AgResults
INNOVATION IN RESEARCH AND DELIVERY

Evaluator's Lessons Learned Series

Issue 8, September 2022

Impact Evaluation Methods: Evaluating the Effect of Market Interventions on Targeted Beneficiaries

This brief describes why impact evaluations of market systems interventions are rare and challenging and presents examples from the AgResults portfolio of pay-for-results competitions. It then outlines methods available for conducting impact evaluations of market interventions and highlights key lessons for evaluators.

AgResults: an opportunity to evaluate the impact on smallholder farmers of pay-for-results competitions to increase private sector engagement in markets providing high-impact technologies

Market system development approaches are of interest to international development organizations as a means to achieve broad-scale, sustained impact. Yet donors face challenges in understanding how these interventions affect market systems and their participants. It can be nearly impossible to know what the *counterfactual* is: i.e., what would have happened without the investment. Further, market systems development programs often work at the level of private sector firms, so it can be difficult to study outcomes for individuals and households that the programs never interact with directly.

Our experience serving as external evaluators for AgResults highlights important lessons and opportunities for future evaluations. AgResults engages the private sector in providing high-impact technologies to smallholder farmers and promotes development of sustainable and inclusive markets for these technologies. Since its inception in 2010, the AgResults initiative has gathered evidence about the impact of its interventions on its intended beneficiaries. In this brief, we focus specifically on the evaluation of whether the market systems intervention has resulted in intended outcomes for intended beneficiaries (not on the market system itself). In AgResults, intended beneficiaries were smallholder farmers.

Despite international attention to market system development approaches and despite excellent work by groups such as the Beam Exchange, the Springfield Centre, and USAID to develop ways to evaluate market systems projects, causal impact evaluation of projects' impact on intended beneficiaries has received relatively little attention. Meanwhile, the last decade has simultaneously seen tremendous growth in the academic focus on field experiments. Field experiments, in research, are studies that use an experimental design in a natural setting, as opposed to a laboratory. In 2019 the Nobel Prize in Economics was awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer for their leadership in bringing the science of field experiments to poverty alleviation efforts. Field experiments are the gold standard for learning how an intervention impacts outcomes because experiments create treatment and control groups that differ only in their exposure to the intervention. Field experiments in market interventions are often infeasible because market impacts are far-reaching. Regardless, there are several 'next best' approaches. Here, we seek to contribute our own experience to the knowledge base, examining how field experiments and quasi-experiments can be applied in the evaluation of market system approaches.

The AgResults initiative sprang from the 2010 G20 summit, when leaders called for the development of "results-based mechanisms" to "harness the creativity and resources of the private sector" in achieving

development goals related to food security and agriculture. Two years later, the governments of Australia, Canada, the United Kingdom, and the United States, in partnership with the Bill & Melinda Gates Foundation, pledged funds to establish AgResults. AgResults is an initiative to identify, design, implement, and evaluate pay-for-results schemes intended to catalyze the development of markets for high-impact agricultural technologies that benefit smallholder farmers in developing countries. As of 2022, completed AgResults projects include efforts to develop markets for aflatoxin-compliant maize in Nigeria, improved on-farm storage devices in Kenya, and emissions-reducing rice production technologies in Vietnam. To overcome market constraints such as incomplete information and missing linkages between potential suppliers and potential customers, AgResults offers cash prizes to targeted market actors (or "competitors") that achieve pre-specified outcomes (typically sales of the beneficial technologies). The availability of the prize reduces the risk to suppliers of entering these markets. The theory is that if a prize is well designed and attractive to private sector competitors, they will find and implement creative solutions to constraints that otherwise inhibit the development of a market for the technology.

Since 2014, Abt Associates—a firm with over 50 years' experience in rigorous program evaluation with a focus on disadvantaged populations—has been engaged in evaluating the AgResults portfolio. In this brief we share our experience designing and implementing evaluations of AgResults projects, including how we have worked around some key challenges. As one of the builders of AgResults' substantial evidence base, we hope that this brief will demonstrate the value of rigorous evaluation and how to apply it in market systems contexts. In the AgResults context of the world's first payment-for-results approach to spur growth in agricultural markets, this article explores *why* causal attribution is so difficult to achieve when estimating the impact of market systems approaches, *which* alternative evaluation methods are available, and *what* lessons we've learned about evaluation methods.

