60__decibels

Exploring Lean Evaluation Methodologies for Digital Farmer Services

Lessons and Limitations

address > 228 Park Ave S, PMB 54413, New York, NY 10003 USA / 60decibels.com / ellie@60decibels.com

Acknowledgements

Thank you to all the companies that worked with us for their support throughout the lean evaluation projects.

Thank you to all stakeholders we collaborated with along the way.

This work was generously sponsored by The Bill and Melinda Gates Foundation through the Busara Center for Behavioral Economics.

1. Introduction

Evaluations are important tools for assessing the impact and effectiveness of interventions or programs. There are various evaluation methodologies, but when it comes to identifying causal impacts, two main types stand out: Randomized Controlled Trials (RCTs) and Quasi-Experimental Designs (QEDs).

RCTs involve randomly assigning individuals to a treatment group, which receives the intervention, or a control group, which does not. This random assignment ensures the groups are comparable, so any differences in outcomes can be attributed to the intervention itself.

In contrast, QEDs do not rely on random assignment. The most common QED is the difference-in-difference (DiD) analysis, which involves a treatment group and a comparison group with similar baseline characteristics, despite not being randomly assigned. While not as rigorous as RCTs, QEDs are valuable when randomization is not feasible due to ethical or practical reasons.

Both RCTs and QEDs aim to determine whether an intervention causes a hypothesized outcome by comparing results between the treatment and control/comparison groups. RCTs provide rigorous evidence due to randomization but can be resource-intensive, time-consuming, and ethically challenging. In some cases, QEDs offer a viable alternative.

In the growing digital farmer services (DFS) sector, there is a demand for rigorous evidence of impact that does not require RCTs. Under the Dig-It-Al grant, 60 Decibels and Busara tested new methodologies to meet this demand. 60 Decibels' core methodology, Lean Data, relies on short, phone-based interviews with end users to assess the impact of a program or intervention. These studies are low-cost, completed in under four months, and identify directional shifts in outcomes such as production and earnings without quantifying impact.

Leveraging this expertise, we sought to develop a methodology with more rigorous causal inference than Lean Data but less resource-intensive than traditional RCTs and QEDs. This led to the concept of a lean evaluation.

Initially, we aimed to conduct a difference-in-difference analysis in a ‘lean’ way. DiD studies typically involve a treatment group and a control group, with data collected at baseline and follow-up to compare changes in outcomes over time. The Lean Evaluation approach incorporated modifications aligned with the 60 Decibels Lean Data methodology. Data collection was conducted through phone interviews rather than in-person surveys. A retrospective approach was used to establish the baseline, with farmers reporting outcomes from the season before their enrollment in DFS. The comparison group was selected from farmers engaged with DFS providers but not enrolled in DFS. The sample size was relatively smaller, with 600-800 respondents, balancing statistical robustness with operational constraints and resource efficiency.

Under the grant, we implemented two Lean Evaluations using this methodology and conducted a third study to test a cross-sectional lean methodology, described in the following sections.

We learned valuable lessons by testing the lean evaluation methodology. This review aims to share these insights to help readers choose the most suitable methodology for measuring the impact of their DFS.

- Section 2 delves into the three methodologies: lean evaluations, RCTs, and traditional QED studies. We compare these methods with the 60dB flagship Lean Data studies and a new methodology, the lean cross-sectional analysis. Table 1 summarizes the pros and cons of each methodology and discusses which DFS are best suited for each approach.
- Section 3 focuses on implementing the lean evaluation methodology for DFS and discusses the lessons learned from each experiment. It provides background information on the evaluated DFS, outlines the steps to conduct a lean evaluation, and discusses the limitations of this approach uncovered during these trials.

2. Standard vs. Lean evaluation

a. Characteristics of Lean Evaluations

The Lean Evaluation methodology adopted the QED method of DiD analysis. The key features of DiD analysis include:

Assignment of treatment and control groups: These groups are selected based on their exposure to the intervention. The treatment group consists of individuals or entities that receive the intervention, while the control group comprises those that do not receive the intervention. The sampling strategy aims to ensure that the treatment and control groups are as similar as possible in terms of their baseline characteristics. This can be achieved through random sampling or by carefully matching individuals or entities based on relevant characteristics.

