Part 4

EVALUATING INCLUSIVE VALUE-CHAIN DEVELOPMENT[1]

Maximo Torero

Summary

This chapter introduces the various qualitative and quantitative methods used to conduct value-chain program impact evaluations that will be discussed in the following four chapters. We provide a brief overview of each method, as well as its benefits and limitations, and the scenarios in which it should and should not be used. While each of these methods has its uses, significant research remains to be done to ensure that impact evaluations of value-chain interventions truly capture program effects and take into account the challenges faced when trying to scale up successful programs in different locations and across different populations.

Competitive and efficient markets are key to successful economic growth, and well-functioning value chains are in turn key to successful markets. However, constraints such as limited market power, high transaction costs, poor incentives, variable risk, and a lack of access to credit can hinder the development of high-value agricultural markets, as well as markets for staple crops. This introduction discusses how interventions designed to establish more inclusive value chains for smallholders, and thus more successful markets, should be evaluated so that their impacts, costs, and benefits can be better understood.

Traditional methodologies to assess the performance and impact of value chains have focused on techniques such as participatory data collection, case studies, or different mechanisms to collect data for point price estimates or identification of inefficiencies across the value chain. While these methods serve the needs of commercial actors, they do not identify an intervention's welfare benefits, nor do they provide measures of the performance of the

1 The authors thank PIM for supporting innovative work on impact evaluation for inclusive value-chain development and on the use of quantitative tools to measure gender differences within value chains, as part of their cross-cutting gender program.

whole value chain. We provide concrete examples of different methods that can be used to fill these two gaps.

In essence, impact evaluation will accumulate credible knowledge of what works and what does not work. The overarching goal behind impact evaluation is to maximize the impact of development projects in reducing global poverty by generating information which will help: (1) to improve the design of projects based on experience; (2) to improve accountability, by clearly identifying the causal links from intervention to impact; (3) to identify successful projects to be scaled up, and (4) to allocate resources across programs by better understanding what works well, and how and which interventions are more cost-effective than others.

Finally, it is not feasible to conduct impact evaluations for all interventions. We need to build a strong evidence base for all sectors in a variety of contexts to provide guidance for policymakers and practitioners. Some examples of the types of value-chain intervention for which impact evaluation would be useful are (1) innovative schemes to upgrade value chains; (2) pilot programs that are due to be substantially scaled up; (3) interventions for which there is scant solid evidence of impact in the given context; and (4) when there is a clear need to prioritize projects based on cost-effectiveness.

What We Know about Impact Evaluation for Value-Chain Interventions

Impact evaluations measure the change in a development outcome that is attributable to a defined intervention; they are based on models of cause and effect, and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. Impact evaluations are structured around one question: What is the impact (or causal effect) of a program on an outcome of interest? The same general principles apply to impact evaluations of value chains and innovation platforms. Evaluation assesses whether the program has affected the key indicators of interest, such as poverty and nutrition, among a sample of project beneficiaries and across other dimensions of interest, such as gender.

There are different designs and methods of impact evaluation. Qualitative methods are normally used to understand the knowledge, attitudes, priorities, preferences, and perceptions of target beneficiaries and other stakeholders. These methods include, among other things, the organization of focus groups, informal interviews, semistructured interviews, and structured interviews (for further details see Lawrence 1999; Garbarino et al. 2009; Chung 2000a,b). In

addition, these methods are also useful to understand the mechanisms behind impacts and the channels through which observed effects emerge. The idea is that important information about perceptions, attitudes toward the program, incentives to participate, and the program's unexpected indirect effects on household or community dynamics may be missed by the use of purely quantitative methods. Qualitative methods are particularly useful for acquiring a more in-depth understanding of the factors influencing a program's operations or impact. Several examples are presented in parts 2 and 3 of this book. For example, the 5Capitals method (Donovan and Stoian 2012) spells out why impact evaluation is different in the context of value-chain development (VCD) and provides an example of how it can respond to the needs of nongovernmental organizations (NGOs) and value-chain actors for learning what works and what does not work for achieving inclusive VCD. There is, however, a trade-off between depth and breadth, and smaller sample sizes in qualitative studies mean that findings are rarely statistically representative of a broad population. Quantitative and qualitative evaluation methods compensate for each other's weaknesses, and each approach provides more value when used in a mixed-method design, providing information and conclusions that are more coherent, reliable, and useful than those from single-method studies.

