Evaluating Financial Products and Services in the US

A Toolkit for Running Randomized Controlled Trials

Innovations for Poverty Action
poverty-action.org

ipa
INNOVATIONS FOR
POVERTY ACTION

Authored by Julia Brown, Lucia Goin, Nora Gregory, Katherine Hoffmann, and Kim Smith, 2015.
Please send comments and feedback to usfi@poverty-action.org.

Disclaimer: This document provides general information related to the law and is intended for informational purposes only and not for the purpose of providing legal advice. You should contact your attorney to obtain advice with respect to any particular issue or problem. Use of this information does not create an attorney-client relationship between IPA and the user or reader.

Acknowledgments

This work was made possible through the generous support of the Ford Foundation. This toolkit also benefited greatly from input and comments from David Batcheck, Matthew Bombyk, Leah Bridle, Jeremy Burke, Katy Davis and others at Ideas42, Aaron Dibner-Dunlap, Laura Feeney, Hugo Gerard, Elijah Goldberg, Katherine Hoffmann, Julian Jamison, Dean Karlan, Shawna Kolka, Kata Mihaly, Rohit Naimpally, Kristi Post, Beniamino Savonitto, Marc Shotland, Joshua Sledge, and Jonathan Zinman. Any errors are our own.

About Innovations for Poverty Action and the US Finance Initiative

Innovations for Poverty Action (IPA) is a 501(c)(3) US-based non-profit organization founded in 2002. Our goal is to serve as the link between academic research and problems faced by the poor. IPA's core competency is designing and managing rigorous evaluations to discover which interventions are most effective at reducing poverty. We have conducted over 350 research projects with partners in 51 countries around the world.

The Financial Inclusion Program (FIP) at IPA seeks to identify effective solutions to promote healthy financial behavior and to share our results to inform the work of financial service providers and governments around the world. FIP partners with financial institutions, governments, and researchers to design and test financial services and products that help households better manage their finances. Within the United States, our work is managed under the US Finance Initiative (USFI). USFI focuses on using insights from behavioral economics to develop, test, and scale new approaches to financial products, pricing, marketing, and education for low- to moderate-income American families.

Table of Contents

This toolkit is targeted at researchers seeking to conduct randomized controlled trials with financial service providers. Despite this focus, many of the lessons in this guide are applicable to randomized evaluations for a broad variety of topics. The ***Introduction*** *section covers:*

1. **Motivations for the Toolkit**
2. **Audience**
3. **How to Use the Toolkit**

Introduction

In recent years, the influx of available consumer data has presented corporate firms, non-profit organizations, and governments alike an opportunity to increase the efficacy and targeting of their products and services. Doing so, however, is not always straightforward; even companies that dedicate significant resources to data analytics have trouble interpreting results or translating results into action. The key to identifying what works is to build experimentation into the design of products and services. Rather than making decisions based on intuition or tradition, organizations that run simple experiments are able to accurately measure the impact of their products and services and refine their actions accordingly.

Randomized controlled trials (RCTs) are the best method for impact testing. Even organizations that lack the capability to conduct sophisticated data analysis can still effectively run and interpret the results from RCTs. While many organizations can conduct randomized experiments on their own, there are added benefits to collaborating with external researchers. Researchers can help sort through some of the nuanced experimental design choices, and they lend objectivity and credibility to the work. Perhaps most importantly, researchers bring the creativity and experience necessary to identify the causes and mechanisms through which products and services can impact users.

Motivations for the Toolkit

Innovations for Poverty Action (IPA) works at the intersection of academics and practitioners. In this toolkit, we have compiled many of the best practices for running randomized controlled trials (RCTs) that IPA has developed over the past decade. We focus specifically on using RCTs to develop and test new financial products and product features for consumers in the United States. We focus on this for two reasons:

First, to our knowledge, there is no guide to running RCTs in the finance sector. Finance related RCTs are unique in that they seek to not only answer questions related to improvements in financial well-being for consumers, but also to identify how these solutions can improve a financial institution's bottom line. In this toolkit, we discuss the development of partnerships between researchers and organizations that are profit-oriented as well as those that are social mission-oriented.

Second, there are aspects of implementing RCTs that are specific to the financial services sector in the US. These include challenges related to the strict regulatory environment in the US (relative to most developing and emerging countries), and issues related to the collection and use of financial data. Our hope is that, by highlighting many of the potential pitfalls involved in managing this kind of RCT, we will help others reduce the impacts of these pitfalls and encourage the broader use of RCTs.

Audience

We have targeted this toolkit at researchers seeking to conduct RCTs with financial service providers. By researchers we mean both established and junior academic researchers (including graduate students, post-doctoral researchers, and junior faculty), research staff at organizations similar to IPA who may be conducting research independently or under the supervision of a senior researcher, or research-minded practitioners interested in experimenting and evaluating their own products and services.

We chose this audience, rather than focusing exclusively on practitioners, because while there are excellent existing technical guides to running RCTs, there is a dearth of resources containing deeper insights into the development of research partnerships and the specifics of experimenting with financial products and services.[1] Clearly the technical components of RCTs are extremely important, but the "softer" skills of managing an RCT, including the logistics of implementation and the interaction between the researcher and the partner institution, are equally central to the success of the experiment. Thus while we do detail the technical considerations of designing RCTs, this toolkit assumes a certain level of research design knowledge and data skills. Where appropriate, we include links to further resources for those who would like to read more in depth.

Although this toolkit is focused on experiments with financial products and services, many of the lessons in this guide are applicable to randomized evaluations for a broad variety of topics and to a broad variety of people. Our aim is to make this a living document that

[1] Excellent technical guides to running RCTs include Rachel Glennerster and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide*; Alan S. Gerber and Donald P. Green, *Field Experiments: Design, Analysis, and Interpretation*; Esther Duflo, Rachel Glennerster, and Michael Kremer, *Using Randomization in Development Economics Research: A Toolkit*; Paul Gertler et al., *Impact Evaluation in Practice.*

compiles tools, experiences, and advice from among the many organizations that engage in this work. We welcome your comments, ideas, and contributions as we continue to improve and expand on this document. Our overall goal is to make RCTs more accessible to all, and increase the use of rigorous impact analysis and improving policy and social services.

How to Use the Toolkit

The toolkit opens with a brief refresher on RCTs. We then discuss developing a new research partnership, creating a research design, preparing to launch a project, collecting data, managing a project once it is off the ground, and finally, wrapping up.

While the structure of the toolkit roughly follows the chronological steps involved in running an RCT, each section is meant to give a more conceptual overview, covering the various aspects to take into consideration at each stage without prescribing *when* each should take place. Depending on the project, some aspects of research design may happen, for example, long before you start conversations with potential implementing partners, or vice versa. While each section can be used by itself, we recommend reading the entire toolkit before launching your project, as each section will provide greater context for the other sections.

Throughout this guide, we reference additional tools and documents that we have found to be helpful for various aspects of designing and running RCTs. These are compiled at the end of the toolkit, in the Additional Resources section. We have also created some templates for common project documents and agreements, which are featured in the Appendix.

The ***Why Randomized Evaluation*** *section briefly covers why randomization is important and how to randomize in the context of an evaluation.*

Why Randomized Evaluation?

In recent years, RCTs have gained a great deal of traction in social science, in large part because of their elegance and simplicity. The ability to use controlled experiments to accurately measure the impact of a program has also changed the dialogue around public and philanthropic spending, shifting the focus from the sheer number of people touched by a program to a focus on a program's ability to achieve measurable improvements in well-being.

The following section is intended to be a quick refresher for people who are already familiar with RCTs. It is not intended to be comprehensive. For those who want a simpler explanation, we recommend checking out the "What is a Randomized Controlled Trial" handout, included in the Additional Resources listed at the end of the toolkit.

Why Randomize?

Randomized controlled trials (RCTs) most rigorous methodology available to evaluate what works in fighting poverty. Also known as randomized evaluations, RCTs are considered the gold standard of program evaluation because they produce the most accurate results. Other evaluation techniques such as propensity score matching, difference-in-differences, or multiple linear regression require researchers to make large—and untestable—assumptions about how well they have controlled for any confounding variables (see Table 1 on page 12 for a comparison between RCTs and other evaluation methods).

The largest problem in claiming a causal impact of an intervention on a set of outcomes when randomization isn't used is selection bias. Selection bias occurs when the group of people receiving a program or service differ in some way, either observable or not, from the people who do not receive the intervention and against whom outcome measures are evaluated. For example, if small-dollar loans are offered only to people whose incomes fall below a certain cutoff, and their outcomes are compared to those who don't fall under that cutoff, we would expect that the results would pick up both the impact of the loans (if any) plus any effects of being lower income.

Randomization solves this problem by assigning people to treatment and comparison (also known as control) groups independently of any individual characteristic. By randomizing group assignment, we are able to ensure that the two groups do not differ in any meaningful way. This means that, on average, both the treatment and comparison groups are the same on all observable characteristics (e.g., same gender composition, same average income, and same average age), and on all unobservable characteristics (e.g., personal ability, internal motivation, or other factors that cannot be measured). Therefore, any measurable differences in outcomes after the test group has received the product or service can be attributed to the intervention itself, rather than to something inherent to the recipients, or to some other external factor.

How to Randomize

The first step in RCT design is to identify the study population. It is worth noting that, while many people assume that the population itself should be randomly selected to ensure that it is representative of the population as a whole, this is not in fact a requirement for an RCT. Random selection impacts the external validity of the study—how generalizable the study results are to other people or settings—but not its internal validity—or, how confidently we can conclude that there is a causal relationship between the dependent and independent variables.

Random assignment is the core of RCT methodology. For example, in a study looking at the impact of a new credit scoring model on the risk profile of potential borrowers, the study population might be all eligible borrowers. This study population is determined by the target population of the implementing partner and the researcher. Once the study population is determined, participants are randomly assigned to the treatment and control groups. Randomization can be as simple as flipping a coin or drawing names out of a hat. In some of our RCTs, IPA has held public lotteries in order to make the selection process transparent. In others, we use a computer program to assign individuals to one group or the other.

Criticisms of Randomized Controlled Trials

Randomized trials have been subject to quite a lot of criticism over the years. RCTs cannot and should not be used in all situations, but when feasible, RCTs are the best tool for evaluating program impacts. Below we give some of the most commonly heard criticisms as well as some responses to them, in the hopes that this will help address many of the concerns about randomization that are often raised.