Data collection: Data is collected from both the treatment and control groups at two points in time - before the intervention (baseline) and after the intervention (follow-up). The baseline data provides a reference point for measuring changes, while the follow-up data captures the outcomes after the intervention has been implemented.

Survey tool: In difference-in-difference (DiD) studies, the surveys are designed to collect comprehensive data, including outcome measures and profile information. This baseline information is crucial for ensuring comparability between the treatment and control groups and establishing parallel trends, accounting for any pre-existing differences. Given the level of detail required for DiD studies, the surveys tend to be extensive, often spanning 1 to 2 hours, sometimes longer. Researchers aim to maximize the data collected during these undertakings, capturing a wide range of variables that may influence or be influenced by the intervention under evaluation.

Sample size: The sample size is determined based on factors such as the expected effect size, desired statistical power, and the level of precision required. Generally, for DiD studies larger sample sizes of 1000 or more are preferred since they increase the statistical power and precision of the study.

Analysis: The key analysis in a DiD study involves comparing the changes in outcomes over time between the treatment and control groups. The difference in the changes between the two groups is attributed to the impact of the intervention, assuming that other factors affecting the outcomes are constant across the groups. In other words, Difference-in-difference compares the change in the outcome of interest for the treated group before and after the treatment, with the change in the outcome for a control group over the same period.

For the lean evaluation, we took the DiD methodology described above as our foundation and made the following changes to make it leaner.

1. Brief, telephonic interviews: To optimize cost and time efficiency, the lean evaluation utilized telephonic interviews for data collection instead of in-person surveys. This approach leverages the 60 Decibels' lean data approach, enabling efficient data gathering from geographically dispersed respondents without incurring substantial travel costs. The interviews ranged from 20-30 minutes in length and collected only the minimum information necessary to conduct the analysis.
2. Baseline: Since we could only identify a farmer as being "treatment" after they had already registered for and gained access to the DFS, we used a retrospective approach, asking farmers to describe the season immediately preceding their registration. This established a baseline reference point without the need for a separate pre-intervention data collection phase. We aimed to conduct these surveys as soon as possible following farmer registration to minimize the recall period.
3. Selection of comparison groups: We relied on DFS providers to identify comparison groups and provide phone numbers for those farmers. This meant that the provider was limited to farmers that they had engaged with in some way but had not enrolled in the DFS. This limits comparability between groups, as these individuals may differ systematically from those who do register, potentially due to factors like willingness or ability to pay for the service or the relevance of the offering to them compared to users in the treatment group. We impose the assumption that these differences are time-consistent and therefore accounted for in the DiD analysis but recognize it as a limitation.
4. Sample size: While traditional DiD studies often require larger sample sizes, the lean evaluation aimed for a sample size of approximately 600-800 respondents. This sample size was determined based on operational constraints (ie. the number of newly registered users of the DFS) and considerations of statistical power, expected effect sizes, and resource constraints, balancing the need for robust analysis with cost and time efficiency.

b. Features of standard evaluation methodologies

The standard evaluation methodologies considered for the purpose of this handbook are Randomized Controlled Trials and Quasi-experimental designs. In the context of evaluation methodologies, Randomized Controlled Trials (RCTs) and Quasi-experimental designs (QEDs) stand out as two prominent approaches with distinct features and applications.

RCTs are characterized by the random assignment of participants to either a treatment group that receives the intervention or a control group that does not. This random assignment ensures that the groups are comparable at the outset, reducing the likelihood of bias in the results. By conducting RCTs in controlled environments, researchers can isolate the effects of the intervention under study, minimizing external influences and focusing on how the treatment impacts outcomes. RCTs are considered the gold standard and provide high internal validity due to their rigorous design. However, ethical concerns arise when some participants are denied access to the treatment solely for research purposes, raising questions about fairness and participant well-being. Additionally, RCTs can be resource-intensive, requiring significant financial investment, dedicated staff, and time commitments to execute effectively, making them less feasible for certain studies with limited resources.