On the quantitative side, there are experimental and quasi-experimental methods (for a detailed review of all methods see Khandker et al. 2010; Gertler et al. 2011) A fully experimental approach takes a subsample of the population of interest and randomly assigns them as participants in the program (the so-called treatment group); a second subsample is randomly assigned to the so-called control group, which does not participate in the program. The control group provides a proper counterfactual by showing the conditions for the treatment group had they not participated in the program, thus allowing for a comparison that identifies the impact of the program. With a sufficiently large sample, this type of design relies on the correct implementation of the randomization and on the full exclusion of the control group from the program (when a control group is not properly excluded, this is called contamination).

A quasi-experimental approach may be used when it is not possible to conduct randomized evaluations. In such approaches, instead of creating treatment and control groups by random assignment ex ante (that is, prior to the beginning of the program), these groups are created ex post (that is, once the program has begun or even after it has ended). This is done by using observed sociocultural, economic, ecological, and geographical characteristics to ensure that the comparison groups are sufficiently similar, at least in observable

characteristics. In this way, it can be argued that any observed impact is due to the program as opposed to other confounding factors. Ex-post design is typically used when ex-ante randomization is not possible—for example, if the program has already begun or if ethical or targeting considerations rule out such randomization. A nonexperimental method may be used to generate a control group; this would involve the comparison of program beneficiaries and nonbeneficiaries who had similar observable characteristics before the project was implemented.

Finally, in a nonexperimental evaluation, a program is nonrandomly established across units (individuals, households, villages, and so on) to identify an appropriate counterfactual. The various nonexperimental methods can be classified into two groups. The first group assumes that the unobservable characteristics of the program's beneficiaries and control group participants have nothing to do with the individuals' decisions to participate in the program. This is also known as conditional exogeneity of program placement—a strong assumption. Such methods include single-difference methods and double-difference methods. The second group is comprised of matching (including propensity-score matching, PSM) methods, discontinuity design methods, and instrumental variables; these methods do not make the exogeneity assumption, but rather address the possibility that, even after controlling for observable characteristics, unobservable characteristics may still make participation nonrandom. As a result, these methods evaluate the impact of interventions by comparing the outcomes among participants to the outcomes among comparable nonparticipants, but without randomization of participation. If both groups are exposed to similar other external events, then they allow the analyst to disentangle the effect of the intervention from the effect of all other *confounding factors*. A second class of difficulties arises when the project is purposively targeted at particular classes of beneficiaries, leading to an external selection bias. Assume for instance that an intervention in the value chain is targeted at the neediest households. In this case, comparing the poverty rate between beneficiaries and nonbeneficiaries after the project may wrongly conclude that the overall impact is zero or negative. In such cases, a more valid control group would be households that were similar to the beneficiaries at the start of the intervention. One strategy may be to compare the changes—instead of the level—of a given indicator (what we refer to as double-difference methods) between the group of beneficiaries and the control group. Assuming that the change in the indicator in the control group is a good representation of what the change in the indicator would have been among the beneficiaries, this "difference in differences" estimate may provide a valid way

to neutralize the external selection bias among observables and unobservable characteristics that are fixed over time, and hence provide an unbiased assessment of the program's effect.