TABLE 1: METHODS COMPARISON

	Methodology	Description	Who is in the comparison group?	Required assumptions	Required data
Quasi-Experimental Methods	Pre-post	Measure how program participants improved or changed over time	Program participants (before receiving the intervention)	The program was the only factor influencing any changes in the measured outcome over time	Before and after data for all participants
	Simple difference of means	Measure difference between program participants and non-participants after the program is completed	Individuals who didn't participate in the program but for whom data were collected after the program	Non-participants are identical to participants except for program participation, and were equally likely to enter program before it started	After data for all participants
	Difference-in-differences	Measure change over time of program participants relative to the change of non-participants	Individuals who didn't participate in the program but for whom data were collected	If the program did not exist, the two groups would have identical trajectories over this period	Before and after data for all participants
	Multivariate regression	Individuals who received treatment are compared with those who did not, and other factors that might explain differences in the outcomes are controlled for	Individuals who didn't participate in the program but for whom data were collected. Data includes other explanatory factors (covariates)	The factors that were excluded do not bias results because they are either uncorrelated with the outcome or do not differ between program participants and non-participants	Outcomes and control variables for all participants
	Statistical matching	Individuals in control group are compared to similar individuals in treatment group	**Exact match:** for each participant, at least one non-participant who is identical on selected characteristics **P-score match:** non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants	The factors that were excluded do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants	Outcomes and variables for matching for all participants
	Regression discontinuity design	Individuals are ranked based on specific criteria. There is some cutoff that determines whether an individual is eligible to participate. Participants are then compared to non-participants and the eligibility criterion is controlled for	Individuals who are close to the cutoff, but fall on the "wrong" side of that cutoff, and therefore do not get the program	After controlling for the criteria, the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary but sufficient requirement for this to hold is that the cut-off criteria are strictly adhered to	Outcomes as well as measures on criteria (and any other controls)
	Instrumental variables	Participation can be predicted by an incidental factor, or "instrumental" variable, that is uncorrelated with the outcome, other than the fact that it predicts participation (and participation affects the outcome)	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate	If it weren't for the instrumental variable's ability to predict participation, this "instrument" would otherwise have no effect on or be uncorrelated with the outcome	Outcomes, the "instrument," and other control variables
Experimental Method	Randomized Evaluation	Experimental method for measuring a causal relationship between two variables	Participants are randomly assigned to the control groups	Randomization "worked." That is, the two groups are statistically identical	Outcome data for all participants

ETHICS

Because RCTs require that participants be randomly assigned to receive an intervention or be in a control group, one of the most common criticisms of RCTs we hear is that it is unethical to deny services to people who need them. There are two principal rebuttals to this. First, to state that denying services is unethical assumes that the service has already been shown to have a positive impact. In cases where a positive impact is known and conclusive evidence exists, then it is indeed unethical to deny that service in the interests of research if it is possible to provide it to everyone who needs it. However, most products or services have not already been proven to work. It could be the case that the intervention has no impact, in which case the funds supporting it could be better used elsewhere; or, the intervention might even be causing harm, in which case denying the service might even be better than providing it.

Second, it is often the case that there are not enough resources available to provide services to all the people who need them. As a result, it is impossible to avoid denying service to some people. Under resource constraints, randomization—essentially a lottery—can be a more fair and transparent way to choose who will and will not gain access to the services. Indeed, some of the best "natural" experiments have occurred when practitioners chose to implement a lottery to fairly allocate scarce resources.

EXTERNAL VALIDITY

The external validity of a study refers to how broadly the results of the study can be generalized. A study demonstrating that group-based microlending is effective in rural India might not carry a lot of weight with a credit union in Indiana. We often hear concerns that investing in an RCT of an intervention in one context is a waste of resources, since we don't know how relevant the results will be outside that chosen context. While there is truth to this claim, we would argue that this is a problem with *all* research, not just RCTs.

To know if a program or service truly has a positive impact, it is necessary to conduct multiple studies, in multiple places. When possible, we highly recommend the replication of existing studies to determine if the results hold, and if so, under what conditions. Along these lines, RCTs are also criticized as being atheoretical, but rich RCT designs—and in some case even simple ones—can be used to test, inform, and refine theoretical models from various fields; this is key to developing an approach that utilizes evidence in a way that is mindful of local contexts.

COST

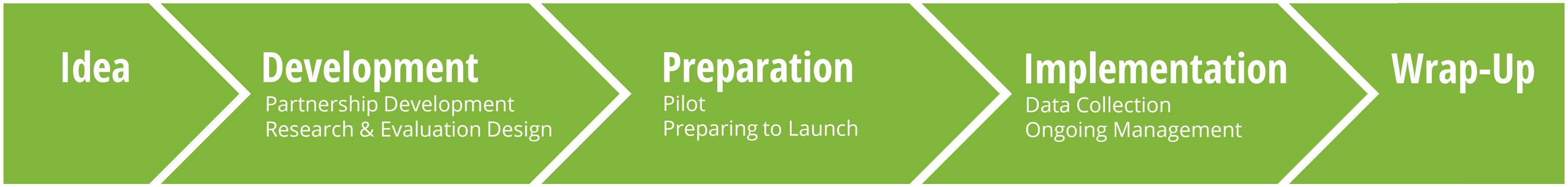
RCTs are known for being expensive. It's true that RCTs can be more expensive than some other evaluation methods, but they don't necessarily have to be, and it's important to distinguish the cost of the RCT from the cost of data collection more generally. Any research requires data collection costs, and the marginal cost of randomization may actually be quite low. The sample size necessary for an RCT is often lower—and the data collection costs are therefore lower—than that of quasi-experimental quantitative evaluation tools. Lastly, it's important to measure the costs of an RCT against the benefits. Billions of dollars are spent on delivering unproven products and services—whether the metric for "unproven" is profitability, social impact, or some combination of the two. So the potential savings from investing based on evidence are enormous. Since the ultimate goal of research is to provide evidence that can be used to shape better policies, the better the research—in terms of both internal and external validity—the greater its value.

The Life Cycle of an RCT

An RCT is typically conducted in four stages: Development, Preparation, Implementation, and Wrap-Up.

- **Development** is the process of taking an initial idea, fleshing out its design, making sure partners and researchers are all on board, and securing funding for the project. Because this is such a large topic, in this toolkit, we have divided it into two sections: the Partnership Development section, which focuses on getting partners, researchers, and funders on the same page and vetting the feasibility of the proposed research, and the Research and Evaluation Design section, which focuses on developing the intervention.
- **Preparation** work includes administrative details, such as signing a Memoranda of Understanding and obtaining Human Subjects Approval, and developing materials needed for the intervention. It also includes developing surveys and any other data collection instruments, and training implementing partner staff and surveyors in how to use them. Lastly, it includes piloting your intervention and any randomization or survey protocols you plan to use, in order to ensure that you've ironed out all the details before the launch. These topics are covered in the Preparing to Launch, Data Collection, and Pilot sections.
- **Implementation** refers to the process of putting your RCT into effect. This includes monitoring your project to make sure that the intervention and the research protocols are being implemented as planned, keeping track of any feedback from study participants, ensuring that all data coming in is clean and usable, and re-training staff as necessary. It also includes making sure that your partner stays engaged with the project through regular communication. This is covered in the Ongoing Management section.
- **Wrap up** of your project includes data analysis and dissemination of results, as well as tying up loose ends with your implementing partner and ensuring your datasets are in line with research transparency guidelines. This is covered in the Wrapping Up section.

FIGURE 1: RCT LIFE CYCLE



*The **Partnership Development** phase of an RCT involves starting with an idea, determining if the idea is feasible, and making sure that both implementing partners and funders are on board.*

Partnership Development

The overall goals of partnership development are to (1) flesh out the design of the research and intervention, (2) determine if the project should move forward, and (3) set the project up for success if it does. This section focuses on goals (2) and (3) of partnership development, since goal (1), fleshing out the project design, is large enough to deserve a section of its own. We start by discussing how to identify a potential partner and initiate a conversation about your research idea. Next we delve into the details of assessing the feasibility and advisability of conducting the proposed research. This includes determining whether the proposed intervention is ready for evaluation (or might need further piloting), gauging partner capacity and fit, and understanding the legal implications of both the proposed intervention and the evaluation (randomization and data collection). Lastly, we touch on some of the things to consider in terms of financing and budgeting for your project.

Getting Started

The initial stages of a project can develop in a few different ways. Sometimes researchers have an idea and need to find a partner to test it with. Other times, researchers are approached by potential partners interested in addressing a specific need or question. It can be challenging to identify potential partners, but typically a good way to frame your search is to seek out organizations that are interested in research and innovation in general.

IDENTIFYING A PARTER

We typically identify partners in one of the following ways:

- **Check with organizations that award funds for research in your area of interest.** Practitioners interested in research have likely previously received grants for their own projects, and are therefore listed as awardees on grant websites. Even if a partner does not have the bandwidth for or interest in joining your project, they might be able to point you to a similar organization that does. If your project is already funded, then your grantor might be able to provide contact information for organizations they have previously worked with or know want to participate in the research.
- **Network at conferences and events.** Partners might be attending panels that are related to your research, or they might be presenting innovative products they have developed through past research partnerships. Even if you do not have a specific research question in mind, simply sharing your research with a partner and inquiring about their organizations' present goals might spark a mutual interest in answering a question or testing a product.
- **Review published papers in journals.** Partners that have previously been part of research projects are sometimes listed. Even if they are not explicitly listed, the researchers on the paper might be able to share more information about the organization or provide a reference to similar organizations.

INITIAL CONVERSATIONS

After reaching out to a partner that seems interested, the development process for an RCT can begin quite informally, often through a brief phone call or set of emails to learn more about a potential implementing partner's existing product offerings, size, and interest in research. In these early conversations, the most important thing to decide is whether or not the idea and the partner are worth pursuing; the questions you ask at this point will depend on your research goals and potential partner, but in general you want to keep in mind:

- **If the potential partner is interested in a specific topic, is this something that fits within your or your team's research agenda?** If not, is this organization willing to experiment with something other than this particular idea?
- **Is this organization willing to randomize access to their product or service?** An answer of "no" is not necessarily a deal breaker at this point in the conversation, since it may be possible to design a randomization strategy that meets the partner's needs further down the line (See the Research and Evaluation Design section for more detail). However, it is important to assess the partner's willingness to randomize at this stage.
- **How many people does the partner serve or expect to serve?** For an RCT, you will need to have a sufficient sample size in order to be able to draw any conclusions about the intervention (see the Research and Evaluation Design for more detail).

CONCEPT NOTES

Concept notes are short documents designed to give your potential implementing partner a brief overview of the research you want to conduct. Drafting a concept note can be a helpful exercise in thinking through the idea you want to test and how it might actually be implemented. It also gives your contact at the partner organization something that can be shared with other stakeholders to build buy-in and facilitate the decision to move the project forward. In general, a concept note should include:

- Motivation for the proposed research project—what problem does the intervention seek to solve?
- A description of the intervention.
- A description of the research questions that will be addressed.
- Background on your research team—who you are, what your approach is, and what your experience has been.
- Information on how the proposed intervention would fit within the organization's mission.

See the box below for additional tips on composing a concept note.

Project Feasibility (To RCT or not to RCT?)

Once you and the partner organization have expressed interest in working together, it's time to start digging deeper to evaluate the potential of the proposed intervention and of the partnership. There is no clear line between this deeper vetting and the initial conversations described above. In many cases some of the issues discussed below come up earlier in project development, while in others they come up later. The following is intended to give you a sense of the best practices and potential pitfalls to look out for. Setting expectations for all parties is really important at this stage. Being aware of potential conflicts early on can help you avoid wasting your and the partner's time on projects that will not be successful. But even more importantly, for projects that do launch, knowing about these conflicts in advance can help you to mitigate them.

VETTING THE PROPOSED INTERVENTION

Typically, we want to conduct RCTs on products and services that:

TIPS FOR CONCEPT NOTES

Be brief! Financial institutions are used to quick and pithy pitches. In general, your concept note should be limited to two pages, or one page per product if you are pitching multiple ideas at once.

Know your audience. A pitch to small non-profit will typically differ from a pitch to a large bank. For example, you would likely emphasize the social value of your idea to a non-profit, while you might want to focus more on the business case for your idea with a commercial bank.

Research your partner's context. Vocabulary is important. Credit unions, for example, have "members" rather than "clients" or "customers." Using terms that don't make sense or aren't appropriate for that partner's work betrays ignorance and can inadvertently communicate disrespect.

Skip the jargon! Don't assume that your partner knows what an RCT is, and make sure that the terms you use are easy to understand.

Remember the business case. For every intervention the goal should not only be achieving desired outcomes (such as improved financial health), but also developing a product business case for the partner. Generally a good business case includes: (1) the purpose of the project, the opportunity it addresses, and its benefits; (2) the strategic fit within the business; (3) risks; (4) affordability; and (5) value for money. See the table on page 18 for some business case pitches that partners are typically interested in.