On the other hand, Quasi-experimental designs (QEDs) offer an alternative approach to evaluating interventions by testing causal hypotheses without random assignment. Instead of randomization, QEDs rely on creating comparison groups that mirror the characteristics of the treatment group to assess outcomes that would have occurred in the absence of the intervention. Techniques such as regression discontinuity design and propensity score matching are employed in QEDs to mitigate selection bias and enhance the validity of findings. One key advantage of QEDs is their applicability in real-world settings where randomization may not be ethically or logistically feasible. By conducting counterfactual analyses that compare outcomes between treatment and comparison groups, QEDs enable researchers to attribute observed effects to the intervention without relying on random assignment. However, QEDs are prone to selection bias due to non-random assignment and may have difficulty controlling for all confounding variables, potentially impacting study validity.

In summary, while RCTs offer rigorous control over group assignment and environmental factors for evaluating interventions, they come with ethical considerations and resource challenges. On the other hand, QEDs provide a practical alternative for causal inference in situations where randomization is not viable, offering flexibility and applicability in real-world contexts through innovative comparison group creation and bias mitigation techniques. Both methodologies play vital roles in advancing evidence-based decision-making by providing valuable insights into the effectiveness of interventions across diverse settings and research scenarios.

c. Evaluation method selection matrix

Lean evaluations, randomized controlled trials (RCTs), and quasi-experimental designs are distinct research methodologies used to assess the impact of interventions. Each approach has its own set of advantages and limitations that influence their suitability for different evaluation contexts.

Table 1: Evaluations methods selection matrix

Evaluation Method	Pros	Cons	Suitability
RCTs (Randomized Controlled Trials)	High Internal Validity: Randomization minimizes selection bias for strong causal inferences Gold Standard: Rigorous design considered the benchmark for evaluating intervention effectiveness	Ethical Concerns: Perceived denial of services to control groups can raise ethical issues Resource Intensive: RCTs are often expensive and lengthy to conduct Operationally Complex: Requires complete control over who receives the intervention and when	Research in which attributing causal impact is the top priority (eg. programs with large-scale public policy relevance) Programs or interventions where it is feasible to randomize assignment to treatment and control groups.

Quasi-Experimental Designs	Practicality: More feasible than RCTs in real-world settings with ethical/logistical constraints Ethical Considerations: Does not restrict a group's access to an intervention for the purposes of research	Limited Internal Validity: non-random assignment can introduce some selection bias Confounding Factors: Difficulty controlling all confounding variables, impacting study validity Resource Intensive: Similar or slightly lower time and cost requirements as an RCT	Research in which attributing causal impact is important but random assignment is not feasible or ethical However, some level of control or comparison is still possible
Lean Evaluations	Practicality: Does not require control over who receives an intervention Less Resource Intensive: It takes less time to conduct a phone-based lean evaluation	Need for phone numbers: Reliance on phone interviews can introduce bias, especially obtaining contacts for comparison groups , and can introduce operational challenges in capturing a baseline Limited Validity: Constraints in obtaining comparison groups may impact result validity. Spillover/contamination is very plausible Attrition: More vulnerable to attrition Outcomes: Outcome measurement is less precise due to shorter interviews, and longer-term outcomes are less likely to be captured due to timeframe	Assessment of digital programs/ products/services aimed at driving short term behavioural change They require some control over the comparison group and work best when contact information for respondents is available as part of the product or service implementation
Cross Sectional Comparison Studies	Practicality: Does not require control over who receives an intervention. Two groups are compared at one point in time. Less Resource Intensive: One-time survey is lower cost, and results are available in the same year	Measures differences, not causal impact: The methodology does not capture changes over time due to an intervention, but rather describes the differences in practices (or outcomes) between two groups, while controlling for observables to minimize selection bias.	Research where robust causal inference is not a top priority When adoption of practices or behaviors is an important outcome When data and feedback is needed

