ONE ACRE FUND

# Getting the Counterfactual Right

Farmers First

## Executive Summary

Good impact measurement is absolutely critical for organizations driven by social missions. Measurement enables organizations to understand their effectiveness, improve on their impact, and actually achieve their mission.

If impact studies are not designed well, program staff will have a false sense of their success and a poor understanding of where impact needs to improve. One area where measurement is often weak is selection bias – the problem of comparing individuals who have self-selected into a program with those who have not, and who might therefore be fundamentally different.

While One Acre Fund had historically assessed our impact quite rigorously – weighing harvest and using a comparison group – we were still concerned that selection bias could be affecting our impact assessments. This memo outlines our experience addressing selection bias through experimenting with four different approaches:

1. *Propensity Score Matching:* matching program and comparison farmers on important characteristics to test for obvious biases.
2. *Randomized Controlled Trial:* randomly assigning some enrolled farmers to receive our program (treatment) and randomly assigning some to not receive our program (control).
3. *Newly enrolled farmers:* creating a control group from farmers who have just enrolled in the program (resulting in high similarity to the test group), but who have not yet benefited from a harvest.
4. *Difference-in-difference:* looking at the year-over-year benefit of program participation for treatment farmers, compared against a year-over-year baseline of non-treated farmers.

One Acre Fund currently uses newly enrolled farmers where possible, propensity score matching throughout to test for bias, and randomized controlled trials (RCTs) in targeted instances to test the accuracy of the above approaches. We are also retooling our systems to allow for multi-year farming tracking, and to then introduce difference-in-difference where possible.

By testing each of the approaches above, we have learned lessons about how to make them better in our context, and also about the true magnitude of our impact. We believe these lessons are applicable to a broad community of peers who similarly struggle to understand and describe their true program impact.

## The Problem of Selection Bias

At its core, program evaluation rests on one central question: What would the world look like if your program did not exist? The impact of our program is what the world looks like for our client farmers, minus what it would have looked like had we not existed. The fundamental problem is accurately describing “what the world would have looked like absent our program,” otherwise known as “the counterfactual.”

# Getting the Counterfactual Right

Farmers First

At One Acre Fund, we have historically measured our impact by weighing harvest of One Acre Fund farmers and comparing those yields to the harvest among neighboring farmers. We also collected input cost data and calculated farmer profit. The physical weights allowed us to have an unbiased estimate of harvests and using neighbors ensured that both groups would be subject to the same agro-ecological conditions.

But the question remains: What if One Acre Fund farmers are fundamentally different from their non-program-joining peers? On one hand, One Acre Fund farmers may be inferior farmers as compared to their neighbors, and in need of greater assistance. In this case, measuring the harvest differences between program farmers and their neighbors would *under*-estimate our impact. On the other hand, it is possible that One Acre Fund farmers are better farmers than their neighbors and more motivated to improve their lives (as expressed by their decision to join the program). In this case, measuring the harvest difference between program farmers and their neighbors would *over*-estimate our impact. This is the problem of selection bias.

Our traditional methods of assessing impact did not address the issue of selection bias. Because we are committed to understanding our true impact, we have experimented with several techniques to address this selection bias problem. Below we outline what we have learned through this process and how those lessons might help other organizations struggling with similar challenges.

## Decision Rubric

There are a number of techniques in an evaluator’s toolkit to address the problem of selection bias, each with their own benefits and drawbacks. Organizations interested in social impact should make decisions about which is the best fit based on their goals and resources. When One Acre Fund decided to investigate which evaluation technique would work best for us, we outlined our four priorities. They are:

1. Good geographic representation of our program: Measurement has to be simple enough to repeat in many different regions within each of our countries.
2. Minimal selection bias.
3. One Acre Fund and comparison farmers facing similar benefits and hurdles (agro zones, access to other supports).
4. Resource reasonability of running M&E.

We then evaluated the four techniques below and how well they fulfilled our four priorities:

**Propensity Score Matching**. This is a statistical technique in which we match comparison farmers with One Acre Fund farmers on key demographic characteristics (age, education, wealth, family size, etc.) in order to help address selection bias. Historically we have not collected this data nor matched the same farmers across our cost input and harvest survey. However, we will do so going forward.

**Randomized Controlled Trial.** The most rigorous comparison group selection strategy is a randomized controlled trial (RCT). In this method, enrollees in our program are randomly selected to be excluded from our program (control) or included in our program (test). We cannot do this on a regular basis due to the ethical and operational problems associated with denying our program to people who have signed up, as well as the difficulty of doing an RCT in more than just a small geographic area. However, we make targeted use of RCT to occasionally verify the accuracy of our other measurement methods.

