CGIAR Standing Panel on Impact Assessment

Interpolating and Extrapolating Impact Estimates: Implications of Ex Post Impact Evaluation Evidence for Projecting Benefits Ex Ante

Ricardo Labarta, Travis J. Lybbert, Astewale Melaku, Madeleine Walker, Chunhao Yang

September 4, 2025

The purpose of this note is to summarize empirical literature in economics addressing two key evidence challenges: (i) scaling up innovations or interventions that have shown promising impacts at smaller scales, and (ii) assessing the external validity of impact estimates. In this context, scaling refers to expanding an innovation's reach—across larger populations, geographic areas, or time periods—based on existing evidence. External validity concerns the extent to which study results can be generalized to different contexts. We aim to provide guidance on the challenges and limitations of transferring impact estimates both within and beyond their original study settings.

Aknoweldgements: SPIA colleagues

Citation: Labarta, R., Lybbert, T.J., Melaku, A., Walker, M., & Yang, C. (2025). Interpolating and extrapolating impact estimates: Implications of ex post impact evaluation evidence for projecting benefits ex ante. Technical Note N.13. Rome: SPIA.

Cover image: Grass to Cash, Kenya, Credit: 2016/CIAT Georgina Smith

Design and layout: Luca Pierotti and Macaroni Bros

Interpolating and Extrapolating Impact Estimates: Implications of Ex Post Impact Evaluation Evidence for Projecting Benefits Ex Ante

Ricardo Labarta, Travis J. Lybbert, Astewale Melaku, Madeleine Walker, Chunhao Yang

September 4, 2025

Contents

1 Introduction 1

2 CGIAR Context: Background and Motivation 2

3 Scaling Evidence 3

3.1 Government-sponsored scaling 3
3.2 Challenges to maintaining personnel quality at scale 4
3.3 Lessons from simulations 4
3.4 General challenges and possibilities at scale 4

4 External Validity Evidence 5

4.1 Introduction 5
4.2 Methodological approaches to conceptualize external validity 5
4.3 Meta-analyses of external validity using published studies 6
4.4 Empirical findings 6
4.5 Lessons learned 6

5 Discussion 7

References 9

1 Introduction

Reliable evidence about the impact of past development investments on target outcomes can enhance institutional learning and improve future funding decisions. But such evidence is always to some extent or in some dimension incomplete. Existing studies, no matter how rigorous generally do not encompass the entire geographic area, time period, or population funding organizations and policymakers would like to understand before scaling a policy, program, intervention, or innovation. While the desire to see the complete evidence 'landscape' across time, space, or population is understandably strong, producing such an evidence landscape is not cost-effective in most settings. Instead, those hoping to make informed, evidence-based decisions must cope with significant evidence gaps.

The gap in evidence across all desired dimensions is not a new problem. Applied researchers have long grappled such gaps and proposed various remedies. For example, there are many statistical techniques to interpolate missing data from existing samples, such as the response to a question that went unanswered for an individual who was surveyed and completed other questions. While these techniques are generally accepted, they are imperfect and subject to abuse. Expanding interpolation of data points to new samples, outside those for which data have already been collected, such as new regions, populations, or time periods is naturally even more difficult. Interpolating or extrapolating reliable estimates of impact of an innovation or policy beyond the specific context that researchers used to produce the evidence is far more challenging.

Reliable impact estimates emerge from the careful combination of a research design that conceptually enables the identification of causal effects and appropriate measures of relevant outcomes. In practice, combining these two elements often reflects key features of the study context, including how the innovation was used, or the intervention was implemented. Within this context, estimates from a well-designed and well-executed impact assessment are meaningful, but typically lose their reliability the further one ventures from this native context. As one attempts to teleport impact estimates from one setting to another, the political and economic systems, culture, climate, and norms likely change, potentially changing the impact of the innovation in a new setting. If the results of an impact assessment are extended to a longer time period, demographic composition, weather patterns, and overall development of the study area could all change dramatically depending on the length of time period analyzed. Finally, in a case where one wants to extrapolate the results of one innovation to an entire country, the pure magnitude of the intervention could mean that it is implemented differently in the entire country case. Similarly, widespread implementation of the intervention could mean a new level of equilibrium prices or practices is reached at the country-level, changing the impact of the innovation altogether.