In each project, AgResults used causal impact evaluations to gain new insight

Nigeria Aflasafe™	Kenya On-Farm Storage	Vietnam Emissions Reduction
Goals of the projects		
The goal of the Nigeria Aflasafe project was to catalyze development of a sustainable, smallholder-inclusive market for maize treated with Aflasafe™, a biocontrol agent that prevents aflatoxin contamination of crops. Aflatoxins are naturally occurring toxins that affect crops such as maize and are not safe for consumption at any level. In each harvest season from 2014 through 2018, AgResults offered per-unit prizes to qualifying Nigerian traders who aggregated Aflasafe-treated maize from smallholder farmers.	The Kenya prize competition aimed to create a market for improved on-farm storage devices for smallholder farmers to store staple grains, particularly maize. Improved on-farm storage was expected to improve food security by reducing post-harvest loss of maize and other food staples, reducing farmers' expenditures on purchases of staple foods once their stores were exhausted, as well as reducing the use of pesticides and aflatoxin levels. AgResults offered cash prizes to qualifying storage providers that achieved specific benchmarks in making sales of improved storage to smallholder farmers.	AgResults used a two-stage competition to promote the development and uptake of low-emissions rice cultivation technologies: it first offered prizes to private sector companies that developed rice cultivation systems that both reduced GHG emissions and increased rice yields compared to conventional rice cultivation practices. Next, AgResults offered prizes to qualifying competitors for effective dissemination of their emissions-reducing systems to farmers, as well as for their actual success in GHG emissions reduction and rice yield improvement.

Nigeria Aflasafe™	Kenya On-Farm Storage	Vietnam Emissions Reduction
What our impact evaluations revealed about projects' impact on beneficiaries		
• The Nigeria program increased farmer incomes significantly by an estimated 16% compared to what their incomes would have been in the absence of the program. Without a rigorous comparison group, the project would have reported an erroneously larger impact, because yields and revenues for the comparison group, and across the region, were 35% higher at endline compared to baseline. • The program increased smallholder families' consumption of aflatoxin-compliant maize by a statistically significant 13%. Consumption was only studied by the evaluators, not by the program.	• The Kenya program was effective in increasing uptake of improved on-farm storage devices by smallholder farmers. • The increase in revenue for adopters of the technology was small (less than US$2). The evaluation found that farmers, in large part, grew maize for household consumption rather than for sale, limiting the potential for maize revenue increases. • Storage devices did not reduce post-harvest losses because adopters of storage devices (and non-adopters similar to them) already successfully prevented in-storage losses by applying pesticide dust to their grains. The evaluation did, however, indicate that the farm families valued the storage devices because they were labor saving and because they preferred grain that had not been treated with pesticide dust.	• The Vietnam program resulted in 14% higher yields, as demonstrated by the statistically significant difference between AgResults and comparison farmers in a matched comparison analysis. The third party verifier, on the other hand, relied on simulations to estimate a counterfactual. The simulation, vulnerable to untested assumptions, suggested that the program resulted in little to no increase in yield. The evaluation, therefore, discovered yield impacts that the verifier was not able to detect. • The evaluation, thanks to a randomized controlled trial, was able to distinguish how farmers in areas participating in AgResults differed from those in areas that did not participate. The evaluation found that the AgResults areas reduced planting density by 7%, straw burning by 14%, and nitrogen use by 13% compared to common practice. • The evaluation found that participating farmers generated 10% higher net value per hectare, but it also found that this increase would not be statistically significant if competitors' discounted inputs were not considered in the calculation. Evaluators indicated a need for further investigation (within a broader follow up study) of whether companies would likely continue to discount inputs after the competition.

