts. See the review by Poets et al. (2020) for an overview. In general terms, at the time of publication, $US10 per sample for lab costs is a reasonable guide for budgeting (this includes DNA extraction from tissue, genotyping, and reporting), with some important variation on either side of that figure, depending on the crop, size of the job, etc. Arguably more of a hurdle are all the implications for fieldwork timing (i.e., the need to collect leaf or grain samples at a specific moment of maturity during the crop season), the need to engage a combination of skills across disciplines, meticulous sampling handling steps from fieldwork to lab, and the time and negotiation required in compiling reference material of varieties of the crop in question.

however, significant potential to harness the dynamic technological change taking place in data collection (e.g., remote sensing) in the service of assessing the technological change taking place in agricultural development.

Aker & Jack (2021) examine barriers to adoption of *demi-lune* water harvesting technologies (an NRM practice that is observable with remote sensing) in Niger through a large RCT. They find little evidence of liquidity or credit constraints limiting adoption. Training increases the share of adopters by over 90 percentage points, and adoption in turn raises output and results in spatial spillovers up to three years after treatment. Remotely sensed imagery will allow for further analysis of dynamic changes in adoption/disadoption over the long run. Very-high-quality reference data are needed to train remote sensing approaches for such an application, work that is fraught with potential pitfalls (as highlighted by Alix-Garcia & Millimet 2023), but with significant potential for positive results when done correctly. Studies are starting to use remote sensing to analyze the outcomes of changes in NRM practices. For example, Jayachandran et al. (2017) use remote sensing information on forest cover to document the impact of a payment for environmental services program in Uganda.

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. Anthropometric measurement provides an objective indicator that is standardized and comparable across settings, and other standardized biophysical markers (e.g., anemia tests) can also be appropriate (depending on the intervention studied) and scalable. By contrast, 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. There is ongoing work to identify better ways to measure certain social outcomes, such as women's empowerment (Doss et al. 2020, Quisumbing et al. 2021, Calvi et al. 2022, Jayachandran et al. 2023), in ways that can be systematically compared across contexts, and to reduce the cost of using those best-practice metrics that have been tested and confirmed.

3. CAUSALITY AND BIAS

3.1. Adoption Is a Choice

Establishing whether a farmer's outcomes improve because she adopted a new agricultural 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 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 adoption. Such selection on observables methods, such as propensity score matching, are hence particularly ill suited to shed light on the impacts of agricultural technologies. This was clearly argued by de Janvry et al. (2011), who highlight the need 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 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 to support the assumptions underlying the empirical estimates.

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 (Oakes 2018). However, randomization in and of itself does not guarantee orthogonality, but instead only buys balance between treatment and control in expectation. Deaton & Cartwright (2017) argue that the orthogonality assumption must be defended on a case-by-case basis, and most good RCT studies include checks for balance on observables. Focusing specifically on RCTs in agriculture in developing countries, Barrett & Carter (2010) also point to the importance of correctly characterizing the environmental and structural conditions to draw the appropriate inference from the often highly stylized experimental designs. Rosenzweig & Udry (2020) go further and point to methods to incorporate information on external shocks into impact assessment to increase external validity.

Violations of SUTVA (Stable Unit Treatment Value Assumption; Angrist et al. 1996) are the other obvious concern in RCTs that relate to agricultural technology, especially when they are not conducted at the right scale. As Imbens (2018) argues, such violations may sometimes simply need to be accounted for through relevant design adaptations, or in other contexts may actually be the main focus of the analysis. Randomizing over relatively large geographic units can allow researchers to specifically test for learning spillovers into the control group (Behaghel et al. 2020). Saturation designs that experimentally vary the density with which access to a new specific technology is offered can also lead to important insights into social learning and diffusion (Baird et al. 2018, Bernard et al. 2023). The chapter by de Janvry et al. (2017) discusses in more detail these and other considerations for applying RCTs in agriculture and provides specific guidelines for how to avoid common pitfalls and maximize on lessons learned.

The literature has shown that heeding these guidelines has enabled important advances in our understanding of information diffusion and farmers' learning. Farmers' learning about agricultural innovations from more than one source can be important, as shown using social network data and a field experiment in Malawi (Beaman et al. 2021). Demonstration through field days to observe and hear about experiences and outcomes with a new rice variety can also increase adoption (Dar et al. 2020, Emerick & Dar 2021). However, social learning can lead to slower adoption than direct exposure if networks are segregated or small (Beaman & Dillon 2018), and heterogeneous benefits can limit diffusion (Magnan et al. 2015). Trial farmers that more closely resemble those farmers that are targeted (BenYishay & Mobarak 2018) or private input suppliers with for-profit motives (Dar et al. 2020) may be more effective.

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 generally do not want to switch off such adjustments because impact evaluations ought to be designed to measure the different potential behavioral adjustments implicit in the process of adoption (de Janvry et al. 2017). Behavioral adjustments by active economic agents with access to a new agricultural technology can be anticipated 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 that we are interested in knowing about from a policy perspective and to quantify overall impacts. Emerick et al. (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 increased their cultivated area, fertilizer usage, and credit demand. It is through these behavioral responses that the returns to the new technology were substantially increased.