The purpose of this note is to summarize a few strands of the empirical literature in economics that speak to these challenges, namely, the evidence (i) on scaling up innovations or interventions that showed promising impact at smaller scale (e.g., during a piloting phase) and (ii) on assessing the external validity of impact estimates. Scaling an innovation means expanding its reach to a larger geographic area, longer time period, or bigger population. Scaling, in the context of this note, typically refers to the application of innovations with existing evidence to larger contexts. External validity is the degree to which the results of a study can be applied to other geographic areas, time periods, or populations. Through our review of these two strands of the literature, we aim to provide guidance on the challenges and limitations of interpolating and extrapolating impact estimates between and beyond their native study contexts, respectively.

2 CGIAR Context: Background and Motivation

For over 50 years, CGIAR has maintained its critical role in advancing international agricultural research for development. This achievement has been possible due to a continuing process of institutional change and a research agenda that has adapted to an evolving funding environment. With the decline of unrestricted funding, CGIAR is under greater pressure to prioritize the allocation of limited unrestricted funds and to justify new investments in the system. Prioritization is never easy, and these processes have faced a host of challenges.

During the creation of the CGIAR Research Programs (CRP) in 2011, CRP teams were asked to set outcome targets of the program as a contribution to System Level Outcomes. The proposed approach for the estimation relied on expert knowledge and literature-based estimates of expected form and quantity of change (i.e. productivity change) combined with target populations to be reached by 2022. Unfortunately, this process did not provide the right incentives to base projections on rigorous evidence of the reach and impacts of CGIAR research, leading the exercise to set unrealistic targets for most of the CRPs (dozens of millions of farmers assisted to exit poverty, dozens of millions of hectares of land restored, etc.)

The One CGIAR reform in 2021 organized the core research portfolio into 33 research initiatives. To justify the investment of each of them and as part of the review of their research proposals, initiatives were asked to estimate projected benefits as plausible levels of impact to which CGIAR contributes. While this exercise was an attempt to link each initiative's theory of change with the projected benefits, this exercise did not incentivize initiatives to clearly establish the mechanisms that map outputs into outcomes and impacts.

The projected benefits promoted the use of modeling and mainly subjective criteria to justify changes in the parameters of the models used in the exercise, which made it difficult to validate model predictions. One of the main assumptions used was that CGIAR could impact areas with large numbers of poor households, but the estimation of projected benefits lacked rigorous evidence to back up this assumption. Although the order of magnitude of projected benefits was not as excessive as estimated for the CRPs, the underlying reach estimates still seemed difficult to justify.

With the new (as of 2024) Science Programs and accelerators, there is again an appetite for prioritization of the CGIAR research within these programs and accelerators. This time the process has been planned in two steps. During the proposal preparation, research teams were asked to identify high level outcomes that according to their high-level theory of change, could lead to achieve outcomes and impacts. Then the programs were asked to identify indicators that can help them to measure changes towards different impact areas. The next steps planned bring a heavier consultation of the Science Programs and accelerators, to first based on expert knowledge, assess the expected level of success of their high-level outcome in achieving potential impacts (very high, high, medium, low). Then based on programs subjective expectation, a team of researchers at IFPRI using modeling will build scenarios on the projected benefits for the whole system (not for science programs). Again, the outlined exercise does not seem to provide the right incentives to use rigorous evidence of the reach and causal impacts of CGIAR research.

As this process is still being implemented, guidance on how to better use existing rigorous evidence on the reach and impacts of CGIAR in this prioritization exercise and to help the system to come up with a

more realistic estimation of the potential impacts of the system seems to be timely. This literature review aims to contribute to this ongoing dialogue. Specifically, it surveys specific recent evidence that speaks to the challenges that are inherent to any attempt to project benefits. This note focuses on rigorous ex post impact evidence that is sheds light on these challenges in the hopes of informing the daunting task of ex ante modeling to generate projections that are sufficiently credible and reliable to be used for investment prioritization within the CGIAR.

3 Scaling Evidence

Evidence of projects with positive results later implemented at scale varies in success. The most successful outcomes of scaled projects are typically those which are attached to existing government programs with established infrastructure. However, government involvement does not necessarily guarantee program success.