- **Have already gone through piloting.** If the logistics of offering the product or service have not been ironed out, or the content of the intervention is still subject to ongoing iteration, it can be really difficult (if not impossible) to conduct an RCT. For more information on piloting the intervention, see the Pilot section of this toolkit.
- **Yet are not so established that they cannot change.** If a product or service is already at a large scale and is so set in stone that changing it in response to the results of the evaluation would not be possible, then it is not a good candidate for an RCT.
- **Have reached or will be able to reach sufficient sample size.** Especially when the product has only been through a small pilot, it is important to find out how your implementing partner plans to reach the scale necessary, and over what time horizon. In the Research and Evaluation Design section of this toolkit, we discuss considerations for sample size in more depth.

In addition to these considerations, it is important to ensure that your proposed intervention and randomization are legal. Financial Institutions in particular are subject to strict oversight, so it is good to make sure early on that your study is not violating any laws. The pullout on page 23 highlights some of the laws you should be aware of if you are planning to test a financial product or service.

TABLE 2: BUSINESS CASE

Business Pitches (This product will...)	**Example questions to consider**
Build consumer loyalty	What is the value of increasing customer retention to X%? What is the value in terms of (a) time lost and (b) lost revenue for a lost customer, to replace a lost client?
Change consumer behavior	What is the value of increasing customer referrals? What is the value of decreasing defaults by X%? What is the value of decreasing charge-offs by X%? What is the value of decreasing debt-servicing costs by X%?
Grow “wallet share” (or products per customer)	How many new accounts will be opened? How many additional deposits will be made? How will account balances change? How many new loans will be opened?
Attract new customers	How many additional cross-selling opportunities are created if (financial institution) has X additional touch points each month? What is the cost/benefit of acquiring each new customer currently? Will that change with the intervention?

VETTING THE PARTNER

Even if the intervention agreed to by you and the partner seems perfect, it may turn out that the partner is not the right fit for the evaluation. Projects can end up stalling for many reasons; management is excited about the intervention but frontline staff are not bought in; a key person on the project has an issue with randomization; or, the partner isn’t able to handle evaluation-related data requests. The following list offers some things to look for as you are developing your project. One potential red flag is probably not enough to derail the evaluation, but if any of these are particularly severe, or you are facing several at once, you may want to re-think partnering with this organization.

- **Is there a good fit between the partner and the intervention?** At one organization where we were evaluating a small-dollar loan, it turned out that staff did not want to offer the product because it had a higher interest rate than other similar products, and staff wanted their clients to save money. Other kinds of product cannibalization can lead to significantly reduced uptake of the product or service you are offering. It's important to assess the existing product offerings at the partner institution, to make sure the intervention you are proposing doesn't inadvertently conflict.
- **Will the partner be able to achieve the required sample size?** You and the partner should be clear on the expected sample size and plans for reaching it. It is also helpful to decide in advance at what points along the way you will check in and change course (e.g., increase marketing efforts, or abandon the study) if the sample does not seem to be getting to where it needs to be. Does the partner have access to the right population for you study? It is also important to consider whether the population you have access to through the partner is the "right" one. For example, if you are focusing on the poor, you might want to consider whether the available population is below your income threshold.

CASE STUDY: A LESSON IN SAMPLE SIZE

We worked with a credit union on an evaluation of a loan for low-income consumers. We proposed to roll the product out to 2,000 individuals over the course of 18 months. However, several months into the project, only 50 people had taken up the loan. A look at the credit union's loan volume for previous years explained the problem: the total number of loans originated across all the credit union's products (not just the ones in our evaluation) was less than our projected sample, meaning that reaching 2,000 people for our study would have required significant changes to operations. This was a failure of communication on the part of both the research team and the implementing partner. Ultimately, we were unable to reach the necessary sample size and, therefore, we did not complete our planned analysis. A quick look at the credit union's historical data before the start of the project could have helped shape expectations about sample size and pointed out a big drawback of evaluating the product.

- **Is the partner able and willing to randomize?** Some organizations have systems in place that make randomization very difficult and/or require changes to how the product or service is offered in order to make randomization feasible. While it is sometimes possible to get around this by having randomization conducted by the research team or by adjusting the research design, it's still important to be aware of your partners' concerns about the logistics of randomization. Not all potential partners are willing to randomize, especially if it means denying services to some of their clients. While there are evaluation designs that don't require denial of service, many partners will still have concerns about the impacts of the evaluation on their clients and on their public image.

  Depending on who is involved in the implementation of the RCT, it may also be an issue if managerial staff are willing to randomize, but frontline staff don't see the value of randomization and either complain about it or circumvent it. You should do your best to talk to people at all levels of the organization about the purpose of the evaluation and why randomization is important. If it is too difficult to get frontline staff on board, you could also design the randomization so that it is invisible to the frontline staff, and only a few key people are involved in implementing the randomization. It is also crucial to pilot the randomization process, as this will help you identify these possible sources of friction.

CASE STUDY: RANDOMIZATION LOGISTICS

One of our projects sought to look at the impacts of different savings add-ons to a small-dollar loan. Ideally, the loan officers (LO) issuing the loans would have been able to randomize and offer the appropriate add-on. However, due to limited ability to train and oversee the LOs, we had the loan underwriters, who sat at the lender's headquarters, conduct the randomization. This slowed down the disbursement of the loans, since the LOs could no longer make decisions on the spot. It also required the decision making to be more centralized than the organization's typical operations. This convoluted process had implications for the external validity of our results.

- **Does the partner have sufficient staff capacity?** Running an RCT can be onerous for the implementing partner. Even so-called "low-touch" RCTs can require large time investments that the partner may not have anticipated. The partner should have at least one person who serves as the primary point of contact for the project and who has the bandwidth to support it. If multiple people (e.g., tellers offering products) are involved, time should be carved out for them to support research-related tasks. This may mean adding implementing partner staff time to your evaluation budget to offset the costs to the partner of participating in research.
- **Does the partner have the technical skills and data systems necessary?** As early as possible, find out (1) what data the partner has, (2) who the primary person responsible for accessing data is, and (3) who can access the data if the primary person is not available. Even large, successful financial institutions often have systems for keeping track of administrative data that are clunky and difficult to use.

  Request a sample batch of data from the potential partner both to see how their systems operate and to see what the data itself looks like. Sometimes the data aren't available onsite and must be requested from an external database provider. This can present large challenges to getting the data you need, as any errors in the data can take a long time to correct if you have to wait for one person (who is often busy with many other tasks) to respond and fix the query. Sometimes the data you want may be in a format that is very difficult to use. For example, we worked with one partner who was only able to send us the credit report data we needed via PDF, which meant a large time investment on our end to convert the PDFs into a structured dataset format.

  In other instances the partner may have the proper data systems in place for your study, but there may only be one person who knows how to run the queries necessary to provide you with the data you need. Talk to the "data person" or IT department early on and get a download of a data report as soon as possible.
- **Can the partner send you the data securely?** When dealing with financial data in particular, it is crucial that the data be protected while it is in transit from your partner's server to yours. This can present challenges if the institution in question is used to transmitting data only internally and is unfamiliar with software for data encryption and transmission. While fortunately we've never had a data security breach, we've had some close calls, and no one likes the idea of Social

Security numbers and bank account information floating unencrypted over email. At the other end of the spectrum, though, are partners whose security protocols are so tight that transmission of data becomes virtually impossible. For example, data from tax returns is guarded very carefully by the IRS, so working with tax preparation service providers can be prohibitively difficult, depending on what you are trying to do.

- **Are there any major upcoming changes to the organization and/or its management that could disrupt the project implementation?** In one of our evaluations, the entire database for processing financial transactions was due to be switched just a few months in to the start of the research project. This made data collection risky, as there was no guarantee that data wouldn't be lost in the transition, and also meant that staff had more limited capacity than usual. Management changes can also impact the organizational buy-in for the project. While these changes can be overcome, it is important to be aware of them in advance and be able to plan appropriate strategies to overcome the potential negatives.
- **Does the intervention require new systems or capacity?** If the proposed intervention requires new data systems or a new module to the IT system, for example, or if it means that staff will be taking on new work, is the partner willing and able to do this? Also, it is important to make sure that there is a budget to both build and support this new capacity throughout the life of the project.
- **Is the partner willing to take action in response to the results of the evaluation?** We have met with organizations that were interested in RCTs as long as it would reinforce what they already "knew"—i.e., that their program was effective—but when results came back showing zero or negative impact, refused to believe the data. Not only does this mean that your evaluation has very little chance of having an impact on actual organizational policy, but it can also mean that the organization could try to stop you from publishing results that cast doubt on their model. When vetting an organization, it is crucial to state up front that the organization needs to be prepared for possible negative results as well as positive ones, and to protect your ability to publish by setting up legal agreements in advance.
- **Is there someone who is willing to be a project champion?** Because RCTs can take a long time to run and be a burden on staff, it can help hugely to have a "project champion" who works for your implementing partner. The champion should be someone who is excited about the project and will be able to motivate other staff at their agency to stay engaged as well as to stick to timelines and deliverables. Ideally this person would be someone who is senior enough that they are able to influence others, but not so senior that they have too many other projects on their plate.
- **Is there buy-in from all levels of staff?** It is important to make sure that any staff involved in either the offering or administration of the product or service are bought in to the research proposal. In one project, we asked loan officers to administer a brief survey to borrowers before closing the loan. Partway through the project, we discovered that the loan officers were given financial incentives for closing the loans quickly, and our survey interfered with that process. Not only were the loan officers reluctant to do extra work, but they also had a financial incentive to skip the survey.

BUILDING INTERNAL DATA COLLECTION SYSTEMS

Government officials, decision makers in financial institutions, researchers, and donors are increasingly looking for ways to measure the impact of financial products and services. Without the appropriate resources or data to do it well, however, designing a randomized evaluation to measure the impact of a product or service can be quite wasteful. While a randomized evaluation might not always be appropriate, organizations should still strive to develop data collection systems that produce actionable and timely data.

With this in mind IPA launched the Goldilocks Project, an initiative designed to provide guidance to donors and NGOs interested in developing strong monitoring and evaluation systems. The Goldilocks principles are organized around creating credible, actionable, responsible, and transportable data collection systems, or CART for short:

- **Credible**: Only collect data that accurately reflects what you are intending to measure
- **Actionable**: Only collect the data that the organization is going to use. Will the data collected be used to change the course of action at the organization? If the answer is no, do not collect it.
- **Responsible**: Match data collection with the systems and resources your organization has to collect it. Think about the available resources. Don't overreach, as doing so could compromise data quality.
- **Transportable**: Apply what you learn to other programs and contexts—either your own program in future years or locations, or those of other organizations working on similar problems.

For more information about the Goldilocks Project and "right fit" monitoring and evaluation, visit IPA's website.

In order to assess this, we recommend that you map out how the product or service is offered and how many touchpoints there are with different staff, then talk to the staff involved about why they would or would not want to offer the product or service and implement any research-related tasks, like surveys or randomization protocols.

- **Is the partner responsive?** Partnership development is the most important time to assess whether your partner is going to be responsive. Partners tend to be most excited about the project at the beginning. If you're having trouble getting them to respond to you and give you the information you need now, then it is likely that in a year from now, they will be even less responsive.

Project Finance & Budgeting

Finding funding for your RCT is one of the most crucial aspects of launching a project. There is no clear path to securing funds; sometimes the money comes before the research design or the partner, and sometimes it comes far after. Creating a reasonable budget, receiving early feedback on the proposal, and applying widely to grants can help secure funding.

Is it Legal?

Financial institutions are subject to strict oversight from state and federal governments. If you are proposing a new evaluation to a potential partner, make sure that the intervention and the randomization are legal. Failure to acquire the appropriate sign off early on can lead to project delays and cause headaches when you are trying to launch. One way to avoid this is to have a call with your partner's compliance or legal team during the partnership development phase to talk through your plans. It is also helpful to be familiar with common legal pitfalls in the types of work you are proposing. Some examples from our experience include:

Gambling Laws: When conducting surveys, we sometimes offer an entry into a prize drawing as an incentive to participate. These drawings can be considered lotteries, which is problematic because lotteries are illegal in the US if not operated by the state. To be compliant with gambling laws, we structure our survey incentives as sweepstakes, meaning we must offer a way to enter the drawing without taking the survey (e.g., by sending a postcard requesting entry into the sweepstakes). While most aspects of sweepstakes are standard across the country, some states have additional laws and requirements so rules may need to be verified by an attorney to ensure state compliance.