Managing the quality of the design and implementation of RCTs is essential to avoid pitfalls and assure that relevant lessons can be learned. 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. 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. Agronomic trials conducted with farmers are often still highly controlled by the researchers 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 could potentially be profitable. The conclusions of such calculations guide not only dissemination efforts, but also further research efforts. Yet yield gains obtained in typical agronomic trials are not very representative of yield gains the average farmer could obtain in real-world settings. Laajaj et al. (2020) 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. They focus on the role of farmer and plot selection, but also on measurement questions, and the role of effort and complementary technical advice—all sources of bias that are regularly overlooked. Accounting for these factors leads to large adjustments in yield and yield increment calculations. These results in turn help understand the dynamic learning and adoption patterns by different types of farmers following the trials (Laajaj & Macours 2016). Similar collaborations between biophysical 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.

3.2. Methodological Pluralism

Well-designed and implemented RCTs can generally provide more rigorous causal identification than a reliance on observational data and econometric tools that are necessarily based on more stringent assumptions. Because the researcher directly manipulates the treatment assignment and hence by design can assure that treatment assignment is not correlated with possible confounders, causal inference requires a more limited set of assumptions than alternative microeconometric methods. Therefore, when the objective is to learn the impact of a newly developed technology at the microlevel and before it is widely diffused, RCTs can provide a level of rigor higher than that attainable with other methods. Meaningful lessons can particularly be learned from the RCTs when they are preceded and informed by good diagnostics through preliminary qualitative work and piloting.

Sometimes, however, the key impact question one aims to answer is about impact and effectiveness once innovations have scaled across broad agricultural landscapes, well beyond the geographic

scope of an original causal design. Or we may want to know about long-term effects some years after initial rollout when farmers have been able to learn and we could expect their behavioral reactions to have changed. In certain cases, the exogenous variation in exposure created by the initial randomized assignment will persist even after many years, allowing for a longer-term analysis of dynamic adoption and disadoption decisions. In other settings, rigorous long-term causal estimates need to rely on methods other than randomized assignment. Meenakshi et al. (2021) discuss designs of a series of nonexperimental studies, using panel methods and two-way fixed-effect estimations relying on roll-out data of technologies, or agronomical advice, and provide promising examples. Finally, in examining aggregate impacts over time or space, there is potential promise in combining estimates of impact from RCTs with estimates of adoption established using observational studies. This is an area that needs further validation research.

Measuring the impacts of policy influence work or institutional changes can also be challenging with an RCT, except where these have localized effects. Research focused on agricultural policies and institutions typically does not have large populations of potential users/adopters, as is the case for farmer-focused technologies. For these research investments, theory-based approaches—defined by White (2009, p. 272) as "examining the assumptions underlying the causal chain from inputs to outcomes and impact"—offer much potential. Even so, they need to account for the fact that many policy interventions and/or institutional innovations are often part of complex designs that aim to address multiple development challenges. Focusing on a subset of such packages risks failing to evaluate synergies between program parts (Stern et al. 2012). For those agricultural, food, or natural resource policy innovations targeting individual households or communities, lessons can be learned from rigorous studies around social protection programs, where there is now a long tradition of evaluating the impact of complex packages and policy innovations (Quisumbing et al. 2020), going from the original body of evidence on conditional cash transfers (reviewed by Parker & Todd 2017) to the more recent evidence on graduation approaches (Banerjee et al. 2015).

Impact assessment of research aimed at landscape-level NRM, policy, and institutions often needs to tackle two additional challenges. Namely that there is rarely (*a*) a large number of observations (e.g., nation states adopting a new policy or watersheds taking on a specific approach to NRM), and (*b*) homogeneity in the treatment because contextual adaptation of institutions, NRM practices, and policies is the norm, not the exception. As this will make it hard to identify a source of exogenous variation needed for microlevel causal inference, the methods that can apply where these two conditions hold (small N, i.e., very limited population of units to observe, and varying or context-dependent treatments) will be quite different from situations where they do not. Defining what it means to do rigorous impact evaluation in such cases is an urgent and important task. Given the scale of investment in policy, institutions, and landscape-level NRM research, this is a space where methodological advances with and by the international agricultural research community can make a major contribution to the scientific community more broadly.