Simulated projects at scale are mostly successful yet importantly involve many researcher degrees of freedom. Finally, there are notable challenges to scaling projects that repeatedly appear throughout the literature, such as deterioration of personnel quality with larger programs, and sample selection bias.

3.1 Government-sponsored scaling

In two studies, light-touch early childhood development (ECD) interventions have been added onto existing government-funded outreach programs targeted at child nutrition and parenting with reasonable success. Early childhood development interventions had already been found successful in extremely limited samples in Jamaica in 1991, in the form of .8 SD improvement on the Griffiths cognition test 1 year after a play stimulation intervention by community health workers (Grantham-McGregor, Powell, Walker, and Himes, 1991). Bos, Shonchoy, Ravindran, and Khan (2024) study the impacts of a program in Bangladesh that similarly provided educational materials about the importance of everyday play and communication as well as children's picture books alongside early stimulation counseling by nutrition workers. Workers were already implementing a national nutrition program by distributing micronutrient supplements and deworming medication to the same households. Child outcomes improved in terms of cognition (.17 SD), language (.23 SD), and socio-emotional development (.12-.14 SD) 1-3 months after the intervention. In Colombia, a similar program was attached to an existing parenting program known as FAMI (Attanasio, Baker-Henningham, Bernal, Meghir, Pineda, and Rubio-Codina, 2018). The intervention improved children's development by .16 standard deviations, a very similar magnitude to the Bangladesh study yet smaller than the original Jamaica study. While both positive original results and those at scale are encouraging, their difference in magnitude exemplifies the difficulty in extrapolating impact estimates at scale.

Governments are not always successful in implementing projects at scale, however. In Kenya, Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2018) show that teachers on identical fixed term contracts produce higher learning gains if they are hired by an international NGO in an experiment as opposed to the government at scale which produced zero impact. Political opposition to the fixed term contracts ultimately led to differences in implementation between the two groups. Additionally, DellaVigna and Linos (2020) show that nudge trials, which are behavioral change strategies that subtly influence people's choices by altering the way options are presented, have much higher magnitudes of impact in smaller studies published in academic journals relative to those run by Nudge Units which are housed in governments or organizations and study interventions at scale. In academic journals, the average impact of a nudge is 33.5 percent over the control, relative to the Nudge Unit average impact of 8.1 percent.

While all but one of the aforementioned studies produce positive treatment effects, it is important to note that magnitudes differ greatly between original and scaled studies.

3.2 Challenges to maintaining personnel quality at scale

Mitchell, Mobarak, Naguib, Reimao, and Shenoy (2023) compare pilot and scaled results of a migration loan program in Bangladesh. A loan offers increased temporary migration by 25-40 percentage points, which decreased to 12 percent at scale. The authors demonstrate that the expansion lowered the alignment between the program goals and loan officers' incentives. In the pilot, implementers had more discretion over who received a loan. When the program was scaled, loan officers were more able to respond to their incentive to approve loans based on repayment probability, which entailed more returns for the officers themselves, relative to approving loans for those who might benefit the most from migration enabled by the loan. A state-level intervention to lower class size in California, inspired by the success of the smaller Project STAR in Tennessee, showed similar improvement in math and reading score achievement (Jepsen and Rivkin, 2007). However, these benefits were dampened by the high share of teachers who were hired quickly to allow for smaller classes yet did not have prior experience nor full certification. Both of these studies indicate that personnel quality is a serious challenge to maintaining impact at scale, yet does not completely nullify program effects.

3.3 Lessons from simulations

Simulations provide the exciting ability to model impact on a whole country or city, though there are many research degrees of freedom which should spur some caution in these cases. Bergquist, Faber, Fally, Hoelzlein, Miguel, and Rodr´ıguez-Clare (2022) develop a quantitative model of farm production and use variation from field and quasi-experiments on subsidies for chemical fertilizers and hybrid seed varieties in Uganda to estimate the parameters of their model. They find that both the average and distributional effects differ meaningfully between local interventions and the full country simulation "at scale". The average effect of the subsidy at scale is a 4.4 percent increase in the household's real income. At scale, more than 80 percent of households experience a positive or negative change greater than 10 percent of their local experiment treatment effect, and a third experience a 50 percent change or greater in either direction. The distributional effects change as well; land-poor households gain more at scale than from local interventions, while land-rich households gain more from local interventions.