In other cases, however, the confounding factor will affect the beneficiaries and the control group differently—for instance, if one would like to assess the effect of microcredit targeted at the poorest households in an area. Assume that the program occurs at a time of relatively high economic growth or weather conditions from which all households in the targeted area (rich and poor) benefit. It is likely that the economic growth will also contribute to the improvement of the income among the poor program beneficiaries, while the effect will be limited on the richer households. In such a case, a "difference in differences" measure between the richer (control) and the poorer (beneficiaries) groups will tend to overstate the effect of the program on income generation for the poor. A valid control group is one that provides a valid representation of what the average poverty level among program participants would have been without the program. Several methods may be used to generate such control groups. For instance, if the program selection criteria are known, information may be collected on nonbeneficiaries who also satisfied the selection criteria but were not included in the program for reasons independent of the outcome of interest.

A third type of bias may, however, occur when the selection process is not fully observable. Such is the case, for instance, when not all targeted households decide to benefit from the program, leading to self-selection bias. The problem of biases linked to unobservable characteristics may be resolved by "natural experiments." Such methods rely on the availability of some variable(s) that help predict participation in the program but are not related to the outcome variable (for example, income). Such methods include instrumental variables approaches, regression discontinuity designs, pipeline comparisons, and others as previously mentioned.

The following four chapters detail several distinct approaches to conducting value chain-intervention impact evaluations. The authors of Chapter 11 (Saenger et al.) implemented a randomized controlled trial and field experiment in Vietnam to improve dairy farmers' quality measurements. Chapter 12 (Cavatassi et al.) examines the Plataformas program in Ecuador using quasi-experimental methods. Chapter 13 (Horton et al.) analyzes the experience of Participatory Market Chain Analysis (PMCA) using qualitative methods in several case studies. Finally, Chapter 14 (Madrigal and Torero) provides several quantitative tools and metrics from the labor economics and discrimination literature, and gives examples of how these could be applied

in a value-chain context. All of these methods seek to connect smallholders and other marginalized groups to high-value markets. The approaches provide complementary views of the value chain and of methods to improve both the rigor and the nuance of impact evaluations for value-chain interventions.

Chapter 11 (Saenger et al.) provides a perfect example of a randomized controlled impact evaluation. The authors conducted a randomized controlled trial and field experiment with dairy farmers and a milk-processing company in Vietnam. Their approach, designed ex ante, is a theoretically ideal approach to constructing a valid counterfactual and to ensuring that there is no selection bias, given that the farmers are randomly assigned to treatment (beneficiaries) and control groups. This randomization ensures that all farmers have the same chance of participating in the program and that the distribution of the two groups' characteristics (both observed and unobserved) are statistically indistinguishable. The authors tested whether the quality-control procedures used by the processing company were leading farmers to underinvest. The risk on the farmers' part came from the possibility that the company would manipulate the process and say that the milk delivered was of low quality and therefore deserved a lower price. By introducing vouchers for third-party quality measurement, the program improved the company's credibility with the farmers. With this increased trust, the farmers then had more incentive to invest in techniques to improve milk quality and increase revenue. This chapter is unique in that it focuses on the mechanisms and incentives for different value-chain actors to contract with one another. The authors' proposed contract-farming designs make both parties better off, rather than trying to cut out the intermediary or encourage smallholders to take over other capacities in the value chain.

Although the intervention reported in Chapter 11 (Saenger et al.) affected the whole milk-production value chain, there were some specific characteristics of the intervention that enabled the use of the randomization procedure. First, the intervention was directly targeted to milk producers, which made it simpler to randomize; second, it was one single intervention rather than a package of interventions, which is normally the case with innovation platforms, as in Chapter 12 (Cavatassi et al.), or with participatory approaches, as in Chapter 13 (Horton et al.). Therefore, it is important to stress this given there are in general few value-chain impact studies that use experimental and randomized controlled trial (RCT) methods because value-chain development usually involves many different partners (public and private-sector institutions) and often complex interventions, which might make RCT and experimental approaches particularly difficult and in many cases not feasible to implement.