3.3. Meta-Analyses and Ensuring the Correct Baby:Bathwater Ratio

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 "appraise and synthesize the available high-quality evidence" (**https://www.3ieimpact.org/evidence-hub/publications/systematic-reviews**) on a specific question, are increasingly influential in evidence-based policy in medicine (through the Cochrane Collaboration), social programs (Campbell Collaboration), and the international development sector writ large (3ie). They can indeed be powerful tools to aggregate evidence across studies conducted in different contexts and using different methods and to derive conclusions with wider external validity than individual studies. To

conduct such reviews, researchers typically comprehensively and systematically search the universe of published evidence on a specific question and can filter out studies based on specified quality criteria. The devil is in the details of these quality criteria—if they are too loose we increase the risk that low-quality or biased studies will influence the overall result; when they are too strict 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 the UK government's Department for International Development, 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. However, only five of these papers passed the second-stage screening. Rejecting 209 out of 214 papers hampered the ability to answer the broad question on heterogeneity that the review aimed to answer. Herdt & Mine (2017) revisited the same set of studies using a slightly less restrictive set of criteria, retaining 30 studies—still a considerable cull of 184 papers that were relevant but simply not of sufficient quality for inclusion. They also set out to answer a more modest question and ultimately established that of the 30 studies, most pointed to a positive relationship between technology use and productivity and income.

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. Given the methodological diversity of the retained studies—a mix of RCTs and econometric analysis from observational data—an excellent feature of this 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. Garbero et al. (2018) take a further step. In reviewing the impact 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% 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 issue of standards of evidence when comparing across methodologies remains a challenge, but the risk-of-bias scoring carried out by Garbero et al. (2018) and Stewart et al. (2015) show us productive ways forward. There is much more work to be done on the specific approaches to statistical meta-analyses that are appropriate for agricultural innovations for which a critical mass of rigorous studies have been carried out. Gechter & Meager (2022), for instance, provide a meta-analysis method to combine observational studies with possible bias in causal estimates with experimental studies with possible site selection bias. Given the context specificity of many agricultural innovations, such advances are likely particularly valuable for moving toward more credible and actionable conclusions from impact assessments on innovations resulting from international agricultural research.

Finally, the existing meta-studies also raise a different type of concern: Even if they can exclude studies based on the lack of rigorous evidence, or include bias corrections, their conclusions are necessarily derived from studies that have been conducted and published. However, the technologies subjected to impact assessment are typically cherry-picked in the first place. Moreover, the population of published studies may have been subject to the so-called file-drawer problems and publication bias (Christensen & Miguel 2018, Andrews & Kasy 2019). Together, these factors bias the distribution of published results upward, relative to a strategy of selecting, evaluating,

and publishing assessments from a random sample of candidate technologies. The development of trial registries and preanalysis plans aims to address these concerns (Olken 2015, Banerjee et al. 2020), but to date, these approaches are largely used only for RCT studies, pointing to important room for improvement in studies relying on other methods for causal inference.

4. UNDERSTANDING AGRICULTURAL RESEARCH IMPACTS ON A LARGE SCALE

4.1. Accurate Diffusion Data at a Policy-Relevant Scale

Empirical studies establishing causal evidence on farmers' behavioral responses to the availability of new technologies and practices are a key part of establishing credible evidence of impacts of agricultural research. The low take-up of innovations that is often observed in real-life settings contains equally important information about the potential profitability of innovations that is all too often ignored. If farmers decide not to adopt a new technology or practice, or when policy makers decide not to pursue proposed institutional innovations, 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 adoption studies are commissioned for individual technologies, they often involve small, nonrepresentative samples and short time frames (Doss 2006) and are consequently of limited value. The question should not be whether some selected farmers adopted a particular innovation once, but rather whether a large share of farmers representative of a population targeted by the innovation decide to adopt and continue to use the innovation 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. Such a vision is best achieved through partnerships with institutions that have a comparative advantage in surveys in countries of highest priority to the international agricultural research community, such as the national statistical institutes and the World Bank. By building on existing investments in surveys designed to help countries monitor their progress on the SDGs, such partnerships that help improve and build in measurement of innovations resulting from international agricultural research also form the basis for subsequent analytical work linking innovations to the final outcomes targeted by the research investments.

Such a philosophy underpins a series of long-run studies in progress by the CGIAR Standing Panel on Impact Assessment in Ethiopia, Uganda, Vietnam, and Bangladesh—all countries with high past and current research investments. The rationale of these country studies is to work with the statistical agencies of these high-priority countries to embed new data collection protocols (including DNA fingerprinting, new modules on NRM, livestock, water management, etc.) into already existing nationally representative household panel surveys. The first major output from this strand of research is the "Shining a Brighter Light" report on Ethiopia (Kosmowski et al. 2020).

This study starts from a compilation of comprehensive information on agricultural research activities in Ethiopia over a 20-year period (2000–2020) through interviews with research leaders, scientists, and government officials and an extensive review of published studies and project documents. This stocktaking exercise led to the identification of 52 innovations and 26 claims of policy influence. Data collection protocols for a subset of 18 research-related innovations for which background evidence suggested they had scaled were then integrated into the Ethiopian Socioeconomic Survey (ESS), a regionally and nationally representative household panel survey in 2015–2016 and 2018–2019.[3] Protocols were based on validation experiments and other

[3]A further round (2021–2022) was recently completed, with most of the same data collection protocols retained. Data should be in the public domain in late 2023.