Bobba, Frisancho, and Pariguana (2023) also study an intervention at whose positive, significant results are maintained at simulated scale with some heterogeneity in magnitude. Providing students with disadvantaged backgrounds information about their academic performance before filling out a high school preference form contributed to placing applicants in schools that better matched their skills. Better matches allowed students to graduate from high school on time and at a higher rate. Those who received the feedback completed high school on time at a 13 percent higher rate than those who did not in the first implemented study, while the completion rate in the simulation is between 2 and 9 percent. The authors document that general equilibrium effects, such as displacement of spots in schools as everyone changes their choices, likely offset the effectiveness of the policy suggested by the original experiment.

3.4 General challenges and possibilities at scale

Dff

In addition to the challenges to scaling documented by the literature, basic considerations such as sample selection must factor into any return on investment calculation of scaling an innovation. For example,

Allcott (2015) show that site selection bias is a key consideration when scaling up a local intervention. Utility companies in more environmentally-conscious areas were more likely to adopt an energy conservation program. When the program expanded, efficacy dropped.

On a broader level, Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, Shotland, and Walton (2017) provide a comprehensive overview of challenges in drawing conclusions about policies at scale based on results from experimental interventions. Specifically, they identify and discuss six major challenges, along with explanations on how to overcome them. These challenges include market equilibrium effects, spillovers, political reactions, context dependence, randomization or site-selection bias, and piloting bias.

Overall, expecting impact assessment results to scale proportionally is unrealistic. While the majority of the studies discussed in this section maintain positive and significant effects at scale, the magnitude of the effect typically decreases and the distributional effects shift. Attaching a scaled intervention to existing infrastructure seems to be the best strategy for continued positive impacts at scale.

4 External Validity Evidence

4.1 Introduction

External validity captures the extent to which inferences drawn from a given study's sample apply to a broader population or other target populations. Achieving robust external validity is particularly challenging because high internal validity—often obtained in controlled experimental settings—does not automatically guarantee that the underlying causal mechanisms will persist under diverse real-world circumstances. To address these challenges, recent methodological innovations have been developed to assess and enhance the generalizability of findings.

4.2 Methodological approaches to conceptualize external validity

Methodological advances have led to the development of structured frameworks and innovative techniques to assess the conditions under which causal inferences can be extended beyond a study sample. Egami and Hartman (2023) dissected external validity into four dimensions: population validity (X-validity), treatment validity (T-validity), outcome validity (Y-validity), and context validity (C-validity). This decomposition provided a structured approach to assessing when and how experimental results might generalize beyond their original contexts. Similarly, Findley, Kikuta, and Denly (2021) developed a multi-dimensional framework called M–STOUT, building upon previous literature to evaluate the external validity of studies. The M–STOUT framework encompasses mechanisms, settings, treatments, outcomes, units, and time – emphasizing that robust external validity hinges on the extent to which these dimensions are accounted for.

Bo and Galiani (2021) proposed a method that employed 1-to-1 matching to construct treated–control pairs and then generated reweighting vectors that are uniformly distributed over all possible such vectors. They also differentiated between a global measure that assessed overall external validity with respect to an overarching population and a local measure that captured how sensitive the treatment effect is to slight changes in the covariate distribution.

Kowalski (2023) discussed how treatment effects vary with selection into treatment can inform external validity. The additional information about heterogeneous subgroups from noncompliance (always takers, compliers, and never takers) can be informative about the range and distribution of treatment effects. If

the local average treatment effect differed significantly from what would be expected for always or never takers, then the treatment effect may not be globally generalized across all policies.

4.3 Meta-analyses of external validity using published studies

Peters, Langbein, and Roberts (2018) systematically reviewed all the policy evaluations based on RCTs published in top economic journals between 2009 and 2014. Only 46 percent of all papers discussed how effects might change in the longer term and whether some sort of adjustments might occur. 65 percent of the reviewed papers examined impacts less than two years after the randomized treatment. Potential changes in treatment effects in case the intervention is scaled were hardly discussed.