The authors of Chapter 12 (Cavatassi et al.) performed an ex-post evaluation using econometric techniques common in impact evaluations. They assessed whether participation in Ecuador's Plataformas program, which establishes alliances between small-scale farmers and a range of agricultural support-service providers, had any effect on income. The chapter finds that the program had a positive impact on yields, prices, and gross margins. The authors conducted baseline household and community surveys in two Ecuadorean provinces and then identified treatment communities. Using data from the most recent census, they constructed a counterfactual control group with similar geographical, agroecological, and sociodemographic characteristics to the treatment communities. They then used PSM to identify which control communities were most similar to each treatment community. In addition to creating control communities, they also factored in households in treatment communities that did not participate in the program. The PSM procedure allowed the authors to select a control community that was very similar to each treatment community in all observable aspects except for the treatment status, thus providing a proper counterfactual for each treatment community.

One of the major concerns regarding the PSM approach is that there might be other observable and unobservable differences that could explain a community's selection into the treatment group. To minimize this problem, the research (Chapter 12) implemented an instrumental variable (IV) approach to control for observable and unobservable differences in the control and treatment groups. The IV technique identifies a factor that predicts participation in a program but that does not influence the program's outcomes of interest. This factor is then used to simulate which participants would have been in the treatment group and which would have been in the control group had the project been based on that factor. The difference in outcomes between these simulated treatment and control groups constitutes the project's impact.

For an IV estimation approach to be viable, as it is in Chapter 12, the instrumental variables used must be strong predictors of whether or not a participant will receive the treatment; however, we must also be sure that the variables themselves will not determine the program's outcome. It will likely be difficult to identify variables that meet both these criteria since the factors determining whether a potential beneficiary wants to participate in the program are likely to also be factors that will affect the outcome of interest. IV methods estimate a program's impact on people who participate in the program *because* of the program's instruments. It is thus important to know which precise groups will be affected by those instruments, and whether these

groups are of interest for the program. IV estimation does not easily allow for generalizing to other groups.

The IV technique is useful in determining local average treatment effects (LATE) rather than average treatment effects (ATE), which are usually the effects examined in impact evaluations. The IV estimator is a weighted average of the LATE of different subpopulations; the subpopulations that are more responsive to the program's instruments carry a higher weight in the final IV estimate. These issues could severely bias the results or the conclusions that can be drawn from them if the subpopulation is not correctly identified; thus great caution is required when interpreting the results of the IV technique.

Finally, one important thing that the authors of Chapter 12 (Cavatassi et al.) did that can help strengthen the interpretation of a program's results is an assessment of the program's impact pathways. The authors analyzed the ways in which farmers might benefit from the program and found that the program significantly increased yields and gross margins for the treatment communities.

Chapter 13 (Horton et al.) tries to assess the impact of a PMCA. The chapter provides a clear example in which neither experimental nor quasi-experimental approaches could be implemented. In PMCA, practitioners gather various market-chain actors together to brainstorm ideas for new agricultural products and better ways to market existing crops. PMCA was created both to link smallholders to markets through innovation, and to evaluate participatory interventions. Chapter 13 (Horton et al.) evaluates eight PMCA interventions, four of which they exclude from in-depth analysis because of significant departures from the PMCA protocol. Attempts were made to conduct an impact evaluation using quasi-experimental methods; however, delays in conducting the baseline surveys prevented the data from being useful for evaluation purposes. Instead, the authors implemented a case-study evaluation following the methodology of Yin (2009). Drawing on the definitions of Chen (2005), they stressed the importance of the action model, "a systematic plan for organizing resources, staff, and relationships in order to deliver the intervention faithfully." They also identified the program's change model, which is the "broader conceptual framework that links the intervention's activities and outputs to the expected outcomes and impacts and explains how and why the intervention is expected to lead to the desired changes." Their evaluation is based on the "fidelity of implementation," which "refers to the extent to which a program's implementation is consistent with its action model." They discovered that PMCA needs to be adapted

to local country and market contexts, while still remaining consistent. The economic benefits of the four PMCA interventions were small, but by identifying both the action and the change models, the authors were able to distinguish creative adaptations to the program from lapses in implementation.