measurement improvement work (Kosmowski et al. 2017, 2019) assuring that best-practice, objective, and reliable data were collected. These then yield credible estimates that between 4.1 and 11.0 million Ethiopian households had been reached by CGIAR-related innovations. The lower bound estimate is for those innovations with observable features that allow for a confident link back to CGIAR research efforts. The upper bound should be interpreted as the potential reach: the number of households that in theory could benefit. Although many innovations are being adopted by some farmers, only a few are reaching large numbers of households in Ethiopia. This is arguably as expected, given the large uncertainties underlying agriculture research for development and inherent to any innovation system. Even so, by considering the innovations stemming from the broad effort of the international agricultural research efforts in Ethiopia all together, the evidence provides a system-level picture across many different innovations and science areas. The three innovations with the largest reach—soil and water conservation practices, improved maize varieties, and crossbred poultry—span three major domains of agricultural research (i.e., NRM, crop breeding, and livestock research). There is also a common thread of supportive government policies, in turn influenced by policy research, that undergirds these success stories. As such, the evidence clearly points to the importance of analyzing socio-technical bundles together (Barrett et al. 2020), while allowing for testing of synergies and substitution at the farm level.

Embedding measures of innovations into socioeconomic surveys allows one to not just calculate the aggregate numbers, but also analyze who is adopting, which can provide crucial feedback evidence into assumptions underlying the theories-of-change of the individual innovations. Relatedly, the panel nature of the surveys facilitates analyzing dynamic changes, tracing scaling where it happens, and also documenting disadoption when that is the case. Measuring disadoption is arguably much more important than most agricultural researchers acknowledge. It can help reveal where new technologies and practices failed to deliver returns for the farmers who tried them or otherwise became obsolete. 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 researchers or centers to pursue such an endeavor may be unclear, which argues for an independent entity to facilitate the kinds of longitudinal country-level analyses that can help to understand these dynamics.

4.2. Causal Linkages Across Scales

Some impact pathways from agricultural research are complex, particularly those mediated by markets and over borders, suggesting the need for models at a macrolevel. The microfoundations 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 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 TFP, and the models then allow simulation of how a new technology or innovation leads to changes in the TFP level or growth rate and further outcomes. Wiebe et al. (2021) present a thoughtful cross-commodity application of this approach.[4] 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

[4]The widely used IFPRI IMPACT model (as presented by Robinson et al. 2015) offers an example of this kind of analysis; the structure of the model makes it well suited for altering productivity in one or many commodities at the level of a single country (De Pinto et al. 2017, Mason d'Croz et al. 2020) or multiple countries.

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 weather-related factors). As noted above, even with the richest data, measurement of TFP is challenging, given the heterogeneity and stochasticity that are intrinsic to agricultural production systems (Gollin & Udry 2021). To some extent, these challenges are reduced when micro data are aggregated (as pointed out by Aragón et al. 2022), but macro data are not always constructed from aggregation, at least in many low-income-country contexts. National-level statistics on agricultural production have frequently been criticized on the grounds that the underlying methods are opaque and the data cannot be linked to any explicit sampling strategy.

A different problem is that macro models used for impact assessment often require estimates of average TFP changes over broad geographies. Calibrating these models with data from micro studies conducted in narrower geographies and location-specific estimates is problematic. For instance, if the micro studies have been carried out in locations that are particularly favorable to the technology, the models will tend to overestimate the benefits of the technology. Thus, a challenge remains in finding appropriate micro evidence to feed into the macro models.

A further problem is that macro models necessarily and inevitably build in strong assumptions about functional forms as well as model closure assumptions. These assumptions are difficult to test or 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 labor 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.

Similarly, 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, such as the widely used GTAP model (Corong et al. 2017) or the MIRAGRODEP model (Laborde et al. 2013). 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.

How should we understand rigor in the context of models? The challenge is not a new one; see, for example, reflections from 20 years ago by Devarajan & Robinson (2002) or Valenzuela et al. (2007). The generally accepted best practice is to validate models by testing them against data other than those that were used to calibrate them. Because 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 (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 test these deeper structures is to engage in a kind of historical forecasting. One might, for instance, take one of these commonly used models and, instead of calibrating it to the data from 2000 to 2015, calibrate it instead to data from 1985 to 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 to 2015 and feed it with base year data from 1985 to see how well it matches observations from 1985 to 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 the 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 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 nonagricultural 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

Cataloguing and tracking the outputs from the international agricultural research system and determining whether and how these outputs lead to development outcomes are an enterprise worthy of significant investment. International and national research centers are incentivized to advocate for their own effectiveness. In the absence of strong and consistent demand for rigor from donors, 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.

The primary objective of impact evaluation studies and subsequent systematic reviews should be to help the international agricultural research system reach its goals in orienting new research toward areas with potentially high payoffs. This 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.