Kool, Andersson, and Giller (2020) documented 30 years of papers on the topic of on-farm experiments in Africa. While many on-farm experiments showed statistically significant treatment effects, the magnitude and even the direction of these effects could vary substantially when the experiment is repeated in different sites or seasons. Key impact factors included differences in soil fertility, climatic variability and management practices among farmers.

4.4 Empirical findings

Hotz, Imbens, and Mortimer (2005) found that models trained in one location can predict treatment effects in others only when rich pre-training data is available, underscoring the limits of extrapolation. The differences in outcomes could largely be removed by adjusting for pre-training earnings and other individual characteristics.

Rosenzweig and Udry (2020) highlighted that treatment effects fluctuated with economic and environmental changes, questioning the stability of external predictions. When treatment effects depend on locationspecific factors, RCT estimates from one setting do not generalize well to other locations with different distributions of shocks.

4.5 Lessons learned

Similarity and homogeneity between the experimental sample and the target population are essential for effective extrapolation. Degtiar and Rose (2023) discuss a set of advanced statistical techniques – matching and weighting, outcome regression, as well as methods that use both outcome and propensity regressions – that can be used to generalize and transport findings beyond study samples. Traditional methods, such as subgroup analysis, reweighting techniques, and tests for heterogeneity, only partially resolve the issue. More effort must be directed to the design stage of experiments, where researchers can preemptively incorporate features that enhance generalizability.

External validity is also sensitive to time variations; treatment effects may respond differently to sudden shocks or long-term economic shifts, necessitating experiments conducted over longer periods to properly test this elasticity.

Moreover, when extrapolating models to new settings, the availability of comprehensive pre-existing data and its similarity to the original dataset are crucial to ensure that the estimates remain valid.

5 Discussion

This note synthesizes findings from the literature on scaling and external validity, focusing on how impact estimates change when interventions are expanded across different dimensions. It emphasizes empirical evidence from government programs, simulations, and methodological advancements aimed at improving generalizability.

Scaling up interventions presents a range of challenges that can alter the effectiveness observed in initial studies. The literature highlights that impacts often diminish at scale due to personnel quality constraints, changes in equilibrium conditions, and differences in implementation fidelity. Government-sponsored programs, for instance, have shown mixed results, with some interventions maintaining positive effects while others fail due to political or structural factors.

External validity is similarly constrained by spatial, temporal, and demographic differences between the study sample and the target population. Studies show that treatment effects vary significantly based on local economic conditions, institutional structures, and participant characteristics. These findings suggest that rigorous methodologies, such as reweighting techniques and subgroup analysis, are necessary to assess generalizability before applying results to new contexts.

The application of AI and machine learning in impact evaluation holds promise for improving predictions of treatment effects across diverse settings. Galiani and Quistorff (2024) explore machine learning-based methods for assessing external validity, particularly reweighting techniques that use predictive algorithms to enhance generalizability. Their study demonstrates that machine learning can systematically identify covariate distributions that maximize similarity between the experimental sample and the target population, reducing bias in extrapolating treatment effects. This approach refines traditional external validity tests by allowing for more flexible, data-driven adjustments in estimating treatment impacts.

However, machine learning has fundamental limitations when applied to questions of causal inference. Unlike econometric methods designed for causal identification, such as instrumental variables or randomized controlled trials, machine learning primarily excels in prediction rather than causal discovery. The algorithms optimize for minimizing prediction error but do not establish counterfactual reasoning—meaning they cannot determine whether a treatment causes an observed outcome or if the correlation is driven by confounding variables.

Limited access to sufficient and relevant data often impedes impact assessments. Data collected from small-scale interventions or specific contexts may not provide valid predictions when scaled up or applied to other target populations due to several factors such as spillover effects, selection bias, differences in settings, measurements, and time.

Satellite remote-sensing data is a promising source to fill this gap, as it provides extensive information through wide geographic coverage, high spatial resolution, and access to otherwise inaccessible information at a comparably low cost (Donaldson and Storeygard, 2016; Wuepper, Oluoch, and Hadi, 2024; Jain, 2020). It has become a powerful tool to predict relevant indicators across many fields including agricultural yield estimation, natural resource planning, pollution monitoring, economic development analysis (using nighttime lights data), climate studies, and real-time conflict monitoring (BenYishay, Runfola, Trichler, Dolan, Goodman, Parks, Tanner, Heuser, Batra, and Anand, 2017; Donaldson and Storeygard, 2016).