Fair Debt Collection Practices Act: The Fair Debt Collection Practices Act (FDCPA) was created in an effort to protect consumers from deceptive and abusive debt collection practices. In one project, we designed a loan feature that included informing family members about participants' debts. We felt confident that receiving permission from our study participants and creating an opt-out option in the program kept us in compliance with the FDCPA's laws surrounding contacting third-parties (in this case, the family members) about borrower's debt. On the advice of the institution's compliance team, however, we also added additional disclosures to the marketing and enrollment materials.

Truth in Savings Act: The Truth in Savings Act established requirements for how banks and other financial institutions disclose information about interest and fees to individuals opening a new savings account. Relatedly, the Truth in Lending Act requires financial institutions to provide disclosures about important terms of credit, such as the cost of the loan (APR, monthly payment) and prepayment penalties. If your intervention involves the creation or marketing of a new savings account or credit product, you will need to allow time to draft the appropriate disclosures.

Federal Communications Commission Regulations: If your evaluation includes text messaging or outreach via phone you may need to become familiar with telecommunications laws as the Federal Communications Commission (FCC) has strict rules about calls or texts sent to consumers' phones. For example, telephone solicitation calls to residential homes are prohibited before 8am or after 9pm. If the text or call is considered a "commercial text", senders will need to obtain consent from customers before initiating SMS contact.

Even unexpected laws can be a barrier. For example, in one of our projects, we gave people the option to show evidence that they had spent their loan on certain permitted expenses, among which was medical expenses. It turned out that in order to receive this proof, we had to be in compliance with the Health Insurance Portability and Accountability Act (HIPAA) regulations. Make sure to budget time (and money, if applicable) to get the appropriate approvals and sign-offs.

GRANTS

Typical funders of RCTs include universities, research institutions, government agencies, private foundations, and non-governmental organizations. Oftentimes a donor organization will put out a request for proposals regarding a specific research area or question. Regardless of the circumstances, when you are preparing a proposal, it is important to know your funder's priorities and demonstrate how your project meets their objectives. Some organizations are more familiar with the methodology and value of RCTs than others, so it can be helpful to look at other types of projects your potential donor supports. For proposals to donors with less knowledge of RCTs, you will want to make sure that you clearly explain why a randomized experiment will best address the research questions of interest.

Grant proposal requirements will vary depending on the donor; closely review any instructions, be mindful of deadlines, and make sure you plan accordingly. Overall, your proposal should convey the value of your research and demonstrate the feasibility of the project. Be clear about what you will be delivering to the donor. Most donors require their grantees to provide regular reports on project activities. You might additionally commit to preparing an academic research paper to be submitted to a peer-reviewed journal at the conclusion of your research. Remember that if your project proposal is accepted and funded, you will be responsible for meeting the commitments you made.

BUDGETING

When it comes to project costs, people often think of salaries, administrative costs, or the cost of the intervention itself. Although it's sometimes neglected, remember to budget sufficiently for the costs of the data collection. When creating your budget, consider what you will need in order to gather, securely store, and analyze your data as well as how long you plan to collect your data (i.e., your project timeline). Our recommendation: it is always best to overestimate time and costs in a proposed budget. The following are some things to take into consideration when putting together your research budget:

- **Timeline.** What is the length of your project? How long do you plan to collect data for? Upfront project costs (i.e., the expenses you need to cover to get your project off the ground) are often at the top of your mind when creating a budget, so it is easy to forget about the longer term expenses of your project. If you are planning to collect data for two or three years or conduct an endline survey, make sure to include those costs into your budget. Analysis can also take much more time than anticipated, so be sure to build extra data support into your budget.
- **Data collection.** Do you plan to conduct a survey? Running a survey can add significant costs. For example, in one of our projects, we hired a marketing firm to conduct a phone survey. The cost was $2.50 per call plus a $50 per month administrative cost. Once we took into consideration the number of call-backs necessary to reach each person on the phone, the total cost came out to $5,500. If you plan to use data from credit reports, make sure you get a quote from a credit bureau, as the costs of this can be very high.
- **Field costs.** Will your staff need to be on the ground with the partner? Depending on the partner and the intervention, it can be helpful to have a member of the

research team on the ground with the implementing partner for at least the duration of any survey. This requires thinking through the costs of housing and transportation.

- **Partner costs.** Are you offsetting costs for your implementing partner? Depending on the partner, your budget may need to include funds to cover the costs of increased staff time to implement the research portion of the partner's work and/or the costs of adapting data and other systems to the needs of the evaluation. You should discuss these costs with your partner in advance and they should be included in any MOUs or subcontracts. Both MOUs and subcontracts are discussed further in detail in the Preparing to Launch section.

Partnership Development Questionnaire

Included in the Appendix is a checklist of questions that we often ask of implementing partners to start gauging the answers to the questions we have discussed in this section. Depending on your context, you may only need some of them, or you may decide to ask some in a preliminary conversation and others further down the line. While many of these are designed to help you assess risk, others are there to help your research team design the optimal evaluation.

Partnership Development Checklist

- □ Vet the proposed intervention
  - □ Has the product already been piloted?
  - □ Can the product reach sufficient scale?
- □ Assess the partner's interest in, and capacity to conduct an valuation
  - □ Is the partner willing and able to randomize?
  - □ Can the partner send you the appropriate data?
  - □ Is the partner responsive?
- □ Create a 1-2 page concept note for circulation at the financial institution
- □ Finalize the project budget

*The **Research and Evaluation Design** section covers the design of the RCT and how the implementation of the design interacts with the partner institution.*

Research and Evaluation Design

The previous section covered many of the practical elements of developing your project, including identifying an implementing partner, creating a budget, and applying for funding. We now turn our focus to the methodological aspects of setting up a randomized controlled trial.

The research design is the blueprint of your RCT. Developing your research design first starts with defining your research question and formulating a hypothesis to test it. Once you have clarified these, consider the outcomes and indicators you will use to test your hypothesis. After you have a clear understanding of the outcomes you plan to observe, you can begin to flesh out your randomization protocol.

There are extensive resources on the topics we cover in this section, and we provide recommendations for further reading where relevant. Throughout this section, we will refer to a hypothetical case study as an example of how to approach evaluation design (see the box on page 27).

Defining Your Research Questions and Formulating Your Hypothesis

The research design process begins with concretely identifying the questions you and your implementing partner would like to answer. The case study described on the following page could potentially address several questions that might be of interest. For example, we might be interested in learning whether reminders from peers help borrowers make loan payments on time, or if reminders can reduce delinquency rates on loans.

CASE STUDY: PEER REMINDERS ON LOAN PAYMENTS

We are interested in designing an RCT that tests methods of encouraging better borrowing habits among small-dollar loan borrowers. One theory we have is that payment reminders from friends or family members might encourage borrowers to stay on top of their payments and avoid delinquency. We have been in conversation with a mid-size credit union that is interested in working with us to develop a loan that has borrowers elect "peers" to send them a reminder email if they miss a payment. The credit union serves a low- to moderate-income community and has 20,000 members across five branches.

The same scenario could also answer whether repaying the loan on time improves borrowers' credit. Our partner may be interested in learning whether take-up of the loan with the peer feature is an indicator of a customer's credit worthiness. To help refine and prioritize questions you will want to conduct some background research to try to gain a strong understanding of the context of your evaluation, as well as determine how your intervention will work in practice.

Once you have clarified your research objectives and defined your intervention, you will need to specify a hypothesis about how your intervention might have an effect on outcomes. To do this, we suggest creating a Logical Framework model or Theory of Change that traces the possible causal pathways from the intervention to your end goal.

BACKGROUND RESEARCH

To clarify your research questions and define your intervention, it helps to have a strong understanding of the context of your study. What problems face your target population? How does this group deal with these problems? Are there any existing programs (or products or services) in the field that are aimed at addressing some of these problems? How well do these existing programs work? What constraints do these existing programs face, and why? Taking the time to ask these questions at the outset of your research will help ensure that your intervention is addressing a real problem, relevant to your target population, and feasible.

There are several different approaches that social scientists use at this stage in the research process to gain a better understanding of the proposed study's context; both qualitative and quantitative methods can provide useful information. We highlight a few of these below, but we recommend consulting Glennerster and Takavarasha's *Running Randomized Evaluations* for a more in-depth discussion.

A field scan (or market scan) involves determining what similar programs already exist, how they have been implemented, and what has been learned from them. A field scan is of particular importance when designing evaluations around financial products. In the case of the peer loan, a scan would help find out the terms of other small-dollar loans offered in the area (e.g., loan APRs, durations, and amounts) and determine whether our product would be competitive in the market.

A needs assessment is a systematic approach to identifying an unmet need of a specific population and determining the appropriate intervention as a response. During a needs assessment, you might use existing data sources or conduct your own surveys or qualitative interviews. A needs assessment can also serve as a kind of "litmus test" for your planned intervention—it might conclude that there is no problem, the problem is not as high of a priority as previously thought, the timing of the evaluation is not right, or the problem has different causes than those originally anticipated.

A process evaluation is used when researchers are interested in evaluating whether an existing program is being implemented successfully. In the context of financial institutions, this might include, for example, whether participants that ask to receive financial counseling are being called, whether the population receiving the financial counseling is appropriate, and how the counseling is being logistically conducted (tracking of phone calls, meetings, results, and budgets). A process evaluation can also assess when and why a program, product, or service is failing.

Note that we also often conduct process evaluations at the end of an RCT to learn more about how the evaluation itself was implemented. More on this can be found in the Wrapping Up section.

Lastly, a literature review is a compilation of existing research regarding your issue of interest. A thorough review of current literature confirms that the intervention you are interested in testing has not already been studied extensively and helps get a sense for what has already been tested and concluded. It should also help inform the design of your study, highlighting the evidence on which interventions have been effective or not. Additionally, a literature review will help you identify outcome measures and indicators used by other evaluations. Including some of these metrics in your study will allow you to compare your findings to others' and incorporate your research into the broader policy discussion. 3ie provides a helpful checklist for conducting literature reviews.

CASE STUDY: BACKGROUND RESEARCH ON PEER REMINDERS

After further developing our research idea, we have decided to focus on the question of whether peer reminders impact payment behavior in low-income participants. We first conduct a market scan to determine whether the feature currently exists, and we learn that a few credit unions throughout the country offer it but haven't rigorously tested it. In our literature review, we note that there are a growing number of studies in psychology and behavioral economics that discuss the different ways that peers are thought to influence behavior and the mechanisms behind this influence. These studies show that consumers may benefit from "commitment devices," or products with features that allow people to "tie their hands" to a future goal.[3,4] *We thus update our theory to specify the mechanism through which we expect peer support to change payment behavior: we believe peer support may encourage positive payment behavior by acting as a source of accountability, with peers playing the role of "friendly enforcer" to provide an additional impetus for clients to follow through on their commitments.*

THEORY OF CHANGE

After conducting your background research, you should have enough information to create a Theory of Change, or a diagram of the causal chain that demonstrates how the proposed intervention (the input) is intended to produce direct effects (the outputs), which will ultimately lead to the final impact of the intervention (the outcomes). Detailing a Theory of Change is a useful exercise to refine your hypothesis as well as identify the outcomes and indicators that you will measure to test your hypothesis.

[2] Nava Ashraf, Dean Karlan, and Wesley Yin, "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines," The Quarterly Journal of Economics 121, no. 2 (May 1, 2006): 635–72, doi:10.1162/qjec.2006.121.2.635;

[3] Xavier Giné, Dean Karlan, and Jonathan Zinman, "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation," American Economic Journal: Applied Economics 2, no. 4 (October 1, 2010): 213–35, doi:10.1257/app.2.4.213.