The approach followed by Chapter 13 (Horton et al.) is extremely useful in understanding the potential effects of PMCA, but it doesn't allow us to isolate whether the observed changes can be truly attributed to the intervention. Clearly, it would have been better to combine this method with an experimental or quasi-experimental approach.

Finally, Chapter 14 (Madrigal and Torero) sheds light on an important issue that is not captured by any of the previous approaches: Most value-chain impact evaluations fail to look at effects disaggregated by gender. This is an important oversight, because in most value chains men and women play different roles, and failure to account for gender in a randomized controlled trial, quasi-experimental, or participatory intervention may significantly alter the results of these studies. To resolve this gap in the literature, the authors focus on several tools and metrics to incorporate gender in value-chain impact evaluations. The Oaxaca Blinder decomposition analysis allows for proper measurement of wage gaps between men and women by controlling for other observable variables; the Duncan Index and Access to Work Equality Index measure occupational segregation and differential access to employment. Finally, time-use analysis can provide insights into how to improve labor opportunities for both men and women. Provided that gender-disaggregated survey data are collected, these tools can all be applied to value-chain interventions and analyses at low cost.

Gaps that Need to be Addressed

Although the four chapters in this section provide clear examples of ways in which value-chain improvements can be evaluated, there are still some important issues and gaps that need to be addressed in future research. First, even where RCTs are used, as in Chapter 11, there are still concerns on RCTs that need to be looked at, and specific implications as mentioned by Barret and Carter (2010). Second, most of the value-chain improvements being developed include interventions that affect different nodes of the value chain. This creates enormous complexity when trying to assess the impact of a program experimentally or quasi-experimentally. For example, if the unit of treatment is a whole value chain, there will need to be sufficient treatment and control value chains of the same commodity to have enough statistical power to assess

the true impacts of the intervention. This would require an appropriate sampling strategy representative of each node: input suppliers, producers, traders, wholesalers, and retailers. However, it will sometimes not be feasible to find the number of value chains needed in the same geographical area. Similarly, the potential for spillovers of the effects in one node of the value chain to others is important, and methods need to identify ways to control for this.

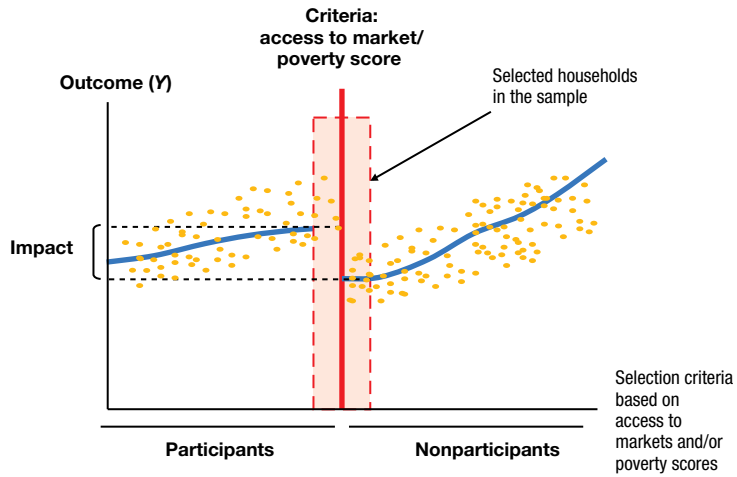
An alternate method which could contribute to partially addressing this problem is a nonexperimental approach known as regression discontinuity design (RDD) (for further details, see Jacob et al. 2012; Bloom 2012; Imbens and Lemieux 2008). If small variations in a specific variable produce a discontinuous change in a person's (or value chain's) eligibility for the treatment or participation in the program, this can be used to identify the program's impact using IV estimation, even if the variable is also a direct determinant of the program's outcome. For example, there may be levels of poverty or of access to roads or technology that determine a farmer's eligibility; these may in turn produce regression discontinuities. As presented in Figure P4.1, by using such discontinuities the impact of an intervention or program can be estimated by comparing outcomes for beneficiaries who just qualify for the project on this index/score[2] with outcomes for individuals who just fail to qualify for the program given their score (the so-called control group), as determined by these characteristics. The logic behind this is that since observations around the cut-off have treatment status, that is as good as randomly assigned.