Related to agriculture, studies use remote-sensing data to predict climatic and weather conditions, yield variability and productivity, and intervention coverage, particularly in contexts where traditional data sources are limited or unavailable. For example, Harari and Ferrara (2018) use sub-national level gridded weather data – Standardized Precipitation-Evapotranspiration Index (SPEI) – to examine how negative climatic shocks during crop-growing season affect conflict incidence in Africa. Using remotely sensed data in this context is particularly valuable given the sparse distribution of meteorological stations across the region. Michler, Rafi, Giezendanner, Josephson, Pede, and Tellman (2024) investigate the impact of adopting stresstolerant rice varieties (STRVs) on rice yields in Bangladesh, utilizing remotely sensed earth observation data on rice production and flooding to address the lack of traditional economic data. Burke and Lobell (2017) assess the agreement between satellite imagery and traditional field survey-based maize yield estimates, demonstrating that satellite-based predictions can achieve accuracy

comparable to survey-based estimates, particularly when precise field measurements are available and field sizes are larger, suggesting substantial potential for rapidly generating reliable productivity datasets. Remote-sensing data – integrated with GIS and GPS technologies – is also increasingly used in precision agriculture to enhance productivity and reduce environmental impact (Liaghat, Balasundram et al., 2010; Brisco, Brown, Hirose, McNairn, and Staenz, 1998).

Beyond their direct use in economic analysis (as outcome or explanatory variables), remote-sensing data provides a valuable source of exogenous variation for identifying causal effects. Duflo and Pande (2007) use instruments constructed using elevation and gradient to estimate the causal impact of large-scale hydropower and irrigation dams on poverty and agricultural productivity. Another example is a study by Rosenthal and Strange (2008) that examines the effect of agglomeration economies and proximity on productivity. They leverage geologic features (i.e., type of bedrocks, seismic and landslip hazard) as instruments, arguing that instruments are correlated with density through height of buildings but have no direct effect on wages in non-agricultural sectors. More examples of the use of satellite remote-sensing data for impact evaluation can be found in Donaldson and Storeygard (2016), Jain (2020), Pelletier, Maue, Karasalo, Jack, and Barros (2023) and Wuepper et al. (2024).

Despite its significant potential for impact assessment, the satellite remote-sensing method presents several challenges, both unique to these data and common in traditional data sources. Donaldson and Storeygard (2016) and Jain (2020) discuss potential issues, such as data complexity, substantial spatial dependence, and measurement errors, all of which need to be carefully addressed to obtain valid estimates.

References

Allcott, H. (2015): "Site Selection Bias in Program Evaluation *," The Quarterly Journal of Economics, 130, 1117–1165.

Attanasio, O., H. Baker-Henningham, R. Bernal, C. Meghir, D. Pineda, and M. Rubio-Codina (2018): "Early Stimulation and Nutrition: The Impacts of a Scalable Intervention,".

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017): "From proof of concept to scalable policies: Challenges and solutions, with an application," Journal of Economic Perspectives, 31, 73–102.

BenYishay, A., D. Runfola, R. Trichler, C. Dolan, S. Goodman, B. Parks, J. Tanner, S. Heuser, G. Batra, and A. Anand (2017): "A primer on geospatial impact evaluation methods, tools, and applications," Williamsburg, VA: AidData at William & Mary Working Paper, 44.

Bergquist, L. F., B. Faber, T. Fally, M. Hoelzlein, E. Miguel, and A. Rodr´ıguez-Clare (2022): "Scaling Agricultural Policy Interventions," NBER Working Paper.

Bo, H. and S. Galiani (2021): "Assessing external validity," Research in economics, 75, 274–285.

Bobba, M., V. Frisancho, and M. Pariguana (2023): "Perceived Ability and School Choices: Experimental Evidence and Scale-Up Effects," SSRN Electronic Journal.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2018): "Experimental evidence on scaling up education reforms in Kenya," Journal of Public Economics, 168, 1–20.