TABLE 3: THEORY OF CHANGE EXAMPLES

Question	Expectation
What is the ultimate desired impact of the intervention?	For our case study, our end goal is improved financial health for loan borrowers, which we might define to mean reduced debt, improved credit, or better borrowing habits.
What milestones may indicate that we are traveling toward our end goal?	With the peer loan, outcomes like improved credit will not be achieved overnight. We assume, however, that immediate outputs of the intervention, such as making payments on time and repaying the loan in full, could indicate that borrowers are moving toward our desired outcomes.
How will our intervention directly produce outputs, and what indicators can we use to measure whether or not the program is being implemented successfully?	In our case, we hypothesize that our intervention—encouragements from peers—will make borrowers more accountable for paying their loans on time and avoiding default. To measure whether or not the intervention is implemented successfully, we may want to track the number of offers made for the peer feature, the number borrowers that sign up for the feature, the number of emails sent from peers to late borrowers, as well as the lag time between the due date of the loan payment and when a member receives a peer reminder.

To develop your Theory of Change, it helps to start with the end result and work backwards. In Table 3 above, we walk through some questions that address the desired impact and the mechanisms through which we expect change to occur from our peer loan case study.

Lastly, consider the assumptions that we've made in building our Theory of Change. Pinpointing these in advance allows you to mitigate the risks that these assumptions will not hold true, and also to diagnose problems after the fact if your RCT does not go as planned.

We provide an illustrated example on page 31 of a Theory of Change framework for our peer reminders study. For additional information, we recommend consulting DFID's *Review of the Use of 'Theory of Change' in International Development*.[4] Glennerster and Takavarasha also have a helpful discussion on creating a Theory of Change and present some examples of how it can be used to select outcome measures and indicators.[5]

Determining Outcomes and Metrics

Determining metrics and outcomes of interest is a balance between what data the research team *can* collect and what data the team *would like* to collect. This necessitates that the research team be very clear about the likely impact of the program or product. Your Theory of Change should be a helpful tool used to give you an idea of the outcomes and indicators you will need to measure. As mentioned earlier, it is also good to include outcomes and indicators from the literature, in order to facilitate comparison between studies. A more in-depth discussion of metrics is included in the Data Collection section.

[4] Isabel Vogel, "Review of the Use of 'Theory of Change' in International Development," Review Report (UK Department for International Development (DFID), April 2012), http://r4d.dfid.gov.uk/pdf/outputs/mis_spc/DFID_ToC_Review_VogelV7.pdf.

[5] Rachel Glennerster and Kudzai Takavarasha, *Running Randomized Evaluations: A Practical Guide* (Princeton University Press, 2013).

- **Brainstorm the possible effects of your intervention.** If you have already worked through a Theory of Change, you should have considered the direct and indirect effects of your intervention. As we explained in our example above, we have defined two positive effects of the intervention: borrowers make timely payments each month, and borrowers are less likely to default. Other effects could include paying above the minimum required amount each month, paying off the loan in a shorter time frame, paying less interest overall, or a reduction in the total number of loans charged off.
- **Determine which outcomes you will need to measure each possible effect of the intervention.** In our example, we would need data that gives us how long each participant took to pay off the loan, the size of their monthly payments and interest rate, how often each peer was sent a reminder, how many times each person made late payments (if at all), and basic information about the randomization and logistics of messaging, including who received offers and messages, as well as when the messages were sent and received.
- **Determine what types of data address these outcomes.** For our study, we would need information on payment history and behavior as well as a way to track message logs and reminders.

SHOULD YOU COLLECT SURVEY DATA?

Financial administrative data such as transaction data, account balances, and payment history might provide a bulk of the data needed for an evaluation. However, the nature of household finances in the US means researchers relying solely on administrative data might not have a complete understanding of participants' financial behavior or habits. Formal savings accounts at multiple institutions, for example, or informal (under the mattress) savings, will not be captured by administrative data from a single financial institution. Additionally, some outcomes of interest, such as perceived financial health and well-being, can only be measured through participant surveys. Where administrative data are not sufficient for your evaluation you may need to collect survey data.

- **Choose only a small number of the most relevant outcomes to analyze.** As demonstrated by our example, often you will have several research questions that you would like to address and likely a desire to collect as many outcomes as possible. Prioritize your questions, be deliberate about the outcomes you collect, and have an ex-ante hypothesis for each outcome. This will help you avoid mining your data for significant results after the fact, and will increase the transparency—and therefore the validity—of your results.

Randomization

Once you have a clearly defined intervention and hypothesis, it is time to create an experiment to test it. When designing your randomization, you will need to consider four technical elements: the treatment, the unit of randomization, the sample size, and the method of the randomization itself.

DESIGNING TREATMENTS

How you select and design your treatments is closely related to the outcomes and measurements you use. Any decision about the number of treatment arms and the randomization strategy utilized across arms hinges on the research question that the team is primarily interested in evaluating. Hypothetically, all treatment arms in a given study could be evaluated and achieve statistical power (more on this in the Sample Size and Statistical Power section below), but in practice this is largely constrained by cost, available sample, logistics, variance in the outcomes, and susceptibility to different

TABLE 4: THEORY OF CHANGE

	Need	Input	Output	Outcome	Long-term Goal
Intervention Logic	• There is a high level of demand for small-dollar credit products, particularly amongst people who may not have access to credit cards or other sources of credit due to poor or missing credit histories • Underwriting small-dollar loans is expensive and risky, leading to limited supply and high cost for available products	• Offer small-dollar borrowers the option to name 1-2 peer monitors who will be notified if the borrower does not repay	• Peers remind borrowers to pay on time • Borrowers don't want to pay late because they don't want their peer to be notified	• Borrowers make loan payments on time more frequently • Borrowers are less likely to default • The financial institution spends less money tracking down delinquent borrowers	• Borrowers improve their credit and become eligible for cheaper credit products • Borrower overall financial well-being is improved • The financial institution is able to offer loans more cheaply due to reduced costs
Indicators		• Number of loan applicants who were offered the peer monitoring feature • Number who chose to participate	• Number of peers notified • Number of peers reporting that they delivered the message • Borrowers reporting reasons for not paying late	• Default and on-time payment rates • Administrative costs	• Credit score • Use of other credit products • Administrative and self-reported data on savings and debt • FI loan data
Sources		• Program administrative data	• Program administrative data • Surveys with borrowers and peers	• Program administrative data • Cost data from FI	• Credit reports • Administrative data • Surveys • FI reports
Assumptions		• Loan officers make the appropriate offers • Loan applicants are interested in the feature	• We have correct contact info for the peers • Loan officers call the peers • Peers deliver message when told to do so		• We are able to reach borrowers for surveys

forms of bias. Your treatment arms should directly produce the outcomes of interest, but perhaps less evident is how they should interact with your partner institution and implementation.

Include the partner in discussions of how many treatment arms will be tested so that they can help you consider the logistics of actually implementing multiple treatment arms. We have found that when frontline staff (such as bank tellers or loan officers) are in charge of making randomized offers, having more than four possible offers becomes difficult to keep track of; staff get confused and make incorrect offers, and the cost of training and re-training is high.

Broadly, there are two situations in which you can include many treatment arms and still ensure that your implementation is successful (assuming your sample size is large enough and your research design is such that you can introduce multiple arms without sacrificing power):

- **If the implementation itself is relatively low-cost, in terms of both time and effort.** For example, in our peer support study we are also interested in the effect of reminder messaging on loan payment behavior. We would like to send two different payment reminders to participants that each have two variations, and then cross-randomize that with assignment to the peer support group. This requires that participants be assigned to one of two peer support groups, and then receive one of four different message types, for a total of eight different possible treatment assignments. A design like this, although complicated in the number of treatment arms, is generally easier to monitor and might be able to accommodate more variations successfully, because the treatments are not widely different and assignment to treatment is not costly.
- **If only one person is needed to implement the design.** The types of evaluations that lend themselves to these designs are typically messaging (SMS) or email interventions, where the treatment can be delivered en masse. Again, it is essential to assess partner capacity in determining whether this is possible—having a couple of false starts of the assignment before the full-scale launch can help gauge the efficacy of implementing many complicated treatments.

UNIT OF RANDOMIZATION

Before determining your unit of randomization, first consider: what is the thing that you will be randomizing? In other words, what is the target of your intervention? For example, you could randomize individuals, households, schools, bank branches, or census districts. There are several considerations to make when deciding what “level” to pick.

- **For starters, it is helpful to think about what your unit of analysis will be.** Randomization should happen at the same level or higher than the level of the outcomes you plan to measure. In the case of the peer loan (and in many other financial product evaluations), we will most likely be interested in measuring individual level data, such as a borrower’s timeliness of payments. As such, it would make sense for us to randomize individuals.
- **You will also want to pick a level of randomization that will best mitigate spillovers**. Spillovers occur when people who do not receive the intervention are still affected by it. For instance, a mailing campaign to encourage debt reduction could impact people who were assigned

to the control group if members of both the control and treatment groups lived in the same house and saw each other’s mail. In this case, we might want to randomize at the level of the household, so that the only way that treatment assignment can affect our study participants is through receipt of the treatment itself.

- **You will also need to consider the feasibility of your desired level of randomization, both from a logistical and ethical standpoint**. Is it fair to randomize? Is it legal? Is your partner willing and capable of randomizing at your desired level?
- **Lastly, you will need to consider statistical power.** Your ability to detect statistically significant effects depends on the size of your sample (the N of your study). This is discussed more in depth in the section below, but as a rule of thumb, the larger the sample, the higher the statistical power. However, randomizing at the level of a bank branch, for example, can lead to a loss of power, as it is necessary to adjust for the intra-cluster correlation between customers at that branch. It is often not feasible to randomize at the branch level because there are not enough branches to reach the necessary power.

SAMPLE SIZE AND STATISTICAL POWER

Consider how many people you should recruit for your study; how many should receive the treatment, and how many should be in the control group? As mentioned above, the size of your sample largely determines the power of your study—the probability of detecting a statistically significant difference in outcomes between the treatment and control groups. A sample size that is too small may make your study “under powered,” meaning that you might not be able to detect an important effect

MDE EXPLAINED

The minimum detectable effect (MDE) is the smallest statistically significant effect size that you can detect at a given level of power, statistical significance, variance in the outcome variable, and sample size. Although the formula for MDE becomes more complicated with more technical research designs, at its most basic, it is a function of t_k (the t-statistic that corresponds to 1-k, where k is statistical power), $t_{\alpha/2}$ (the t-statistic that corresponds to α, where α is statistical significance), P (the proportion of participants in treatment), σ2 (the variance in outcomes or treatment effect), and N (the total number of units in the study, or the sample size).

$$\boldsymbol{MDE = \left(t_{1-k} + t_{\propto/2}\right) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}}$$

This equation has some important implications:

- All else constant, when your population is more homogeneous (lower population variance), the sample size necessary to detect a given effect size will be smaller.
- In general, a larger sample size will allow you to detect a smaller effect size; however this relationship is not linear: the “returns” in terms of gains in power from increasing the sample size are diminishing.
- All else equal, the optimal proportion of participants in your treatment and control groups is 50%.

that actually exists. On the other hand, spending a lot of resources to increase the sample size to the point where you can detect the tiniest of changes may not be cost-effective, if those tiny changes are not meaningful in real terms. To optimize your sample size, you will want to calculate how many study subjects are needed for each treatment group in order to measure your minimum detectable effect (MDE).

In the following paragraphs we touch on some additional considerations to be made when determining the sample size you will need for your study. As with the other areas we cover in this section, there are extensive resources on power calculations. Please refer to our Additional Resources section for more information.