One caveat to this approach is that if the discontinuity (or cut-off range) is too big, those who did not qualify for the program may be sufficiently different from those who did in terms of their observable characteristics. As a result, the impact of the program may be estimated incorrectly.

A variation on this type of evaluation method is called "fuzzy" regression discontinuity (see Jacob and Zhu 2012). In this case, some beneficiaries have scores that place them on the nonbeneficiary side of the discontinuity. This RDD method is termed fuzzy because the cut-off is not clear or strict. When the eligibility criteria for participation are public information, the variable used to establish the treatment group could be manipulated so that a person appears eligible; clearly, this can create difficulties when estimating the effect of the program. Such manipulation would introduce nonrandom selection

2 Note that this score does not necessarily relate to the PSM procedure. The score for the RDD is a variable, either existing or constructed, that establishes a threshold above which individuals are allocated to the program and below which they are not part of the program. The propensity score is one such variable that can be used in this estimation if it is discontinued at some specific point.

FIGURE P4.1 Regression discontinuity design



Source: Authors.

around the cut-off, which would need to be addressed by randomizing the subpopulation around the fuzzy cut-off; if this is not possible, nonrandom assignment can be permitted to adjust for selection into the fuzzy interval in the final estimation. However, as long as this manipulation is not precise, the RDD remains valid.

RDDs require a large sample (and a considerable amount around the cut-off) and the fuzzy interval must be moderate to be able to provide valid and precise impact estimates.

The second issue that calls for significant innovation and research is that in all the impact-evaluation approaches, even RCTs, there needs to be a mechanism to capture heterogeneity and external validity—that is, to understand how much the results identified can be extrapolated to other areas or even other value chains of similar commodities (heterogeneous populations). In a majority of impact evaluations, it is commonly assumed that the estimated treatment effects can be generalized to the whole population or to a new location in which no experiment was conducted. However, since individuals in a new location can have different observable and unobservable characteristics, the ATE can be significantly different from the one obtained from experiments conducted in other locations. Several authors have protested against policy recommendations that they believe are based on implicit extrapolation from a small number of experiments to a wide variety of dissimilar contexts

(Deaton 2010; Pritchett and Sandefur 2013). Empirically, a growing body of work shows that identical policies have different effects among individuals with the same observed characteristics living in different contexts (for example, Allcott 2012; Attanasio, Meghir, and Szekely 2003), because unobserved differences between populations remain. Hence, we need a method that accounts for heterogeneity across locations, or we need to design an evaluation that takes this issue into account from the beginning.

For methods that account for heterogeneity, there has been some progress. Athey and Imbens (2006) generalize the standard difference-in-differences estimator and derive an estimator that can be used to extrapolate results under perfect dependence between the treated and untreated outcomes. Gechter (2014) improves on this work by developing a method for predicting the ATE in a new location under a mild restriction on the joint distribution of potential outcomes. Specifically, he derives bounds on the predicted ATEs by imposing a lower bound on the rank correlation of the potential outcomes. We can then take the case of minimal treatment effect heterogeneity (perfect rank correlation) as a benchmark to further investigate how the predicted bounds on the ATE change by allowing different levels of heterogeneity.