The rigor revolution demands that we do better in the following ways. First, we need to institutionalize reliable data collection related to international agricultural research activities along the results chain from investments to outputs to outcomes. One practical approach toward this goal is to focus on a few key countries where large investments have occurred as a first step toward catching up after years of neglect, following the template set by the "Shining a Brighter Light" study (Kosmowski et al. 2020). This helps the international agricultural research community to reconnect with its historical track record of collecting longitudinal data, best illustrated by the large body of literature resulting from the longitudinal ICRISAT villages data sets. The potential

for a new generation of longitudinal studies lies in combining carefully implemented geolocated surveys featuring DNA fingerprinting of the major crops and livestock, reliable data on farmers' management practices, and detailed socioeconomic data, with information on the policy and institutional environment. Data quality should be of the highest priority. If this is done right, it also allows taking 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. In articulating a theory of change for a new research-derived innovation there typically are a few key assumptions for which there is substantial uncertainty about whether they hold, which in turn can have large implications for the potential of the innovation to scale. These uncertainties should help inform where scarce resources for impact evaluation should be allocated. This suggests shifting away from searching for what works in the abstract and toward finding out why certain things work and others do not in particular contexts. As highlighted in myriad ways throughout this review, 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. Finally, making methodological breakthroughs on tracing policy influence or measuring the outcomes from capacity-building efforts remain challenges in the future. More generally, the integration across data types and methodological approaches offers tremendous potential.

DISCLOSURE STATEMENT

The authors are Senior Researcher, Chair, and former Chair, respectively, of the Standing Panel on Impact Assessment of the CGIAR, a global partnership of agricultural research organizations dedicated to food security and poverty reduction in the developing world.

ACKNOWLEDGMENTS

Comments by Daniel Gilligan and Frank Place on an earlier version of the manuscript are gratefully acknowledged.

LITERATURE CITED

Abay K, Wossen T, Abate GA, Stevenson JR, Michelson H, Barrett CB. 2023. Inferential and behavioral implications of measurement error in agricultural data. *Annu. Rev. Resour. Econ.* 15:63–83

Aker JC, Jack K. 2021. *Harvesting the rain: the adoption of environmental technologies in the Sahel*. NBER Work. Pap. 29518

Alix-Garcia J, Millimet DL. 2023. Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. *J. Assoc. Environ. Resour. Econ.*

Alston JM, Norton G, Pardey PG. 1995. *Science Under Scarcity: Principles and Practice for Agricultural Research Evaluation and Priority Setting*. Ithaca, NY: Cornell Univ. Press

Andrews I, Kasy M. 2019. Identification of and correction for publication bias. *Am. Econ. Rev.* 109(8):2766–94

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91(434):444–55

Aragón FM, Restuccia D, Rud JP. 2022. *Assessing misallocation in agriculture: Plots versus farms*. NBER Work. Pap. 29749

Baird S, Bohren A, McIntosh C, Özler B. 2018. Optimal designs of experiments in the presence of interference. *Rev. Econ. Stat.* 100(5):844–60

Banerjee A, Duflo E, Finkelstein A, Katz LF, Olken BA, Sautmann A. 2020. *In praise of moderation: suggestions for the scope and use of preanalysis plans for RCTs in economics*. NBER Work. Pap. 26993

Banerjee A, Duflo E, Goldberg N, Karlan D, Osei R, et al. 2015. A multifaceted program causes lasting progress for the very poor: evidence from six countries. *Science* 348:1260799

Barrett CB. 2003. *Natural resources management research in CGIAR: a meta-evaluation*. Themat. Work. Pap. 27799, World Bank, Washington, DC. **https://documents1.worldbank.org/curated/en/171131468135928440/pdf/11050277990CGIAR.pdf**

Barrett CB, Agrawal A, Coomes OT, Platteau J-P. 2009. *Stripe review of social sciences in the CGIAR*. Work. Pap., CGIAR Sci. Counc., Rome

Barrett CB, Benton TG, Cooper KA, Fanzo J, Gandhi R, et al. 2020. Bundling innovations to transform agrifood systems. *Nat. Sustain.* 3(12):974–76

Barrett CB, Carter MR. 2010. The power and pitfalls of experiments in development economics: some nonrandom reflections. *Appl. Econ. Perspect. Policy* 32:515–48

Barrett CB, Upton JB. 2013. Food security and sociopolitical stability in Sub-Saharan Africa. In *Food Security and Sociopolitical Stability*, ed. CB Barrett, pp. 323–56. Oxford, UK: Oxford Univ. Press

Beaman L, BenYishay A, Magruder J, Mobarak AM. 2021. Can network theory-based targeting increase technology adoption? *Am. Econ. Rev.* 11(6):1918–43

Beaman L, Dillon A. 2018. Diffusion of agricultural information within social networks: evidence on gender inequalities from Mali. *J. Dev. Econ.* 133:147–61

Beegle K, Carletto C, Himelein K. 2012. Reliability of recall in agricultural data. *J. Dev. Econ.* 98(1):34–41

Behaghel L, Gignoux J, Macours K. 2020. *Social learning in agriculture: does smallholder heterogeneity impede technology diffusion in Sub-Saharan Africa?* CEPR Discuss. Pap. 15220, Cent. Econ. Policy Res., London

BenYishay A, Mobarak M. 2018. Social learning and incentives for experimentation and communication. *Rev. Econ. Stud.* 86(3):976–1009