Bos, J. M., A. S. Shonchoy, S. Ravindran, and A. Khan (2024): "Early childhood human capital formation at scale," Journal of Public Economics, 231, 105046.

Brisco, B., R. Brown, T. Hirose, H. McNairn, and K. Staenz (1998): "Precision agriculture and the role of remote sensing: a review," Canadian Journal of Remote Sensing, 24, 315–327.

Burke, M. and D. B. Lobell (2017): "Satellite-based assessment of yield variation and its determinants in smallholder African systems," Proceedings of the National Academy of Sciences, 114, 2189–2194.

Degtiar, I. and S. Rose (2023): "A review of generalizability and transportability," Annual Review of Statistics and Its Application, 10, 501–524.

DellaVigna, S. and E. Linos (2020): "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," NBER Working Paper.

Donaldson, D. and A. Storeygard (2016): "The view from above: Applications of satellite data in economics," Journal of Economic Perspectives, 30, 171–198.

Duflo, E. and R. Pande (2007): "Dams," The Quarterly Journal of Economics, 122, 601–646.

Egami, N. and E. Hartman (2023): "Elements of external validity: Framework, design, and analysis," American Political Science Review, 117, 1070–1088.

Findley, M. G., K. Kikuta, and M. Denly (2021): "External validity," Annual review of political science, 24, 365–393.

Galiani, S. and B. Quistorff (2024): "Assessing external validity in practice," Research in Economics, 78, 100964.

Grantham-McGregor, S. M., C. A. Powell, S. P. Walker, and J. H. Himes (1991): "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study," The Lancet, 338, 1–5.

Harari, M. and E. L. Ferrara (2018): "Conflict, climate, and cells: a disaggregated analysis," Review of Economics and Statistics, 100, 594–608.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the efficacy of future training programs using past experiences at other locations," Journal of econometrics, 125, 241–270.

Jain, M. (2020): "The benefits and pitfalls of using satellite data for causal inference," Review of Environmental Economics and Policy.

Jepsen, C. and S. Rivkin (2007): "Class Size Reduction and Student Achievement," Journal of Human Resources.

Kool, H., J. A. Andersson, and K. E. Giller (2020): "Reproducibility and external validity of on-farm experimental research in Africa," Experimental Agriculture, 56, 587–607.

Kowalski, A. E. (2023): "How to examine external validity within an experiment," Journal of Economics & Management Strategy, 32, 491–509.

Liaghat, S., S. K. Balasundram, et al. (2010): "A review: The role of remote sensing in precision agriculture," American journal of agricultural and biological sciences, 5, 50–55.

Michler, J. D., D. A. A. Rafi, J. Giezendanner, A. Josephson, V. O. Pede, and E. Tellman (2024): "Impact Evaluations in Data Poor Settings: The Case of Stress-Tolerant Rice Varieties in Bangladesh," arXiv preprint arXiv:2409.02201.

Mitchell, H., A. M. Mobarak, K. Naguib, M. E. Reimao, and A. Shenoy (2023): "Delegation Risk and Implementation at Scale: Evidence from a Migration Loan Program in Bangladesh,".

Pelletier, J., C. Maue, M. Karasalo, K. Jack, and J. Barros (2023): "Remote sensing for impact evaluation of agriculture and natural resource management research: Guidelines for use in One CGIAR," Rome: Standing Panel on Impact Assessment (SPIA).

Peters, J., J. Langbein, and G. Roberts (2018): "Generalization in the tropics–development policy, randomized controlled trials, and external validity," The World Bank Research Observer, 33, 34–64.

Rosenthal, S. S. and W. C. Strange (2008): "The attenuation of human capital spillovers," Journal of Urban Economics, 64, 373–389.

Rosenzweig, M. R. and C. Udry (2020): "External validity in a stochastic world: Evidence from lowincome countries," The Review of Economic Studies, 87, 343–381.

Wuepper, D., W. A. Oluoch, and H. Hadi (2024): "Satellite Data in Agricultural and Environmental Economics: Theory and Practice," Agricultural Economics, e70006.

Photo: 2016/CIAT Georgina Smith

CGIAR Advisory Services – SPIA

Via di San Domenico 1, 00153 Rome, Italy

Email: spia@cgiar.org

URL: https://cas.cgiar.org/spia