Why causal attribution is so difficult in market system development approaches

Causal attribution is important because it clarifies whether a program's specific activities can be credited for generating the outcomes of interest. It helps us understand whether the outcomes we associate with the program would *not* have occurred absent the program.

The ideal way to discover the impact of an intervention is to compare its effects in one setting to a setting that is otherwise identical, except for the presence of the intervention. The laboratory sciences use highly controlled settings to conduct experiments; while social sciences often use "field" experiments, which compare two settings that are less controlled but highly similar except for the presence of the intervention being tested. Field-based experiments are often not viable in the evaluation of market interventions because of "spillover"

that contaminates the comparison group, limited resources, and fear of program failure. We discuss each of these in turn.

When evaluating a market intervention, it can be difficult or impossible to define and protect a comparison group from the intervention's influence. For example, if an agricultural market intervention leads to an increase in maize yields for participating farmers, the outcomes of nonparticipating maize farmers may be affected by an increase in the supply of maize to local markets pushing down market prices. In this example, both the treatment and comparison groups are affected by the intervention, implying that the comparison group is not an effective example of what would happen in the absence of the intervention.

It is difficult to set up field experiments within market intervention programs because donors usually want to rely on the knowledge, choices, and efficiencies of the private sector; donors usually not want artificially to restrict the beneficiary population or geographic area in order to have some "treatment" farmers affected by the initiative and "control" farmers unaffected by the initiative. Such an experimental design constrains the ability of local actors to take advantage of all existing efficiencies. For example, consider a scenario where the private sector influences its customers by reaching broad audiences on radio and internet platforms: it is impossible to contain the spread of the internet announcements.

Another approach could be to offer the same market intervention in many different markets that are isolated from one another, randomizing which of the markets receive the intervention. However, this approach is not common, sometimes for budget reasons, sometimes due to a preference for tailoring interventions to specific settings.

Even with a comparison group in place, it can be very difficult, statistically, to detect differences between a comparison group and a treatment group because the majority in the intended treatment group typically will not experience the intervention. In the AgResults Vietnam and Nigeria projects, the proportion of farmers engaged by AgResults competitors was low—in Nigeria, only 13% of eligible farmers actually took up the treatment, while in Vietnam only 5% of eligible farmers did. Extending the time frame may allow the intervention to reach more farmers. Another approach is to narrow the anticipated treatment group by understanding who may be targeted by the intervention.

What alternative designs are available

Even if it does not use random assignment, an evaluation can still have a comparison group that is similar to the treatment group and is followed over time. Sample inclusion criteria, propensity scores, analysis weights, coarsened exact matching, and other analytic techniques can help the evaluator establish similar treatment and comparison groups. Comparing groups both *before* and *after* program implementation is called a "difference-in-difference" approach. This approach compares the time-trends of the participants and non-participants and produces causal estimates of the impact of the program under the assumption that whatever distinguishes the participants from the nonparticipants does not change over time (for example, unobserved characteristics such as motivation, curiosity, or propensity to take risks). Although the "differencing" reduces concern about unobserved differences between the treatment and comparison groups, it is still important that the treatment and comparison groups be similar on important characteristics likely to affect the outcome.

Another approach, particularly when comparison areas are difficult to identify, is to use a synthetic comparison group. The synthetic comparison group method is an effort to elevate a single case study design to a quantitative evaluation by creating a composite, "synthetic" comparison case from a weighted average of comparison cases. The synthetic comparison group method has been used to estimate the effect of the reunification of Germany on the German economy (Abadie et al. 2015), the effect of California's tobacco

control program on tobacco use (Abadie et al. 2010), the effect of Basque Country terrorism on the local economy (Abadie and Gardeazabel 2003), the effect of the Mariel boatlift of Cuban workers in Miami (Peri and Yasenov 2019) and others. A challenge in using a synthetic comparison group is the need to collect data from multiple comparison areas and in at least three time points. Where primary data collection efforts are required, it can be prohibitively expensive (not to mention unpalatable) for donors to collect data in areas where they are not operating programs. In the context of the AgResults projects described above, synthetic comparison group approaches were not feasible because, as is often the case in developing country settings, primary data collection from multiple areas and multiple time points would have been too expensive and there were no satisfactory secondary data sources that could have been used instead.