**Newly enrolled farmers.** Newly enrolled farmers are farmers who have joined the program but have not yet had any program benefit. They constitute a highly similar comparison group because they have self-selected into the program just like current program farmers, but have not yet received training or harvested. For our countries in which enrollment takes place before harvest (depending on the agricultural calendar), we can compare newly enrolled farmers' harvests with current program farmers.

**Difference-in-difference**. The difference-in-difference approach is another common methodological technique to help address selection bias. With this approach, we compare the change in harvest for One Acre Fund farmers in the year before they joined the program to the year that they had program participation. In order to eliminate the obvious potential bias of specific-year effects, we control this by looking at change in harvest yields among farmers who stayed out of the program in both seasons. This requires a much better tracking of farmers across time periods than we had previously done.
The relative benefits of these four options across our criteria are summarized below, in which more check marks signify a better fit.

| | Good geographic representation | Minimize selection bias | Face similar contexts | Resource reasonability |
|---|---|---|---|---|
| PSM | ✓✓ | ✓ | ✓✓ | ✓ |
| RCT | | ✓✓ | ✓✓ | |
| Newly enrolled | | ✓ | ✓ | ✓ |
| Diff-in-diff | ✓ | ✓ | ✓✓ | ✓ |

## Outcomes and Lessons Learned

In 2014, we experimented with each of these strategies in various country contexts. Below are specific examples of what we found and what we learned about the feasibility of each technique.

***Propensity Score Matching (PSM) shows promise***. In Kenya we use "interested neighbors" as a comparison group. These are farmers who have not joined One Acre Fund's program, but who our farmers tell us are interested in joining the following year. In order to see if these neighbors were truly similar to our farmers, we gathered data on a number of key demographic variables, such as family size, gender, land size, and wealth. We then statistically matched our farmers with similarly situated comparison farmers using PSM. We did this for maize farmers in Western Province, Kenya, and found a similar result to our initial analysis. Without using propensity score matching, we found a 39 percent increase in maize yields per acre and with PSM, we found a 36 percent increase in maize yields, both statistically significant. These results are similar to an RCT that we conducted that year in that area (see below).

While PSM is an imperfect solution (there are unobserved variables, like motivation and risk tolerance, which are harder to measure), it does allow us to mitigate the issue of selection bias. It only requires asking a few more questions on our surveys, and does not require long-term tracking of respondents (like difference-in-difference) or limitation in our geographic representativeness (like RCTs or newly enrolled client comparisons). We therefore plan to do train all our M&E analysts on this technique and to incorporate it in our regular impact assessments going forward.

***RCTs are the most rigorous option, and are an effective way to test our other methods***. In our 2014 season in Kenya, we ran a small randomized controlled trial in which we randomly selected 2 program sites (about 425 farmers) to be unenrolled in the program. These farmers were instead given health products and the ability to enroll the following season. Because both One Acre Fund and control farmers had self-selected into the program, this gave us two highly similar groups to compare. The results of this RCT were overall consistent with our regular M&E, actually showing higher results for that year, possibly because the area in which we located the RCT was a bit above average. However, because we randomized only over six sites, our data lacked a very high level of precision and significance was only at the 10 percent level. To increase precision in any future RCTs, we will work to increase the number of randomization units. Visit the impact page on our website for a more detailed explanation and full write-up.

While RCT is the most rigorous of all of our impact measurement options, it would be impractical to deploy at any meaningful scale across dozens of operating units in each of our four country operations, plus across all our different crops. There are also serious ethical and reputational issues associated with randomly denying program participation to enrolled farmers. Therefore, we have decided to use RCT occasionally to test and verify the accuracy of our other measurement methods (see below). We have significant room to improve in our execution of RCTs in order to improve our statistical power.

***Newly enrolled and likely-to-enroll farmers can be used to good effect in some countries.*** In Burundi, farmers enroll in the program for the next season long before our current farmers have harvested. Because of this, we are able to create a "newly enrolled farmers" control group ahead of harvest, and then compare their harvests to our program farmers. Because both test and control groups have selected into the program, but only one group has actually been through One Acre Fund training and received inputs, this provides an excellent opportunity to assess our program while addressing selection bias. In 2014, we harvested beans with 195 enrolling clients and compared that data with program farmers in a similar area. The program farmers earned 26 percent more profit than newly enrolling farmers. This effect size was smaller than the program impact we found in our regular M&E. This may be because the newly enrolled farmers all came from a small new area, which is not wholly representative of our program, or it may be that the newly enrolled farmers form a better comparison group. We will continue to check our program impact assessment by comparing with newly enrolled in our new program areas in Burundi.