	Direct Feedback: opportunities to ask users directly about their experience with an intervention	Outcomes: Better for measuring adoption of practices and behaviors than quantifying outcomes	quickly to inform adaptation
60dB Lean Data Studies	Rapid: Studies conducted within 16 weeks Low-Cost: Short, one-time phone surveys with a smaller sample Flexible: Can be iterated, adapted in response to changing technology or program Direct Feedback: opportunities to ask users directly about their experience with an intervention	Subjective causal inference: relies on a farmers' own attribution of changes to a given intervention Outcomes: generally directional rather than quantifiable	Research where robust causal inference is not a top priority When data and feedback is needed quickly to inform adaptation

In summary, lean evaluations may offer an alternative to RCTs and QEDs in certain cases but come with operational limitations and challenges. RCTs provide high internal validity but come with ethical and cost considerations. Quasi-experimental designs balance internal and external validity, making them practical in real-world settings but susceptible to selection bias. Understanding the strengths and limitations of each method is crucial for selecting the most appropriate approach based on the specific research objectives and constraints.

3. Lean Evaluation Methodology for DFS

a. Background

Before we dive into the methodology, we will first summarize the three lean evaluations that we explored in the last year.

I. CoAmana:

Digital Solution: CoAmana's "Amana Market" platform is an online marketplace that connects farmers in Nigeria with buyers and traders for their produce. Their primary offering is an online market access tool; however, they also offer SMS farm advisory and aggregation services that allow farmers to bundle their produce for customers buying in bulk. They had recently improved their digital services by adding USSD functionality and two-way SMS.

Primary research question: How does using the Amana Market platform affect farmers' sales volume and the price they receive for their produce? How does using the Amana Market platform affect farmers' post-harvest loss?

Methodology used: Lean evaluation (DiD using the 60dB lean approach)

II. ACRE

Digital Solution: ACRE Africa's BIMA PIMA is a digitized weather-index micro-insurance product in Tanzania that uses satellite and weather station data to monitor rainfall levels. Excess rainfall or drought conditions trigger pay-outs to farmers which are directly deposited into the farmer's mobile account at the end of the season. Farmers can register through:

- Village Champions (Hybrid Model) who host in-person trainings and sell physical scratch cards. These scratch cards can then be registered later using a mobile phone.
- Call Centre Representatives (Digital Model) who reach out to and educate farmers via phone about the importance of insurance and direct them to Village Champions in their respective villages for registration.

Primary research question: Do maize farmers who adopt BIMA PIMA invest more in the productivity of their farms?

Methodology used: Lean evaluation (DiD using the 60dB lean approach)

III. TomorrowNow

Digital Solution: TomorrowNow.io provides developers and businesses with precise hyper-local weather data through its technology and weather API. In Kenya, TomorrowNow.org, the non-profit arm of TomorrowNow.io, has partnered with the Kenya Agriculture and Livestock Research Organization (KALRO) to deliver climate-smart agriculture information and advice to farmers via SMS, completely free of charge. Farmers can receive one of two types of messages:

- **Version 1 (V1):** KALRO's value-chain specific general farming advisory (**Comparison**)

- **Version 2 (V2):** KALRO's value-chain specific general farming advisory enhanced with TomorrowNow's hyper-local weather advisory (**Treatment**)

Primary research question: How does the adoption of maize planting practices differ between farmers receiving V2 messages and farmers receiving V1 messages?

Methodology used: Lean cross-sectional comparison - We used a logit regression, which is a tool for studying the relationship between binary outcomes and multiple predictor variables at a specific point in time.

b. Survey design

For each study, we aimed to develop a survey that could be administered via phone and provide us the information required to answer the research questions. Below are some important considerations in designing a survey.
Outcome Indicators: Choosing outcome indicators that are relevant, measurable, and aligned with the research questions and expected effects of the DFS is crucial. In standard evaluations, both medium-term and longer-term outcomes are typically measured. However, given the quicker nature of lean evaluations, our experience points towards focusing primarily on medium-term outcomes.