Yet another approach is to use a "changes in changes" method that examines several time points *before* the intervention and several time points after an intervention, for both the treatment and comparison group (Athey and Imbens 2006, Xu 2007). Whereas difference-in-difference requires that the evaluator assume the pace of change is constant over time, the changes-in-changes approach does not. In market systems approaches, most donors anticipate that the pace of growth is not constant over time. For example, consider a market systems approach to the promotion of menstrual cups for school-age girls in rural East Africa where menstruation is known to keep girls 'sick' at home from school. Without intervention, the proportion of girls using menstrual cups would increase over time, albeit very slowly at first, then increasing quickly, and then plateauing, likely following an "S" curve. With the intervention, the use of menstrual cups would grow faster, following an "S" curve that is more steeply shaped, and which may plateau at higher level.[1] The challenge in this approach is that it requires several observation periods *before* as well as after the intervention: where original data collection is required, donors are often not in a position to put program implementation on hold in order to first collect several years' of "baseline" data. This approach is most feasible in settings where regularly updated secondary data sources are available.

The rigorous options above require pre-intervention data. Even when they are not feasible, it is still possible to compare the outcomes participants to those of nonparticipants. There's always a chance that participants and nonparticipants have different outcomes not because the participants experienced the intervention, but because of the underlying characteristics that drove their decision to participate (wealth, asset ownership, anticipated future economic gains, risk-seeking habits, etc.) Thus, similar to the methods mentioned above, the evaluator can employ techniques such as propensity scores, coarsened exact matching, and other methods to ensure that the studied non-participants are similar to the participants. Without baseline measures of outcome variables, these approaches are not scientifically rigorous enough to withstand the scrutiny of most academic, peer-reviewed journals. However, for program managers, donors, and other stakeholders who would otherwise only have monitoring data, these nonexperimental approaches add value by providing at least a plausible range of the true impact of the program. The table below briefly describes the evaluation methods we used in the Nigeria, Kenya, and Vietnam AgResults projects, and the types of baseline and follow-up data we used for those methods.

[1] Hall and Khan (2003) explain that the graph of the proportion of households adopting a new technology over time is shaped like an "S" when (a) the cost of that technology is constant or never increasing, (b) the preferences/tastes of consumers is normally distributed (bell-curve shape), and (c) consumers adopt the technology as soon as they begin to value it more than its cost. Strang and Soule (1998) show that technology adoption also follows an "S" curve when initially there is imperfect information about the technology and neighbors tell neighbors about the technology, improving their neighbors' information about the technology over time.

AgResults Impact Evaluation Methods		
Nigeria Aflasafe	**Kenya On-Farm Storage**	**Vietnam Emissions Reduction**
Research question: What is the impact of AgResults on farmers who live in villages likely to be recruited by private sector companies participating in the AgResults challenge project? **Method:** Selection of treatment and comparison groups *after* intervention was under way; propensity-score weights to improve comparability of treatment and comparison groups on stable characteristics that could not have been affected by the intervention. **Data:** Endline survey of smallholder farmers. **Baseline equivalence:** Farm area, volume of typical harvest, asset ownership two years ago, number of persons working on the farm, education, others.	**Research question:** What is the impact of adopting improved on-farm storage devices on smallholder farmers? **Method:** Difference-in-difference impact evaluation of adopters and non-adopters, matched using coarsened exact matching. **Data:** Baseline and endline survey of smallholder farmers. **Baseline equivalence:** Baseline value of outcome measures (maize revenue and yield), farm area, volume of typical harvest, asset ownership, others.	**Research questions:** (a) What is the impact of AgResults on farmers' adoption of an emissions-reducing technology package? (b) What is the impact of technology adoption on rice farmers' incomes? **Method:** (a) Randomized control trial (RCT) (b) Selection of treatment and comparison groups *after* intervention was underway, regression weights to improve comparability of adopters and nonadopters on stable characteristics that could not have been affected by the intervention, as well as baseline cooperative characteristics. **Data:** Baseline cooperative surveys; (a) Endline farmer diaries from random sample of rice farmers; (b) Endline income survey of adopters and non-adopters similar to adopters. **Baseline equivalence:** Selected by agricultural cooperative leader as being serious about rice farming and close to the road, soil type, characteristics of local field drainage systems, baseline cooperative characteristics, education, others.