- **What is the expected effect size given the Theory of Change of your intervention?** In our peer reminders study, we expect the effect size of on-time payments to be fairly small; roughly one-third of the members at our partner credit union currently miss at least one payment per loan, and, based on the commitment device literature we reviewed, we expect that the peer reminders will have the effect of increasing the number of on-time payments by at least one per participant. Make sure you're familiar with the related literature and solicit your partners' input to get a sense of what you can realistically expect.
- **What effect size is meaningful to your partner?** On the other hand, spending a lot of resources to increase the sample size only to be able to detect a minuscule effect is probably not worth the effort. A rule of thumb for policy interventions is to use the smallest effect size that is large enough for the scaled intervention to be cost effective. If any impact of, say, less than ten percentage points wouldn't be meaningful for your partner, don't bother powering your study to detect effects smaller than that.
- **What data are you using to estimate power?** It is not always possible to gather baseline statistics on the target population (although this can sometimes be done during piloting, discussed further below), so power calculations are often based on data from populations that are similar to the target. The closer that your data represent your sample population, the more accurate your power calculations will be. It's a good idea to have an idea of what the differences between your target population and the individuals in the data are, so that you can gauge whether your power calculations are likely to be more or less conservative.
- **What level of compliance with the intervention are you expecting?** Power for intent-to-treat analysis can be done quite simply, but if you expect imperfect compliance and want to conduct a treatment on the treated analysis, it's important to adjust your calculations based on the expected uptake of the intervention.
- **At what level are you randomizing?** Although randomizing groups (school, bank, or branch) can make it easier to ensure compliance and reduce spillovers, correlation within the groups (intra-cluster correlation) increases the variance, which can dramatically increase the number of people—or, more likely, the number of groups—you need to include in your study in order to detect the same effect size.

THREATS TO DESIGN

Treatment noncompliance, spillovers, and attrition can all threaten the internal validity of your RCT. Noncompliance occurs when people assigned to the control group gain access to the randomization, or when people assigned to the treatment group do not take up the intervention. Spillovers occur when there are unintended impacts of the intervention on the control group (or other groups that are not part of the study).

Attrition occurs when any outcome data on study participants is lost; this might occur because of people falling out of the study population, either because they have moved (or even died) and cannot be tracked down for further data collection, because they have requested to be removed from the study, or because of researcher error. Issues of noncompliance and spillovers can be mitigated—if not eliminated—by careful research design, so it is important to think through the potential sources of these threats and design accordingly. Unfortunately, there is no easy fix for biases introduced due to attrition. Attrition leads to bias when "missingness" is correlated with the experimental assignment, so if your study suffered from high attrition rates make sure to assess whether the missingness is independent of randomization. Ex-post statistical corrections exist, but all of these require untestable assumptions about the nature of the missing data.[6] When possible, build in extra resources to track down the data from missing participants.

METHODS OF RANDOMIZATION

In an ideal world, we could randomly assign the study population to as many treatment groups as we'd like, ensure full compliance, and eliminate spillovers. However, this is rarely (if ever) the case. Often, simple randomization is not feasible for logistical, legal, or ethical reasons. In the field of consumer finance, for example, fair lending laws (or at least concern about interpretation of those laws) often prohibit financial institutions from denying products or services to qualifying individuals. Therefore, if we want to evaluate a loan product, we cannot randomize subjects into the control group by rejecting loan applications. The following describes some solutions to common obstacles to randomization in the financial sector.

- **Encouragement design.** Typically, it is not possible to force a person to participate in a program or to deny someone the ability to participate. With an encouragement design, the "treatment" that the person receives is not the product or service itself, but merely targeted advertising designed to encourage her to enroll. This might be a design we'd want to use to implement our Peer Reminders study; we can't automatically subscribe people to the feature nor deny them the option if they really want it, but we can randomly assign people to receive an offer. We might want to randomly select half of our participants to receive a verbal encouragement to take the offer and then discourage the other half, either verbally or through some behavioral friction (additional steps to taking the feature).

  By randomizing which participants are encouraged to take the treatment and tracking outcomes for those who do and do not receive the encouragement, it is possible to obtain reliable estimates of the impacts of both the encouragement and of the product or service itself. Although some of the people in the "encouraged" group may not enroll, and some of the people who don't receive the encouragement will enroll, all that is required is that the encouragement increase the likelihood that participants will follow through with what they are being encouraged to do (i.e., that the "encouraged" group be more likely to take the peer reminders feature than the "not encouraged" group).

[6] See, for example, Gerber Green 2012.

- **Phase-in design.** Sometimes donors or partner organizations are unwilling or unable to exclude some clients from receiving their service. One option for testing an intervention of this kind is to phase in the program in stages. The first group of beneficiaries would receive the program in year one, the second group in year two, and so on. In this way, everyone in the community eventually gains access to the program, but in the initial year(s) of the evaluation, the second and third groups serve as the control.
- **Cluster randomization.** Not all programs can be provided at the individual level. As we mentioned previously, the nature of some interventions require that entire branches or neighborhoods be randomized to treatment or control, to mitigate the possibility of substantial spillover occurring. This kind of randomization changes the number of people that need to be included in the study (the sample size), which is discussed in more detail above.

Documentation

Having a written evaluation plan, a document that details your research design, is crucial for ensuring that the study is implemented successfully. Along with an evaluation plan, a pre-analysis plan can be a great planning document, particularly for the data component of the research design.

EVALUATION PLAN

After finalizing all decisions regarding the intervention, treatment arms, and outcome variables, the research team should create an evaluation plan in conjunction with the partner. The evaluation plan outlines timelines, roles and responsibilities, intervention details, and data collection procedures, and acts as the project's central planning document. Walking through the creation of an evaluation plan with the partner organization can help illuminate key challenges and also help solidify the goals of the project. A template for an evaluation plan is located in the Appendix.

An evaluation plan requires clarifying not only the process of randomization and research questions, but also how the theoretical design interacts with the environment in which it is implemented. Failure to consider this interaction while moving from design to implementation can result in unexpected changes to the research design mid-fieldwork, such as needing to drop an entire treatment arm or not being able to audit or ensure the quality of the data. This becomes especially important while working with financial institutions in the United States; each partner organization has its own set of goals, protocols, and metrics that they consider valid that need to be addressed early on and incorporated into the design.

PRE-ANALYSIS PLAN

The pre-analysis plan outlines the technical components of a project. Basics to include are: the type of study to be conducted, the data sources, how the variables are constructed and how they fit together into a dataset, model specifications, and challenges that might arise during the study. The pre-analysis plan helps raise the credibility and transparency of the study. If a pre-analysis plan cannot be written before the evaluation begins, at the very least it should be created before the analysis is conducted. A template for a pre-analysis plan is included in the Appendix.

Gaining Partner Buy-In

While it is crucial to broadly choose research questions that are academically interesting and policy relevant, it is just as important to consider what questions are important to your implementing partner. One of the biggest challenges of managing ongoing field experiments is ensuring and maintaining partner buy-in. If the partner is just as invested in the design and the outcomes as the research team is, the chance of successful implementation and partner responsiveness is much higher.

CASE STUDY: SIMPLIFYING THE DESIGN

We conducted an evaluation of a credit-building product with a financial institution in St. Louis for which we collected administrative and credit report data. Gathering these pieces required copying and pasting survey ID numbers from excel files, re-merging our data using excel formulas, and sending the ID numbers to the credit bureau. This design seems relatively straightforward, but after our baseline data collection concluded, we realized that many survey participants had fallen through the cracks due to errors in the copy-paste and formula process. Even with perfect copy-pasting, this design is not replicable and thus lends itself to errors. An easier way to implement the same research questions might have been to write a program file using Stata, Python, or another software that outputted the required survey IDs each time. This process would have been easier to monitor because the steps would have been documented and verifiable, and generally less susceptible to human error.

KEEPING PARTNERS' GOALS IN MIND

- **Determine what outcomes the partner is interested in.** Part of the initial conversations with your partner should include an assessment of what their short-term and longer-term goals are for the evaluation. The goal should be a research design that benefits both the partner and the research team—this may mean adding an additional treatment to the original design, or reframing a question to fit both groups' interests.
- **In a similar vein, gauge early on what metrics are important to your partner organization and what quantitative tools they typically use to make decisions internally.** These metrics may be different from the outcomes that the research team considers important or useful so researchers may need to revise the design of the evaluation so that both sets of outcomes are included.
- **Be selective about the data you request from your implementing partner.** It can be tempting to find out what data and measurements the partner organization is *capable* of providing and then try to get as many of them as possible, even if they are not explicitly related to the original research questions of interest. It is important to remember that everything you request from your partner comes at a cost, even if additional steps and deliverables do not immediately appear to be burdensome. Randomized evaluations typically span at least one to two years, and partners can get burned out when data collection is onerous, leading to lower quality data. Because of this, it is typically more beneficial to request data on the outcomes that directly address the research questions.

SIMPLIFYING THE DESIGN

When working with an implementing partner always try to remember the KISS principle—keep it simple, scientists! Keep the capacity of your implementing partner in mind as you are finalizing your research design.

- **Include fewer treatment arms than what your partner has the capacity to include.** Toward the beginning of evaluations, partners and researchers are typically more invested and involved than they will be as the study progresses or shifts to ongoing data collection. If the partner determines that their group has the capacity to launch four proposed treatment variations, for example, at the very most three should be launched. There are *always* issues that arise throughout implementation, no matter how straightforward and streamlined the final design ends up being, and it is essential to include room for the partner to help address these issues.

Creating a design that is below the maximum partner capacity also adjusts for decreased partner interest and investment over time; as was mentioned briefly in Partnership Development, it is difficult for many organizations to stay equally invested in the same questions for multiple years, and between staff turnover and other priorities taking precedence, sometimes the evaluation will be last on their list. If you can create a design that adjusts to this decreased attention, the quality of your data and implementation will be much higher.

- **Treat each evaluation as a long-term relationship rather than a one-time experiment.** It might be tempting to add treatment arms and other measures simply because the design is flexible enough to allow it and the partner has expressed their willingness to do so. If the first evaluation is streamlined, planned well, produces metrics partners are interested in, and runs successfully, there is a higher chance that the partner organization will be interested in working with the same team again. A sequential set of well-organized evaluations are easier to manage than one that collects the same outcomes, but has too many moving parts to keep track of and leaves the partner frustrated. Partner relationships aside, the data will be of higher quality and the results more accurate if the research team is able to monitor the evaluation more closely and ensure that implementation runs as planned.
- **Simplify the process of implementing the treatment assignment and data collection.** Always default to logistically easier ways to assign participants to treatments or gather data you are collecting. Errors often arise in what initially appear to be benign steps in the data collection process. Especially if the evaluation relies on people who are not as familiar with the study design to collect pieces of data (e.g., loan officers, tellers, and other implementing partner staff), it's important to make the process as foolproof as possible.

Research and Evaluation Design Checklist

- □ Conduct background research, needs assessment, and literature review
- □ Outline the intervention's Theory of Change
- □ Determine which outcomes you will use to measure the effect of the intervention
- □ Identify the data (administrative, credit report, survey) needed to address your outcomes of interest
- □ Select the unit and method of randomization
- □ Conduct power calculations to determine the study sample and minimum detectable effect size
- □ Codify the research design in the project Evaluation Plan

*The **Preparing to Launch** section covers the tasks that need to be completed in order to get the RCT off the ground.*

Preparing to Launch

Successfully executing your RCT requires advance planning and good project management. To draw an analogy: you can have the best recipe in the world (in this case, your research design!), but if you don't read directions, measure your ingredients, or keep track of time, you'll end up with burnt brownies.

In this section, we discuss the administrative items you'll need for your RCT, a few suggestions for setting a good precedent for partner communications, some tips for training staff and surveyors, and then some additional points about supplemental materials that you may want to include as part of your intervention. At this stage, you will also want to set up your data systems, which we discuss in detail in the Data Collection section.