In the CoAmana lean evaluation, we tried to measure sales volume of crops and revenues earned from them. However, these outcomes are less likely to change during the follow-up period because one year is a relatively short timeframe to observe changes influenced by multiple variables, such as productivity and market prices. Nevertheless, we included questions on price received for produce and perceived agency in decision-making, which are shorter-term outcomes more likely be affected when starting to use a new market access platform.

Based on these learnings, we recommend that lean evaluation surveys concentrate on intermediate outcomes. For example, to assess the impact of ACRE insurance, we focused on changes in farm investment, rather than farm earnings, which would take longer to realize and may require more statistical power to detect. While longer-term impacts are valuable, the condensed timeline and iterative nature of lean evaluations make it more practical to measure changes in indicators that can manifest within the evaluation period. Instead of long term outcomes, modules on company experience would prove more valuable.

Survey Length: Since data collection is conducted over the phone, it is ideal to limit the conversation to about 30-35 minutes to ensure the best quality data. A survey length of 25-30 questions would be optimal. In the lean evaluations we conducted, the surveys were closer to 50 questions, as they included long-term outcomes, short-term outcomes, profile, and experience modules. Many respondents complained about the survey duration, and enumerators noticed respondent fatigue setting in. To ensure good data quality, multiple rounds of feedback were required, given the length of the surveys.

Question Type: Typically, Randomized Controlled Trials (RCTs) and quasi-experimental design studies have separate quantitative and qualitative components, which are time-consuming and require a substantial budget. However, in our lean evaluations, we attempted to incorporate a mix of quantitative and qualitative questions within the same survey. While this approach increased the survey length, it provided valuable context on why certain outcomes changed or did not change over time. Qualitative insights, especially when received in the short term, can be quite valuable, as they could potentially aid in the iteration and refinement of DFS. The qualitative questions are a regular component of the lean data studies as well.

c. Data collection

Baseline

A critical first step is to conduct a baseline survey before the introduction of the DFS for both the treatment and comparison groups. Ideally, this establishes the pre-intervention status and captures any pre-existing differences between the two groups. However, based on practical experience, it is highly likely that respondents in the treatment group might already have some knowledge about the DFS, as enrolling or registering for it is often a prerequisite for companies to share their contact information. This kind of prior exposure might create a bias in the baseline measurements.

Ideally, the baseline survey would be conducted before any farmer gains access to the DFS, but this is unfeasible. To accommodate the requirement of farmers registering with the survey and thereby being assigned to the treatment group, the baseline survey should be conducted immediately after registration to avoid prolonged exposure to the DFS and, more importantly, prevent the treatment group from using the DFS before the baseline is captured. An alternative approach is to set the reference period for measuring outcomes prior to registration. Even in this case, it is vital to conduct the baseline survey promptly, as recall accuracy might become challenging as time progresses. Regardless of the approach, minimizing the time between registration and baseline data collection is crucial to mitigate potential biases and ensure the most reliable pre-intervention measurements. Operationally we found this to be very challenging. The survey is time sensitive, given the agricultural seasons, so we needed to ensure that each DFS could enroll and share the contacts of a minimum number of new users and comparison farmers within a short window.

Follow-up

The follow-up survey for a lean evaluation should conducted one year after the initial survey, involving the same respondents, and focused on the same agricultural season. However, there is a possibility of attrition, where some respondents may drop out or become unavailable. To mitigate this issue, it is advisable to offer incentives to the respondents, such as airtime or other rewards, to encourage their continued participation, as well as over-sample at baseline. However, this should be lower compared to longer evaluations. Conducting the follow-up survey on time is crucial, as any delay may impact the accuracy of the respondents' recall and potentially skew the data. Timely execution of the survey is essential to ensure reliable and consistent results.

d. Sampling

i. Sampling strategy

The sample frame should be drawn from respondents who have engaged with or are starting to engage with the DFS being evaluated. More specifically, contact details for two different groups should be available:

Treatment group: A randomly selected sample of respondents registering or newly registered to use the DFS from a specific start date onwards. These respondents may or may not have used the DFS yet, and they will have access to it. Ideally, this group should be newly enrolled (from the start date onwards) and have not used the DFS at all. However, if this is not feasible for all respondents in the sample, questions will only cover their experience during a specified time period before they enrolled.