Seven lessons for impact evaluations of targeted beneficiaries of market system development projects

Market systems are dynamic. Large and even small shifts in the private sector's strategy can upend evaluation plans. We offer seven strategies that can increase the chance that the evaluator has the right data and a valid counterfactual to estimate the impact of a market system development project on its intended beneficiaries.

Conduct an initial qualitative assessment to identify the opportunities for impact assessment. In-depth, and typically original, background research on market actors, their business models, and their relationships reveals information critical to the evaluation design team. The evaluation team can use information about the structure and conduct of market actors to assess which evaluation designs are feasible. Initial qualitative assessment can also help the evaluation team, as well as the funder, re-assess whether the intervention's goals and theory of change hold promise.

Random assignment, if planned, must be integrated into incentives or contracts. The evaluator should try to establish a counterfactual whenever possible and protect it from spillover influence from the project. Yet a counterfactual may create challenges for the project that will want to provide private sector firms as much latitude as possible in determining where they work. Unless the evaluator and implementer can balance these competing priorities, project implementation will suffer, or the evaluation will miss an opportunity to produce high quality evidence. An experience from AgResults is instructive in this regard: in Nigeria, AgResults had a promising set-up for an RCT except that cash incentive prizes (i.e., the main intervention) were not linked to the RCT design. Thus, there was no direct incentive for the private sector actors to abide by the RCT design. By the third year of the program, the fraction of farmers reached in the treatment areas was roughly equal to the fraction of farmers reached in the control areas, resulting in near-zero contrast between the treatment and comparison groups. To avoid similar issues, the evaluator and project implementation team should work closely from the beginning, so that the counterfactual, if possible random assignment, can be integrated into the project roll-out.

Large baseline samples will ensure you find enough adopters at endline. Baseline surveys with a large number of respondents improve the likelihood of having baseline data for farmers or consumers ultimately affected by the project, even if the project does not have a large reach. Baseline surveys that cover a large geographic area minimize the risk of missing any geographic area ultimately touched by the program. The evaluator cannot know in advance where the program will operate, and as we learned the hard way, should not take too seriously the stated intentions of the presumed private sector participants before the intervention begins. In hindsight, the Nigeria baseline sample survey, with roughly 1500 respondents, was not large enough. A difference-in-difference design, as a replacement for the Nigeria RCT, was not feasible because only 257 of the baseline respondents had participated in AgResults by the time the impact evaluation took place. A sample of 257 treatment group farmers did not offer sufficient statistical power. The Kenya baseline survey was larger, at roughly 4700, and thus we were able to identify a matched comparison group of smallholder farmers who were similar to treated farmers on a wide range of baseline characteristics and use a difference-in-difference design. We are currently conducting the evaluation of the ongoing AgResults Tanzania project and are recruiting 5000 potential intervention participants to the baseline survey effort. Sample size should be determined based on conservative estimates of the percent of beneficiaries that will be reached.

Take advantage of delays in program implementation to conduct multiple baseline surveys, to enable more rigorous evaluation options. Many projects start late. Anticipating that there would be no comparison area that the evaluators could isolate from the AgResults Kenya program, the evaluation took advantage of a late roll-out to collect data in two pre-intervention periods as well as one post-intervention period, providing a short time series that made the difference-in-differences design possible. If there had been even more measurements for additional pre-intervention periods, a synthetic comparison group design or changes-in-changes evaluation design may have been feasible.