Bernard T, Lambert S, Macours K, Vinez M. 2023. Impact of small farmers' access to improved seeds and deforestation in DR Congo. *Nat. Commun.* 14:1603

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. *Am. J. Agric. Econ.* 96(3):813–30

Calvi R, Penglase J, Tommasi D. 2022. Measuring women's empowerment in collective households. *AEA Pap. Proc.* 112:566–60

Carletto C, Dillon A, Zezza A. 2021. Agricultural data collection to minimize measurement error and maximize coverage. In *Handbook of Agricultural Economics*, ed. CB Barrett, DR Just, Vol. 5, pp. 4407–80. Amsterdam: Elsevier

Carletto C, Gourlay S, Murray S, Zezza A. 2016. *Land area measurement in household surveys: Empirical evidence and practical guidance for effective data collection*. Work. Pap. 147692, LSMS Guidebook, World Bank, Washington, DC

Christensen G, Miguel E. 2018. Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* 56(3):920–80

Corong EL, Hertel TW, McDougall R, Tsigas ME, Van Der Mensbrugghe D. 2017. The standard GTAP model, version 7. *J. Glob. Econ. Anal.* 2(1):1–119

Dalrymple DG. 1978. *Development and Spread of High-Yielding Varieties of Wheat and Rice in the Less Developed Nations*. Foreign Agric. Econ. Rep. 95. Washington, DC: USAID

Dar M, de Janvry A, Emerick K, Sadoulet E, Wiseman E. 2020. *Private input suppliers as information agents for technology adoption in agriculture*. Work. Pap., AFC India Ltd., Kerala

Deaton A, Cartwright N. 2017. *Understanding and misunderstanding randomized controlled trials*. NBER Work. Pap. 22595

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*. Rep., Indep. Sci. Partnersh. Counc. Secr., Rome

de Janvry A, Sadoulet E, Suri T. 2017. Field experiments in developing country agriculture. In *Handbook of Economic Field Experiments*, ed. AV Banerjee, E Duflo, Vol. 2, pp. 427–66. Amsterdam: North-Holland

De Pinto A, Wiebe K, Pacheco P. 2017. Help bigger palm oil yields to save land. *Nature* 544:416

De Weerdt J, Gibson J, Beegle K. 2020. What can we learn from experimenting with survey methods? *Annu. Rev. Resour. Econ.* 12:431–47

Devarajan S, Robinson S. 2002. *The influence of computable general equilibrium models on policy*. TMD Discuss. Pap. 98, Trade Macroecon. Div., Int. Food Policy Res. Inst., Washington, DC