Comparison group: A randomly selected group of respondents who have never used the DFS and will likely not use the DFS version until after the follow-up is complete. This group will consist of respondents who were approached but did not end up using the DFS. Further checks should be done to ensure this group has not used any version of the DFS by matching contact details with usage data.

ii. Sample size and attrition

We conducted power calculations to estimate the required sample size. These were calculated in a way that they were adequate for detecting an effect on main outcomes of interest that were predefined. The calculations considered factors such as the level of significance, proxy measures for the outcomes of interest, and assumed standard deviations based on literature. We estimated a 10% attrition rate between the baseline and the follow-up period of one year from experience. We also maintained a ratio of 1.25:1 for treatment and comparison group sample sizes to account for potential lack of sufficient uptake in the treatment group.

Guided by these assumptions, our initial sample size targets were 400 for the treatment group and 350 for the comparison group. However, we faced challenges as these targets were insufficient due to higher attrition rates than expected. Several factors contributed to the high attrition, including unreachable phone numbers, respondents no longer meeting eligibility criteria (such as selling their crops for the season or growing a certain crop they had cultivated during the baseline), and outright refusals to participate.

e. Analysis

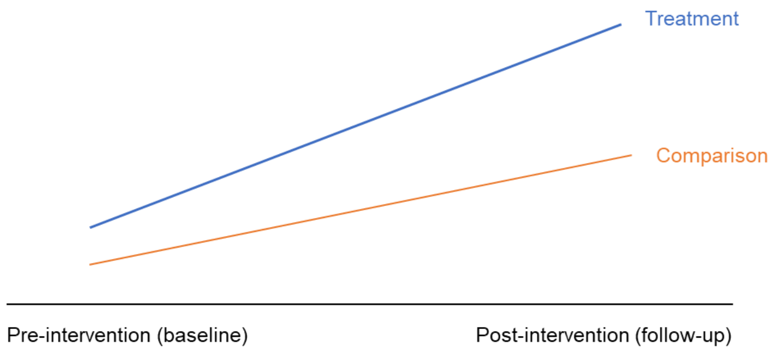
i. Balance test

Companies provided phone numbers for comparison group respondents based on who had declined to enrol. This might have introduced selection bias into our sample if respondents who opted out of a service are systematically different from the treatment group. Doing the balance test after baseline is therefore essential to establishing comparability.

The primary goal of the balance test is to ensure that the treatment and control groups are statistically similar in terms of their observable characteristics before the treatment is introduced. This test compares the means or distributions of key variables, such as demographic characteristics, baseline outcomes, or other relevant factors, across the two groups. If the groups are well-balanced, it implies that any observed differences in outcomes after the intervention can be attributed to the treatment itself, rather than pre-existing differences between the groups. Observable differences between the two groups should be controlled for at endline.

ii. Difference-in-difference estimation

We attempted the difference-in-difference (DiD) analysis to understand the impact of DFS on predefined outcomes. The DiD methodology is a statistical technique used in quasi-experimental research designs to estimate the effect of a treatment or intervention on an outcome of interest. It compares the changes in outcomes over time between a treatment group that receives the intervention and a control group that does not.



The key principle is that any unobserved factors that may influence the outcome are assumed to be consistent across both groups. By taking the difference in outcomes between the treatment and control groups before the intervention (first difference), and then subtracting it from the difference in outcomes after the intervention (second difference), the DiD method accounts for any initial differences between the groups as well as trends over time that may affect both groups.

This double differencing approach aims to isolate the true effect of the treatment, providing a more robust estimation compared to simple pre-post comparisons or cross-sectional analyses. The DiD method is commonly used in policy evaluations, natural experiments, and scenarios where randomized controlled trials are not feasible or ethical.

Lessons from analysis

The analysis has highlighted two key limitations:

- Statistical power affected by low take-up rates

A significant challenge was the low take-up rate of the DFS among the target population. This low engagement directly impacted the statistical power of the analysis, making it difficult to detect meaningful differences between the treatment and comparison groups.