To match/weight accurately, collect data on intended beneficiary characteristics that may be related to their participation decision. All our evaluation designs called for collecting information about demographics, education, past and expected dependence non-farm income, household assets, farm assets, and connections to local agricultural organizations or market players. This information is vital for establishing comparability between treatment and comparison groups, creating weights so that the comparison group can be constructed to resemble the treatment group more closely, or using matching techniques to find a subset of the treatment and comparison groups that are most similar. In the Vietnam evaluation, we also incorporated information on risk-taking and entrepreneurial personality traits.

Consider smaller-scale data collection efforts to obtain representative descriptive statistics of the population, infer the sustainability of the market, and possibly detect spillover. In the Vietnam evaluation, we anticipated that a small percent of farmers in the areas assigned to the treatment group would participate in AgResults. To compare traditional rice cultivation practices with the those recommended by AgResults, we decided to recruit two randomly selected farmers in every community of the targeted province to keep detailed diaries about their rice cultivation over two crop cycles. We used these diaries to conduct the RCT analysis as planned, which revealed the average impact of AgResults on all farmers in the province. In addition, the detailed information from the diaries told us how much behavior change is required of farmers to switch to AgResults practices. If we had only sampled AgResults farmers and farmers who are similar to AgResults farmers, we would not have obtained a representative population-level estimate of current rice farming practices. Using the RCT design, this smaller-scale farmer diary study also provides some insight into whether AgResults influenced the farming practices of farmers geographically close to the implementation sites (i.e. spillover).

Complement the quantitative impact evaluation with an in-depth qualitative analysis. For important nuanced understanding and insight on the knowledge, attitudes, and choices of targeted beneficiaries, researchers should conduct semi-structured interviews or focus group discussions with targeted beneficiaries, local leaders, and other involved parties. Qualitative research approaches offer vital insight into the spectrum of beneficiaries' perceptions and experiences.

References

Abadie A., Diamond, A., & Hainmueller, J. (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association, 105*(490), 493-505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495-510.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, *93*(1), 113-132.

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431-497.

Capacci, S., Allais, O., Bonnet, C., & Mazzocchi, M. (2019). The impact of the French soda tax on prices and purchases. An ex post evaluation. *PloS one*, *14*(10), e0223196.

Hall, B., & Khan, B. (2003). Adoption of new technology, in Jones, Derek C. (ed.), *New Economy Handbook,* Amsterdam: Elsevier Science. [NBER Working Paper No. W9730. UC Berkeley Department of Economics Working Paper No. E03-330.]

Peri, G., & Yasenov, V. (2019). The labor market effects of a refugee wave synthetic control method meets the Mariel boatlift. *Journal of Human Resources*, *54*(2), 267-309.

Strang, D., and Soule, S. A. (1998). Diffusion in organizations and social movements. *Annual Review of Sociology* 24, 265-290.

Xu, Y. (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57-76.

Recommended Citation

Geyer, Judy, Adi Greif, Betsy Ness-Edelstein, and Denise Mainvillle. (2022) Impact Evaluation Methods: Evaluating the Impact of Market Interventions on Targeted Beneficiaries. Rockville, Maryland: Abt Associates.

AgResults is a $152 million multilateral initiative incentivizing and rewarding high-impact agricultural innovations that promote global food security, health, and nutrition through the design and implementation of Challenge Projects, which provide payments for results intended to foster the creation of sustainable markets benefitting smallholder farmers. The AgResults initiative is a partnership between the Australian Government, the Bill & Melinda Gates Foundation, the Government of Canada, the United Kingdom's Department for International Development, the United States Agency for International Development, and the World Bank.

Abt Associates in partnership with Denise Mainville Consulting, is an external impact evaluator of AgResults. Abt Associates uses rigorous evaluation methods – both quantitative and qualitative – to determine whether the AgResults Challenge Projects achieve their objectives. These briefs summarize our lessons learned on individual projects, as well as cross-cutting topics.

The contents of this brief do not necessarily reflect the views of the AgResults partners. For more information about AgResults, visit: http://www.agresults.org.