Dillon A, Karlan D, Udry C, Zinman J. 2020. Good identification, meet good data. *World Dev.* 127:104796
Doss C, Kieran C, Kilic T. 2020. Measuring ownership, control, and use of assets. *Fem. Econ.* 26(3):144–68
Doss CR. 2006. Analyzing technology adoption using microstudies: Limitations, challenges, and opportunities for improvement. *Agric. Econ.* 34(3):207–19
Emerick K, Dar M. 2021. Farmer field days and demonstrator selection for increasing technology adoption, *Rev. Econ. Stat.* 103(4):680–93
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, Fond. Étud. Rech. Dev. Int., Clermont-Ferrand, Fr.
Erenstein O, Laxmi V. 2008. Zero tillage impacts in India's rice–wheat systems: a review. *Soil Tillage Res.* 100(1–2):1–14
Fuglie K. 2018. R&D capital, R&D spillovers, and productivity growth in world agriculture. *Appl. Econ. Perspect. Policy* 40(3):421–44
Fulginiti L. 2010. Estimating Griliches' *k*-shifts. *Am. J. Agric. Econ.* 92(1):86–101
Garbero A, Marion P, Brailovskaya V. 2018. *The Impact of the Adoption of CGIAR's Improved Varieties on Poverty and Welfare Outcomes*: *A Systematic Review*. IFAD Res. Ser. 33. Rome: IFAD
Gechter M, Meager R. 2022. *Combining experimental and observational studies in meta-analysis: a debiasing approach*. Work. Pap., Penn State Univ./London Sch. Econ.
Gollin D, Probst LT, Brower E. 2018. *Assessing Poverty Impacts of Agricultural Research: Methods and Challenges for CGIAR*. Rome: Indep. Sci. Partnersh. Counc.
Gollin D, Udry C. 2021. Heterogeneity, measurement error, and misallocation: evidence from African agriculture. *J. Political Econ.* 129(1). **https://doi.org/10.1086/711369**
Griliches Z. 1957. Hybrid corn: an exploration in the economics of technological change. *Econometrica* 25(4):501–22
Griliches Z. 1958. Research costs and social returns: hybrid corn and related innovations. *J. Political Econ.* 66(5):419–431
Herdt RW, Mine S. 2017. *Does modern technology increase agricultural productivity? Revisiting the evidence from Loevinsohn et al.* Tech. Note, Indep. Sci. Partnersh. Counc. Secr., Rome
Hurley TM, Pardey PG, Rao X, Andrade RS. 2016. *Returns to food and agricultural R&D investments worldwide, 1958–2015*. Brief., InSTePP, St. Paul, MN
Imbens G. 2018. Understanding and misunderstanding randomized controlled trials: a commentary on Deaton and Cartwright. *Soc. Sci. Med.* 210:50–52
Jaleta M, Tesfaye K, Kilian A, Yirga C, Habte E, et al. 2020. Misidentification by farmers of the crop varieties they grow: lessons from DNA fingerprinting of wheat in Ethiopia. *PLOS ONE* 15(7):e0235484
Jayachandran S, Biradavolu M, Cooper J. 2023. Using machine learning and qualitative interviews to design a five-question women's agency index. *World Dev.* 161:106076
Jayachandran S, De Laat J, Lambin EF, Stanton CY, Audy R, Thomas NE. 2017. Cash for carbon: a randomized trial of payments for ecosystem services to reduce deforestation. *Science* 357:267–73
Kilic T, Sohnesen T. 2015. *Same question but different answer: experimental evidence on questionnaire design's impact on poverty measured by proxies*. Policy Res. Work. Pap. 7182. World Bank, Washington, DC. **https://ideas.repec.org/p/wbk/wbrwps/7182.html**
Kosmowski F, Alemu S, Mallia P, Stevenson J, Macours K. 2020. *Shining a brighter light: Comprehensive evidence on adoption and diffusion of CGIAR-related innovations in Ethiopia*. Synth. Rep., Stand. Panel Impact Assess., Rome, It.
Kosmowski F, Aragaw A, Kilian A, Ambel A, Ilukor J, et al. 2019. Varietal identification in household surveys: results from three household-based methods against the benchmark of DNA fingerprinting in southern Ethiopia. *Exp. Agric.* 55(3):371–85
Kosmowski F, Stevenson J, Campbell J, Ambel A, Tsegay AH. 2017. On the ground or in the air? A methodological experiment on crop residue cover measurement in Ethiopia. *Environ. Manag.* 60:705–16
Laajaj R, Macours K. 2016. *Learning-by-doing and learning-from-others: evidence from agronomical trials in Kenya*. Policy Brief, Fond. Étud. Rech. Dev. Int., Clermont-Ferrand, Fr.
Laajaj R, Macours K. 2021. Measuring skills in developing countries. *J. Hum. Resour.* 56(4):1254–95

Laajaj R, Macours K, Masso C, Thuita M, Vanlauwe B. 2020. Reconciling yield gains in agronomic trials with returns under African smallholder conditions. *Sci. Rep.* 10(1):14286

Laborde D, Robichaud V, Tokgoz S. 2013. *MIRAGRODEP 1.0: documentation*. Tech. Note, AGRODEP, Washington, DC. **https://www.agrodep.org/resource/no-20-miragrodep-10-documentation**

Loevinsohn M, Sumberg J, Diagne A, Whitfield S. 2013. *Under what circumstances and conditions does adoption of technology result in increased agricultural productivity?* Rep., Inst. Dev. Stud., Brighton, UK

MacAskill W. 2015. *Doing Good Better: How Effective Altruism Can Help You Help Others, Do Work That Matters, and Make Smarter Choices About Giving Back*. New York: Penguin

Macours K. 2019. Farmers' demand and the traits and diffusion of agricultural innovations in developing countries. *Annu. Rev. Resour. Econ.* 11:483–99

Magnan N, Spielman DJ, Lybbert TJ, Gulati K. 2015. Leveling with friends: social networks and Indian farmers' demand for a technology with heterogeneous benefits. *J. Dev. Econ.* 116:223–51

Mallia P. 2022. *You reap what (you think) you sow? Evidence on farmers' behavioral adjustments in the case of correct crop varietal identification*. Work. Pap. 2022-08, Paris Sch. Econ.

Maredia M, Reyes B, Manu-Aduening J, Dankyi A, Hamazakaza P, et al. 2016. *Testing alternative methods of varietal identification using DNA fingerprinting: results of pilot studies in Ghana and Zambia*. Food Secur. Int. Dev. Work. Pap. 149, Mich. State. Univ., East Lansing

Mason-D'Croz D, Bogard JR, Herrero M, Robinson S, Sulser T, et al. 2020. Modelling the global economic consequences of a major African swine fever outbreak in China. *Nat. Food* 1(4):221–28

Meenakshi JV, Johnson N, Karasalo M. 2021. *Designing quasi-experimental impact studies of agricultural research at scale*. Tech. Note 10, Stand. Panel Impact Assess., Rome

Oakes JM. 2018. The tribulations of trials: a commentary on Deaton and Cartwright. *Soc. Sci. Med.* 210:57–59

OECD (Organ. Econ. Co-op. Dev.). 2005. *Paris Declaration on Aid Effectiveness*. Paris: OECD Publ. **http://dx.doi.org/10.1787/9789264098084-en**