In Kenya, we cannot compare the harvests of program farmers with newly enrolled farmers because the harvest happens before enrollment. However, in Kenya, we can look for “interested” farmers to use as a comparison group (these are farmers who their neighbors tell us are interested in joining the program in the following season), and we have some evidence this is a good selection method. In 2013, we used these farmers as a control group, and tracked how many actually enrolled in 2014. Just under half of these “interested farmers” enrolled in 2014 and more importantly, those farmers who enrolled looked very similar to those who stayed out of the program. Our 2013 impact estimates are similar whether we use the interested farmer group (51% greater profit per farmer) or the farmer who did eventually enroll (45% greater profit per farmer), providing some evidence that the “interested farmer” group is a reasonably valid comparison group.

***Difference-in-difference requires improvement to our data collection systems.*** The difference-in-difference approach requires tracking farmers over time. The goal is to compare harvest of a farmer before program participation and after, controlled for fixed-year effects by doing the same comparison for a farmer that did not participate in our program in either year. In 2014, we intended to do a difference-in-difference analysis in Kenya using data from interested farmers from 2013 who either did or did not join the program in 2014. For a number of reasons, we did not end up with data we could use to do this analysis. First, we lacked a good system for matching farmers from one season to the next. For example, if a farmer had been surveyed in 2013, but we recorded them with a misspelled name, we were not able to easily link information on a specific farmer over two seasons. In addition, we did not adequately communicate to our enumerators (survey agents) the importance of finding specific farmers who we had surveyed the prior year, so many farmers were missed.

However, we learned a number of important lessons and have altered our systems and re-trained our enumerators. We have better data to track farmers over time and have emphasized the importance of tracking specific farmers to our staff. We hope to be able to conduct a difference-in-difference analysis with farmers in Kenya, Tanzania, and Burundi in 2015.

### Learnings and Next Steps

As a result of our focus on improving our impact measurement, we uncovered the following:

- **Our program effect seems to be verified by different measurement methods.** Through experimenting with different methodologies and statistical techniques in 2014, we learned that selection bias does not seem to explain our program effect. All design methods and techniques – RCT, PSM, newly enrolled farmers – agree at least directionally, adding to our body of impact evidence.
- **It’s important to balance feasibility and rigor.** We learned important lessons about the trade-offs in terms of feasibility and rigor of each method. Going forward, we will be able to do PSM

across our regular M&E data collection efforts. This will allow us to address the issue of selection bias in each country with minimal field costs and maximum geographic representativeness. In addition, we can periodically check our data with the other methods as they are appropriate. In 2015, we will have difference-in-difference estimates from three of our countries as well as an analysis of how newly enrolled and likely-to-enroll farmers fare compared to currently enrolled farmers. In the future we may also run additional randomized controlled trials in different program areas to apply the most rigorous evaluation assessment to test our program as we grow and make sure our M&E is on track.

Going forward, we plan to pursue the following next steps

- We will be using propensity score matching in all of our program countries to help address the issue of selection bias.
- Depending on where it is feasible, we will also employ other quasi-experimental approaches to help verify our findings.  In Tanzania, Kenya and Burundi, we will use the difference-in-difference approach. In Burundi we will also test our results by looking selectively at newly enrolling farmers and comparing them to already enrolled farmers.
- In the future, we will consider running a larger scale RCT to most rigorously test our program impact.

## Key Conclusions

Other organizations might wish to consider a similar analysis of the evaluation tools available, even if an organization does not have the resources to experiment with each method. We propose the following steps:

- **Decide on goals.** We wanted to achieve rigor and geographic representativeness within resource constraints. Other organizations might want to test a specific intervention or population and this could have implications for their specific goals.
- **Identify possible methods**. It might require a new sample, survey methodology, or analysis, but more rigorous methods are worth the investment.
- **Test those methods.** If resources are available, it could be worth testing new methodological approaches to identify problems and better understand limitations early.
- **Be transparent about the limitations of data.** When the analysis is complete, it is important to be transparent about the limits of the data. An experiment run over a small geographic area, for example, should not be considered as representative of the program as a whole.