Administrative Items

Once you have finalized your research design, it's time to start dotting your "i's" and crossing your "t's" to make sure that you have everything ready to go for the launch of your RCT. Before you begin any of your project activities you will want to make sure that (1) all the required legal agreements are in place, (2) your research staff has completed human subjects training and your study has been reviewed and approved by an Institutional Review Board (IRB), and (3) you have registered your RCT with American Economic Association's (AEA) Trial Registry.

LEGAL AGREEMENTS

Once you and your partner have agreed on the evaluation design, the recommended next step is to codify this agreement in writing. The various legal agreements involved in research can sometimes feel excessive if all parties are in agreement about the details of the project. However, the process of executing these agreements can be an important moment to help ensure that you in fact are on the same page. These agreements are meant to protect the privacy of the research participants and both you and your partner if something does go wrong.

- **Non-disclosure agreements (NDAs)** are typically used at the beginning of negotiations—before you even agree to work together—and are used to protect the confidentiality of information that must be shared between two organizations in order to determine whether or not they will work together. This information can include data for power calculations, statistics on product use, or the names of available variables in a particular database.
- **Memoranda of Understanding (MOUs)** are agreements to work together. These typically include a statement of the proposed project to be undertaken jointly by both parties, as well as a scope of work defining the obligations (including any reporting requirements) of each organization. They also include statements regarding the confidentiality and ownership of information and other intellectual property generated during the course of the partnership.
- **Contracts and Subcontracts** are similar to MOUs but are used when funds are flowing from one organization to the other. They include similar language to that of an MOU but also specify the amount of funds, the type of contract (fixed cost versus cost-reimbursable), and the schedule of payments (e.g., in tranches or upon receipt of certain deliverables).

In executing these agreements, you may want to consider the following:

- **Right to publish.** As a research organization, we are committed to publishing the results of our studies, even if the results do not align with the expectations of the implementing organization. This is crucial for the integrity of our research. Allowing the implementing organization to view the content of the results before they are released should not be a requirement, and the legal language in any MOU or contract should generally not restrict publication. The only caveat to this is that we do sometimes sign NDAs that restrict publication; however, we include language specifying that the current NDA will terminate and a new agreement will be developed when a project is funded or the organizations decide to pursue a partnership.
- **Protection of Human Subjects.** We don't want to risk publishing anything that would jeopardize the privacy of any of the clients or other individuals who are participating in the study. All guidelines for transmitting and securing data should be explicitly stated in the MOU or contract. If there are additional legal requirements that protect certain kinds of data (e.g., tax law requires that data not be shared except in the aggregate, at a minimum of 10 people per group), these should also be made explicit.
- **Marking confidential data.** We recommend that your agreement state that confidential information is only that which is explicitly marked as being confidential. This makes it easier to avoid accidental breaches of confidentiality. Data identifying research participants is an exception to this, because it should always be confidential.

- **Freedom to disclose conversations.** Some NDAs require that parties not disclose the existence or content of conversations between them. We recommend tracking these so that staff are clear on what may and may not be discussed and do not accidentally disclose something that is in breach of the contract.
- **Intellectual Property (IP).** Your MOU or subcontract should specify the ownership of all IP generated during the course of the partnership. Typically, the researchers own the IP generated by the research team, the partner owns IP generated by the partner, and any jointly developed IP is jointly owned and may be used freely by both groups.

HUMAN SUBJECTS

As discussed in the following section, Data Collection, almost all studies using human beings as research subjects must be approved by an Institutional Review Board (IRB). In addition to submitting your study for IRB approval, US federal regulations require that any study personnel handling Personally Identifiable Information (PII) have certification in human subjects protection. This requirement typically extends to principal investigators, student investigators, research assistants or other research staff—essentially, anyone who will have direct contact with PII. Certification involves completing an online tutorial on Research Involving Human Subjects. Both NIH Human Subjects and the Collaborative Institutional Training Initiative (CITI) at the University of Miami offer free certification programs online. Each course takes a few hours to complete.

We discuss the details of informed consent, protection of human subjects, and data security in the Data Collection section, but it is important to note that you must receive clearance (either approval or exemption) for your study from the appropriate IRB *before* collecting any data. Additionally, make sure to keep your IRB abreast of any major updates to your research design or intervention. Any serious or unexpected problems with your study must typically be reported to your approving IRB within 48 hours. Reportable instances include, but not limited to: subjects dropping out of study beyond anticipated attrition, adverse responses to the survey or intervention, reports of subject dissatisfaction related to any aspect of the study, and loss of data or hardware housing data.

AMERICAN ECONOMIC ASSOCIATION TRIAL REGISTRY

The American Economic Association's (AEA) Trial Registry is used by academics to pre-register projects before they are implemented. It is meant to help solve the problem of publication bias by providing a place where all trials are registered in advance of their start, make access to results easier and more transparent, and address the growing number of requests for registration by funders and referees. The registration process is brief. Registrants are asked to provide the trial title, country, status of the project, and experimental design. Submitting a pre-analysis plan or power calculations is optional. IPA has made registration a standard operating procedure for all new projects and we encourage other researchers to add their new, ongoing, and past projects to the website. For more information please visit the AEA RCT Registry website.

Communications and Partner Relations

The earlier the research team can set expectations for timelines and check-ins, the better. The pre-launch period is when the evaluation will be its most on-time and is when the implementing organization will be at its most responsive and invested, so setting a clear guidelines for when check-ins should occur sets a good standard. Even if there are no outstanding issues with the project, having a standing appointment on your calendar will help make sure that questions are answered in timely manner and any possible issues are identified and addressed quickly.

Additionally, we recommend that you:

- **Find "champions" (at various levels of the organization) to spread excitement prior to the launch of the evaluation.** Invite representatives from different departments to provide feedback in the project and evaluation design phases, and communicate often about the current stage of the evaluation design and preparation. This is another way to generate staff trust: finding key people that support the evaluation and are invested in the outcome can help nudge staff who may be more hesitant or even unwilling to comply with the evaluation protocols.
- **A good way to secure staff buy-in is to tie the evaluation to the staff's interests.** Staff may not initially assume that researchers have the institution's clients' best interests in mind, or share the values of their organization. From a staff perspective, implementing a randomized evaluation can introduce new operating procedures which seem arbitrary and fastidious, so it is important to make the case for your evaluation clear. Foster a discussion about how the new product or evaluation could potentially contribute to the goals of the organization, and demonstrate how the product interacts with the rest of the institution's portfolio of products. We discuss interacting with staff more extensively below.

Training

Just before the launch of your RCT, you will need to make sure that implementing partner staff plus any surveyors involved receive the training that they need to effectively implement the intervention, randomization, and data collection.

IMPLEMENTING PARTNER STAFF

Even if the intervention is very low-touch and involves little participation from staff at your implementing partner, it is a good idea to make sure that staff understand the purpose of the research, so they can effectively answer any questions that may come from study participants. Ideally, implementing staff will not only have an understanding of the purpose of the research, but also an understanding of why an RCT was chosen as the method for the evaluation. Below we provide some things to think about when designing staff trainings.

- **Who is involved in the research and in what ways?** While the member service representatives at the credit union you are partnering with may be the ones making product offers, the tellers may be the ones answering questions later. Consider separate trainings for different staff who may be involved with the research in different ways. Likewise, ensure all partner staff are aware of who is involved on the research team and who they should go to for questions.
- **When and how does training normally occur?** The more that you can piggyback on existing staff training, the better; your training will be seen as more a part everyday operations by the staff involved. Determine how the financial institution trains their staff on new products and procedures, and try to mimic their process as much as possible

when creating your staff training materials**.** Use the organization's new product or procedure training protocols (if available) as a guide to integrate your evaluation into staff's day-to-day roles and responsibilities.

- **Can you talk to everyone?** We worked with one partner that staffed their branches 24 hours a day. While we wanted to meet with all the tellers, it was logistically impractical to have members of our research team onsite at 3am. We therefore had to rely on emails, with reinforcement from the branch managers, to communicate the requirements of the study to all branch staff. This made it much more difficult to be sure that everyone had understood the information, but also made it that much more crucial that the branch managers were on board and in full understanding of the needs of the evaluation.
- **How will new staff be trained on the evaluation?** Consider adding your evaluation manual to your partner organization's new staff on-boarding manual, or scheduling regular new hire orientation trainings so all new staff are aware of the intervention and its protocols.

SURVEYORS

Not all RCTs will require surveyors, but when they do, here are some tips to ensure that they are well trained prior to the launch of your RCT.

- **Training more surveyors than you need is always better than not training enough.** Even if the design only requires a few surveyors, many things can come up throughout the course of the evaluation that cause surveyor turnover. We typically recommend training 30 percent more surveyors than the team actually needs; this tends to be easier when working with a survey firm or marketing organization. In our own evaluations, we have had surveyors leave on short notice because of illness, family emergencies, maternity leave, and other job opportunities. Another common issue that arises is an extension of the fieldwork; for a multitude of reasons, the research team and partner may jointly decide to extend the baseline or endline survey, and many surveyors cannot commit to additional time outside of their original contracts. Both of these circumstances delay the speed of the implementation and once the evaluation begins, it is much more time consuming to schedule and conduct individual trainings of new hires than it is to have the foresight to train a surplus of field staff. If for whatever reason training a surplus of staff is not feasible, then a good backup to have is a ready-to-go protocol for onboarding new staff as they join.
- **A good way to ensure the consistency and quality of your survey data collection is to work a "false start" into your project launch.** The first day or two after the launch of your survey, carefully evaluate the quality of how your data was collected, coded, and consolidated. After this, scrap the data and then restart the data collection and hold a brief re-training and debrief session. The debrief session can be repeated during the early days of surveying; in both, you should gather all questions from surveyors and ask them to report any challenges or glitches they encountered in the field.

  During these sessions, clearly explain grounds for dismissal and expectations for data quality checks. While this is not easy to talk about, any falsification of data from surveyors will undermine the integrity of your study. Let surveyors know that they should tell the project team if there are any elements of the survey that they are uncomfortable with doing (e.g., a question about a sensitive topic like child support arrears). It is better to work through these types of issues early on and avoid of surveyors hiding problems and faking data. To this end, it is also good to communicate that if something happens at any point during the survey period and a surveyor needs to leave (e.g., a family emergency), they can always come to the project team.

- **Retrain!** Retraining is great to do throughout the project life cycle! Have surveyors demonstrate parts of how they administer the survey. Even if you don't do a false start, use the first debrief (no more than one week into the survey period) to discuss what is going well, what isn't, and what questions have come up. It's a good idea to run through any role plays from your original training again, but ask surveyors to add samples of real conversations they've had with respondents. This can help you identify issues you may not have anticipated, answer common questions, and ensure that your surveyors are conducting their work consistently and correctly with all respondents. For frontline staff, try role playing a scenario in which a client asks them about the intervention to assess how much of their original training they have retained. Make sure that you reiterate any information they need and review or revise the survey manual accordingly.
- **Establish survey protocols and secure agreement from survey managers prior to training or launching a survey.** This sounds like a no-brainer, but having the entire research and project team in complete agreement with the survey and the survey process is crucial to being able to control the messaging that your surveyors receive. We ran into a situation where we were working with a survey firm to administer a survey and a credit consent script. During an initial training meeting, the manager of the survey firm made it clear that he did not think our consent script was going to successfully achieve high consent rates, which sent mixed signals to the surveyors and resulted in surveyors skipping the consent script or paying less attention to it in the field. To avoid situations like this, make sure to include each key person in the conversation of what expectations for the survey are prior to trainings and early meetings.

GENERAL TIPS FOR TRAINING

Keep it simple! Include only the information that participants need to complete each individual task and understand where to go when issues arise. A great resource to leave behind with your surveyors is a survey manual, which includes instructions on how to administer key sections of the survey, FAQs that might arise during debriefings and early training sessions, as well as contact information if something unexpected comes up and they need to reach out to the research team. Make sure that the manual you leave behind is self-explanatory; write it in a way that assumes someone with zero knowledge could pick it up and know how to administer the survey and where to go when issues arise.