OECD (Organ. Econ. Co-op. Dev.). 2008. *Accra Agenda for Action*. Paris: OECD Publ. **http://dx.doi.org/10.1787/9789264098107-en**

Olken B. 2015. Promises and perils of pre-analysis plans. *J. Econ. Perspect.* 29(3):61–80

Parker SW, Todd PE. 2017. Conditional cash transfers: the case of Progresa/Oportunidades. *J. Econ. Lit.* 55(3):866–915

Pingali P. 2012. Green revolution: impacts, limits, and the path ahead. *PNAS* 109(31):12302–8

Poets A, Silverstein K, Pardey P, Hearne S, Stevenson J. 2020. *DNA Fingerprinting for Crop Varietal Identification: Fit-for-Purpose Protocols, Their Costs and Analytical Implications*. Rome: SPIA

Pritchett L. 2002. It pays to be ignorant: a simple political economy of rigorous program evaluation. *J. Policy Reform* 5:251–69

Quisumbing AR, Ahmed A, Gilligan DO, Hoddinott J, Kumar N, et al. 2020. Randomized controlled trials of multi-sectoral programs: lessons from development research. *World Dev.* 127:104822

Quisumbing AR, Sproule K, Martinez EM, Malapit H. 2021. Do tradeoffs among dimensions of women's empowerment and nutrition outcomes exist? Evidence from six countries in Africa and Asia. *Food Policy* 100:102001

Raitzer DA, Kelley TG. 2008. Assessing the contribution of impact assessment to donor decisions for international agricultural research. *Res. Eval.* 17:187–99

Rao X, Hurley TM, Pardey PG. 2020. Recalibrating the reported returns to agricultural R&D: What if we all heeded Griliches? *Aust. J. Agric. Resour. Econ.* 64(3):977–1001

Renkow M, Byerlee D. 2010. The impacts of CGIAR research: a review of recent evidence. *Food Policy* 35(5):391–402

Robinson S, Mason-D'Croz D, Islam S, Sulser T, Robertson RD, et al. 2015. *The international model for policy analysis of agricultural commodities and trade (IMPACT): model description for version 3*. IFPRI Discuss. Pap. 1483, Int. Food Policy Res. Inst., Washington, DC

Rosenzweig MR, Udry C. 2020. External validity in a stochastic world: evidence from low-income countries. *Rev. Econ. Stud.* 87(1):343–81

Savedoff WD, Levine R, Birdsall N. 2006. *When will we ever learn? Improving lives through impact evaluation*. Rep., Cent. Glob. Dev., Washington, DC

Skoufias E. 2005. *PROGRESA and its impacts on the welfare of rural households in Mexico*. Res. Rep. 139, Int. Food Policy Res. Inst., Washington, DC

Stern E, Stame N, Mayne J, Forss K, Davies R, Befani B. 2012. *Broadening the range of designs and methods for impact evaluations: report of a study commissioned by the Department for International Development*. Rep. 38, Dep. Int. Dev., London

Stevenson J, Gantier M, Traxler G, Kosmowski F, Macours K. 2023. *The challenge of tracking the reach of post-green revolution crop breeding*. Preprint, CGIAR Standing Panel on Impact Assessment, Rome. **https://doi.org/10.21203/rs.3.rs-3028333/v1**

Stevenson J, Vanlauwe B, Macours K, Johnson N, Krishnan L, et al. 2019. Farmer adoption of plot-and farm-level natural resource management practices: between rhetoric and reality. *Glob. Food Secur.* 20:101–4

Stevenson J, Vlek P. 2018. *Assessing the adoption and diffusion of sustainable agricultural practices: synthesis of a new set of empirical studies*. Rep., Stand. Panel Impact Assess., Paris

Stewart R, Langer L, Rebelo da Silva N, Muchiri E, Zaranika 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 Syst. Rev.* 11:1–224

Valenzuela E, Hertel TW, Keeney R, Reimer JJ. 2007. Assessing global computable general equilibrium model validity using agricultural price volatility. *Am. J. Agric. Econ.* 89(2):383–97

Walker TS, Crissman CC. 1996. *Case studies of the economic impact of CIP related technologies*. Rep., Int. Potato Cent., Lima, Peru

Walker TS, Maredia M, Kelley T, La Rovere R, Templeton D, et al. 2008. *Strategic guidance for* ex post *impact assessment of agricultural research*. Rep., CGIAR Sci. Counc., Rome

White H. 2009. Theory-based impact evaluation: principles and practice. *J. Dev. Effect.* 1(3):271–84

Wiebe K, Sulser TB, Dunston S, Rosegrant MW, Fuglie K, et al. 2021. Modeling impacts of faster productivity growth to inform the CGIAR initiative on Crops to End Hunger. *PLOS ONE* 16(4):e0249994

Wossen T, Girma G, Abdoulaye T, Rabbi I, Olanrewaju A, Alene A, et al. 2017. *The cassava monitoring survey in Nigeria*. Rep., Int. Inst. Trop. Agric., Ibadan, Niger.