When a substantial portion of the treatment group does not engage with the service, it reduces the ability to detect statistically significant effects and potentially underestimates the true impact of the DFS.

- Limited observability of long-term outcomes in 1 year evaluation

Another critical limitation was the difficulty in observing long-term outcomes within the short evaluation period of one year. Outcomes like increased income, enhanced resilience, and overall productivity improvements often require more time to manifest. In our analysis, these long-term impacts showed no major changes within the one-year timeframe. This highlights the challenge of capturing significant long-term benefits in a condensed evaluation period for agricultural DFS.

iii. Cross-sectional analysis

For TomorrowNow, we tested the cross-section lean methodology. In addition, we employed a mixed methods design, combining both quantitative and qualitative data at the customer level. This approach ensured a comprehensive evaluation of TomorrowNow and KALRO services and facilitated the answering of key research questions using the same customer-level data.

To analyse the impact, we used a logit regression, which is a tool for studying the relationship between binary outcomes and multiple predictor variables at a specific point in time.

In the regression analysis, we controlled for various observable characteristics such as gender, age, education, land size, location (county), income source, and others. This helped minimize potential bias arising from unobserved omitted variables and enhanced the reliability and validity of the estimates. We set the confidence level at 95% (p-value less than 0.05) to determine statistical significance.

The limitations of this methodology are summarized below:

- We assume a linear relationship between the control variables and the log-odds of the binary outcome for using the logit model. If this assumption is violated, and there is a non-linear relationship, the model may not accurately capture the true underlying associations.
- Cross-sectional analyses have limitations in establishing causality since the data is captured at a single time point. While they can provide associations between variables at a specific moment, they cannot determine cause-and-effect relationships.

Despite these limitations, the lean cross-sectional methodology was quicker to implement, and provided valuable and more rigorous insights compared to the 60dB lean data studies in a short period of time. We recommend this approach to companies pursuing slightly more rigorous insights on concrete medium-term outcomes like behaviour change, change in investments and have access to contacts of a comparison group (this can be a more basic version of their DFS, or a group that does not use the service).

4. Lessons and Limitations

1. Difficult to Establish Baseline and Comparison Groups

Establishing true baselines for DFS evaluations is problematic as contact information is typically obtained after users have already registered for the service. This prior exposure makes it unlikely to capture genuine baseline values for outcomes before any interaction with the service occurs. The comparison groups are also limited to farmers the DFS provider has previously engaged with, potentially introducing self-selection bias. These farmers may have different characteristics compared to registered users, possibly due to factors like willingness or ability to pay, or the perceived relevance of the service. These limitations can significantly impact the validity of evaluation results, making it challenging to draw definitive conclusions about the effectiveness of DFS.

2. Expect Higher Attrition with DFS

Lean evaluations conducted primarily via phone interviews face higher attrition rates compared to in-person methods. The ease of declining participation or dropping out during a phone interview, along with potential distractions and technical issues like poor signal strength, contribute to this increased attrition. Additionally, changes in participants' phone numbers further complicate follow-up efforts. DFS studies face even further attrition due to service uptake issues, especially if treatment is defined as initial enrollment rather than sustained use and adoption. To accurately measure impact, it's crucial to observe meaningful and continued use of the service over time, ensuring that the effects being measured truly represent the service's impact on actively engaged farmers.

3. Measure Crop-Specific and/or Intermediate Outcomes

Lean evaluations typically focus on medium-term outcomes within a year, as extending beyond this timeframe would make the study similar to a quasi-experimental design. This limitation makes it challenging to measure long-term impacts such as changes in sales volume, revenue, income, and resilience. Phone-based data collection, which is typically shorter interviews compared to in-person data collection, further complicates accurate measurement of outcomes like income and yields. For DFS that are not specific to a particular value chain, measuring outcomes becomes even more challenging due to the variety of crops involved. In such cases, focusing on impacts on practices and behaviors can be a more feasible approach, allowing for the assessment of intermediate outcomes that can indicate potential long-term impacts of the service.