Finally, an alternative way to ensure a certain level of external validity is through the ex-ante design of a scaling-up mechanism that will allow a program to be replicated on the basis of results from rigorous impact evaluations. An example of this potential approach is given by Torero (2014), who essentially develops a typology of rural areas that identifies needs, opportunities, and bottlenecks at the regional level based on modeling of agricultural performance and potential using the economic concept of the production possibilities frontier, drawing on highly detailed household-level survey data and geospatial tools. Such a typology allows program targeting based not only on needs, as is the case when using poverty maps, but also on economic potential against current performance (or efficiency relative to the economic potential) and the associated needed investment gaps to improve the respective performance so that it can reach its economic potential. As a result, projects designed to resolve those gaps can be replicated in similar types within the typology. In addition, combined with appropriate project designs and impact-evaluation tools, this typology can help systematize targeting of development projects in a range of technical domains across the value chains, including financial services. However, because this approach involves an ex-ante identification of similar locations where an intervention can be successfully tested, it will require significant work before implementation.

References

Allcott, H. 2012. *Site Selection Bias in Program Evaluation*. NBER Working Paper No. 18373. Cambridge, MA: National Bureau of Economic Research.

Athey, S., and G. W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74 (2): 431–497.

Attanasio, O., C. Meghir, and M. Szekely. 2003. *Using Randomized Experiments and Structural Models for "Scaling Up": Evidence from the PROGRESA Evaluation*. IFS Working Papers WP03/05. London: Institute for Fiscal Studies.

Barrett, C. B., and M. R. Carter. 2010. "The Power and Pitfalls of Experiments in Development Economics: Some Non-random Reflections." *Applied Economic Perspectives and Policy, Agricultural and Applied Economics Association* 32 (4): 515–548.

Bloom, H. S. 2012. "Modern Regression Discontinuity Analysis." *Journal of Research on Educational Effectiveness* 5 (1): 43-82

Chen, H. 2005. *Practical Program Evaluation*. Thousand Oaks, CA, US: Sage Publications.

Chung, K. 2000a. "Qualitative Data Collection Techniques." In *Designing Household Survey Questionnaires for Developing Countries: Lessons from Fifteen Years of the Living Standards Measurement Study*, edited by P. Glewwe and M. Grosh. Oxford: Oxford University Press for World Bank.

———. 2000b. "Issues and Approaches in the Use of Integrated Methods." In *Integrating Quantitative and Qualitative Research in Development Projects*, edited by M. Bamberger. Directions in Development Series. Washington, DC: World Bank.

Deaton, A. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–455.

Donovan, J., and D. Stoian. 2012. *5Capitals: A Tool for Assessing the Poverty Impacts of Value Chain Development*. Technical Series, Technical Bulletin no. 55, Rural Enterprise Development Collection No. 7. Turrialba, Costa Rica: Tropical Agricultural Research and Higher Education Center (CATIE).

Garbarino, S., and J. Holland. 2009. *Quantitative and Qualitative Methods in Impact Evaluation and Measuring Results*. Discussion Paper. Birmingham, UK: University of Birmingham.

Gechter, M. 2014. "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India." Job Market Paper. Boston: Boston University.

Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2011. *Impact Evaluation in Practice*. Washington, DC: International Bank for Reconstruction and Development/World Bank.

Imbens, G. W., and T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–635.

Jacob, R., Pei Zhu, M.-A. Somers, and H. Bloom. 2012. *A Practical Guide to Regression Discontinuity*. New York: MDRC.

Khandker, S. R., G. B. Koolwal, and H. A. Samad. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank Publications, No. 2693. Washington, DC: World Bank.

Lawrence, M. B. 1999. "The Qualitative Method of Impact Analysis." *American Journal of Evaluation* 20 (1): 69–85.

Pritchett, L., and J. Sandefur. 2013. *Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix*. Working Paper 336. London: Center for Global Development.

Torero, M. 2014. "Targeting Investments to Link Farmers to Markets: A Framework for Capturing the Heterogeneity of Smallholder Farmers." In *New Directions for Smallholder Agriculture*, edited by P. B. R. Hazell and A. Rahman. Oxford: Oxford University Press.

Yin, R. 2009. *Case Study Research*. Thousand Oaks, CA: Sage Publications.