4. Consider Risk with Digital Farmer Services:

Many DFS companies operate in early stages, with their operations not fully scaled or functional. This developmental stage introduces potential risks when evaluating their impact. Conducting lean evaluations requires establishing a baseline that often involves assessing new services, customer bases, or geographical areas. These factors introduce additional risks as the services being evaluated may still be evolving or adapting to their target markets.

5. Plan for Resource Allocation

While lean evaluations are expected to be faster and lower cost than randomized controlled trials (RCTs) and in-person quasi-experimental designs (QEDs), they still require significant resources. These evaluations typically take about 18 months to yield results and can exceed $100,000 in costs, depending on various factors.

Drawing from our experience, we have identified several critical conditions that need to be met for conducting an effective lean evaluation of Digital Farmer Services (DFS). These conditions address key challenges we encountered and aim to enhance the validity and practicality of the evaluation process:

1. Focus on Medium-Term, Observable Outcomes:

Evaluations should prioritize medium-term outcomes that are directly associated with the use of the DFS and can be reasonably observed within the evaluation timeframe (typically one year). Based on our experiences, we recommend focusing on outcomes such as:

- Changes in farming practices (e.g., adoption of recommended planting techniques)
- Intermediate indicators of productivity (e.g., effective pest management)
- Behavioral changes (e.g., increased investment in farm inputs)

We initially included outcomes like sales volume but found these challenging to measure accurately in a lean, phone-based evaluation. Long-term outcomes such as increased income or enhanced resilience are typically not observable within the one-year timeframe and should be avoided.

2. Type of Service and Take-up Considerations:

The nature of the DFS significantly influences the feasibility of a lean evaluation. Services that have a clear, immediate use case and a high likelihood of take-up are more suitable. For example:

- Input financing services where farmers are eligible to finance their purchase of inputs
- Insurance products with clear enrollment processes
- Advisory services with regular, trackable interactions

High take-up rates are crucial for maintaining statistical power in the analysis. In our studies, low take-up rates significantly impacted our ability to detect meaningful differences between treatment and comparison groups.

3. Operational Readiness of the DFS Provider:

The DFS provider must have robust systems in place for:

- Timely identification and sharing of contact information for newly registered users
- Ability to track service usage and engagement
- Capacity to identify a suitable comparison group from their network

Many DFS companies are in early stages of operation, which can introduce additional risks and complexities to the evaluation process.

4. Resource Allocation and Timeline:

While lean evaluations are designed to be more efficient than traditional RCTs or QEDs, they still require significant resources:

- Typical timeline: About 18 months from inception to final results.
- Estimated cost: Can exceed $100,000, depending on various factors

- Need for skilled personnel to design surveys, conduct phone interviews, and perform statistical analysis

5. Realistic Expectations for Causal Inference:

While the lean evaluation methodology aims to provide more rigorous insights than simple before-after comparisons, it has limitations:

- Difficulty in establishing true baselines due to the timing of user registration
- Potential for selection bias in the comparison group
- Challenges in controlling for all confounding variables

6. Unbiased Comparison Group:

Establishing a suitable comparison group is crucial for the validity of the evaluation. As observed in our studies, comparison groups often consist of farmers who have been approached but declined to use the DFS, potentially introducing self-selection bias. Careful consideration must be given to ensure the comparison group is as similar as possible to the treatment group in terms of observable characteristics.

Conclusion:

The lean evaluation approach for DFS offers a middle ground between quick feedback studies and resource-intensive RCTs. However, its effectiveness is contingent on meeting the above conditions. This methodology may be most suitable for NGOs or agricultural development programs with more control over comparison groups

For many commercial DFS companies, particularly those in early stages or with low take-up rates, alternative approaches such as 60 Decibels' Lean Data studies or lean cross-sectional comparisons may be more appropriate. These methods can provide valuable insights with fewer operational challenges, albeit with less rigorous causal inference.

Ultimately, the choice of evaluation methodology should be based on a careful assessment of the specific context, research questions, and operational realities of the DFS being evaluated.