Supplemental Materials

Budget time and resources for the creation of marketing materials. Having all of these materials ready to go at launch is important to ensure that they are utilized and available when needed. Materials might include fliers, websites or web banners, scripts or messages, application forms, or membership and loyalty cards. Staff should be made familiar with the materials during training so that they know how to use them correctly.

It's important to note that sometimes your marketing *is* your intervention. For example, if you are providing nudges in the form of reminder messaging or streamlined enrollment, then the content and design is not just nice to have—it is crucial. In this case you should be looping in your partner and beginning work on material design as soon as possible.

Ensure that data collection systems are 100 percent in place. This includes not only considering what needs to occur in order for the randomization and data reporting to run smoothly, but also drafting and sharing a protocol to follow when the inevitable hiccup does arise. A helpful worksheet to provide at this stage is an external contact list of whom to contact with questions or concerns regarding each component of the project.

At this stage, any additional materials related to your intervention should be finalized. For example, if you're working with surveys, make sure that they are translated, piloted, formatted, and printed before arriving to the field. For more information on writing, piloting, or translating a survey, please see the Data Collection Section.

Preparing to Launch Checklist

- □ **Ensure all research personnel are certified to participate in studies involving human subjects**
- □ **Submit your evaluation to the relevant human subjects board for review**
- □ **Register your evaluation on the American Economic Association Trial website**
- □ **Train financial institution staff and surveyors on the intervention**
- □ **Ensure all legal agreements are in place**
- □ **Create and finalize all supplemental materials**
- □ **If surveying**
  - □ **Train your surveyors on the best way to administer your surveyor**
  - □ **Create a survey manual**

*The **Data Collection** section covers the guidelines for and types of data you may want to collect; this includes administrative data, credit report data, and survey data.*

Data Collection

One of the most valuable outputs of a randomized evaluation is a high-quality dataset. Creating and maintaining this dataset requires careful planning to ensure that data are collected and secured properly. In this section, we first discuss some general guidelines for data collection, including both the planning and security of your data. We then discuss the precursor to (almost) any evaluation's data collection process: the human subjects' protocols. Finally, we highlight some things to watch out for when working with and gathering administrative data, credit report data, and survey data.

Data Planning

While your evaluation plan should outline the basics of your data collection procedures, it is often necessary to take some additional planning steps before launching your intervention to flesh out the nitty-gritty of gathering your data. This can include creating a data sharing plan with your partner, deciding how data will be transferred, creating a data schema to map all of your data sources, and creating a codebook.

- **Create a data sharing plan with your implementing partner.** A data sharing plan outlines the data collection and transfer protocols between the implementing partner and researchers. At a minimum, this should include a description of the data to be shared, the collection and reporting time frame (how frequently will reports be sent and for how long), a breakdown of roles and responsibilities (who will be responsible for preparing and transferring each report), and the intellectual property rights of the evaluation (See the Appendix for a Data Sharing Plan template). Your data sharing plan can also serve as a useful record of the data decisions made between researchers and partner. Since data collection often spans several years, your data plan can be a helpful reference in case of staff turnover.

- **Determine how the data will be transferred.** Ideally, there will be more than one person at the partner institution capable of creating the data reports required, so that if one person is absent or transitions out of the institution, the research team can continue to receive consistent reports without worrying about turnover. Relatedly, make note of how long each piece of data will be available in your partner's database. We once worked with a financial institution that lost a portion of their transactions data at the end of the month so it was crucial that the extraction, transfer, and initial cleaning of the data were conducted in a timely manner. It may be possible to gain direct access to the partner's system so that the research team can take responsibility for creating the reports. While rarely permitted, this may be the best possible setup because it allows for easy troubleshooting while reducing the administrative burden on the partner's end.

- **Test run data reports.** We highly recommend that you gather some sample reports from each of your different data sources before you launch the intervention. Doing a test run will help you get a sense of the level of error or bias that each data piece might be susceptible to as well as the potential hiccups that may come up in preparing and transmitting the data. If your partner can't provide the full report early on, ask for a random sample of the relevant data during the planning phase (e.g., 1,000 transactions) in order to orient yourself to the format and structure of the variables.

  Examine the data and think about whether what you see is what you expected to see. Does the distribution of values make sense? Are there any fields or other variables that might be missing from the report (for example, different types of transactions)? Are there systematic trends in missing values? It is helpful to understand how

GENERATING A UNIQUE ID

Due to privacy or security concerns, financial institution may not be willing—or legally allowed—to grant researchers access to their internal unique identifiers (customer or account IDs). In the absence of another auto-generated identifier (such as a survey ID) the project team will need to generate a unique identifier for each observation in the study sample. However, matching different dataset IDs and maintaining the process with a dynamic sample can be a time-consuming (and error prone) task for both the implementing partner and research team.

Depending on your partner institution, the IT department may be able to generate unique IDs using existing information. For example, many databases have a 'row counter' variable which increases as new entries (e.g., new customers) are added. The counter variable is typically generated automatically, and is unique to each observation in the data set and could act as the study's ID. Content-based unique IDs are another option. Content-based unique IDs are based on another unique identifier, such as customer ID or account number. Using a program, such as MD5 or SHA-1 hash, you can create a string variable that is unique to each combination of the identifiers you have specified, while masking any sensitive information. However, in certain circumstances it could be possible to reverse-engineer this process, revealing the original identifying information. Make sure to consult your IRB and IT team before settling on an approach.

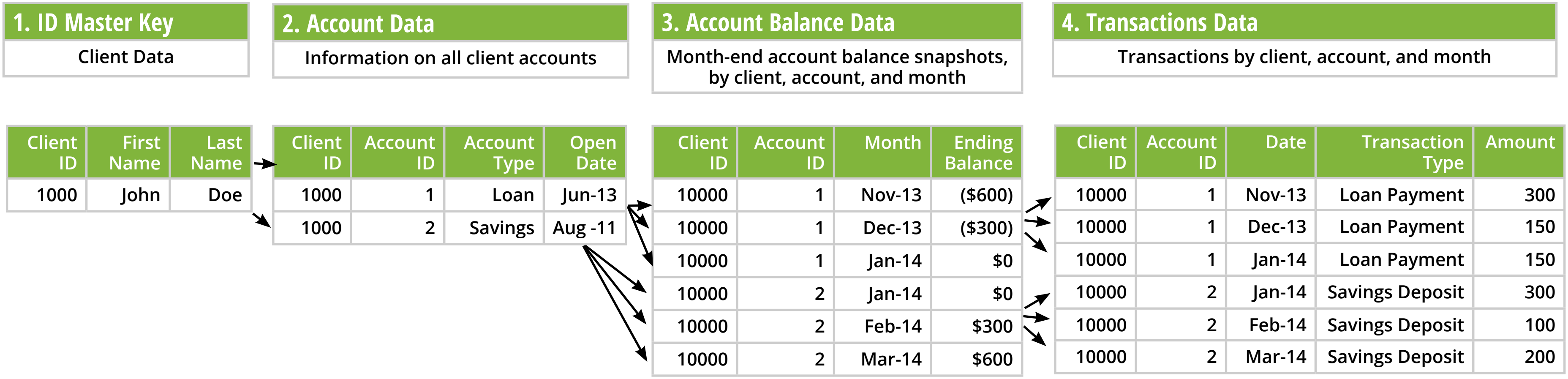
reports are generated—this test run period is especially useful for conducting checks of the queries and systems used to generate the data. For example, administrative data that is entered by hand by financial institution staff or customers (i.e., data recorded during a financial counseling or loan application session) is likely to have more errors than credit report data that is generated automatically.

Test runs are also useful for orienting implementing team members to new protocols (e.g., encrypting datasets or pulling new variables). If possible, test the report for several periods to identify inconsistencies across time; this is especially important if archive (retrospective) data reports are hard or impossible to obtain, or are even more costly than a real-time report.

- **Map out your data sources and how they will connect.** Often you will have more than one data source for your study. For example, you may plan to collect data from surveys in addition to administrative data from an implementing partner, or you might want to merge your data with other publicly available datasets (such as GIS data). Before you begin collection, it is helpful to create a schema that maps out how your different data sources will come together (See Figure 2 for an example schema). The most important element of this is deciding what unique IDs (identifiers) will be used for each dataset and ensuring that their format is consistent across datasets. It is important to consider how the data will eventually merge (i.e., a one-to-one, one-to-many, many-to-one, or many-to-many correspondence between the unique identifier and the observations in your dataset). Keep in mind that different datasets, even from the same source, may be structured differently. You may have demographic data with one observation per client, account balance data with three to four observations per client, account balance data with one observation per client-account per month, and transaction data with dozens of observations per client-account per month. This will determine how the data will eventually merge.

Having the right data schema in place early on will also save a lot of time when it comes to the analysis phase later. Your data schema should also address (1) how data collected at different units of randomization will be mapped together for the purposes of analysis, and (2) how this will change if data will be collected at multiple points in time.

FIGURE 2: SAMPLE DATA SCHEMA



1. ID Master Key

Client Data

Client ID	First Name	Last Name
1000	John	Doe

2. Account Data

Information on all client accounts

Client ID	Account ID	Account Type	Open Date
1000	1	Loan	Jun-13
1000	2	Savings	Aug -11

3. Account Balance Data

Month-end account balance snapshots, by client, account, and month

Client ID	Account ID	Month	Ending Balance
10000	1	Nov-13	($600)
10000	1	Dec-13	($300)
10000	1	Jan-14	$0
10000	2	Jan-14	$0
10000	2	Feb-14	$300
10000	2	Mar-14	$600

4. Transactions Data

Transactions by client, account, and month

Client ID	Account ID	Date	Transaction Type	Amount
10000	1	Nov-13	Loan Payment	300
10000	1	Dec-13	Loan Payment	150
10000	1	Jan-14	Loan Payment	150
10000	2	Jan-14	Savings Deposit	300
10000	2	Feb-14	Savings Deposit	100
10000	2	Mar-14	Savings Deposit	200

- **Create unique anonymous keys to link the datasets.** Sometimes the reports you will receive will include a unique ID, such as a social security number or a partner-generated member number, but ideally you will have a research team-generated unique ID that is on each data piece you collect. A research team-generated unique ID has several advantages over using pre-existing IDs: (1) if constructed corrected, there is no chance that a team-generated ID includes Personal Identifiable Information (PII) and can therefore be shared with parties that are not included on data security agreements; (2) the same unique ID can be on each dataset, eliminating the need for multiple IDs, and (3) there is no chance that the IDs will somehow change over time, the way that a Member ID at a credit union might change. In the Case Study box to the right we provide an example of to include researcher team-generated IDs on multiple data sources.
- **Create a codebook for the data.** A codebook provides information about the names, content, and scope of the data collected throughout the evaluation. The codebook should include a comprehensive list of all the variables collected for the evaluation (noting unique IDs in particular), a description about how each variable is generated, and a list of all potential values and (expected) anomalies. A good codebook will make the data easy to understand and alleviate problems about the future interpretation of the data (See the Appendix for a sample Codebook template).

Data Security

A loss or unintentional disclosure of data could have potentially severe consequences for your study's participants, as well as have implications for the integrity of your findings and the reputations of partner organizations and the researcher team. A comprehensive plan to secure your data, known as a Data Security Protocol, will help mitigate the risks and consequences of a data breach. A Data Security Protocol typically includes details regarding how the data will be collected, handled, and stored. In general, project data should always be password protected, encrypted (both during transfer and storage), and backed up regularly to multiple drives or secure

CASE STUDY: ONE ID FOR MULTIPLE DATA SOURCES

We conducted a study where our data sources included (1) administrative data from a credit union, (2) credit report data, and (3) survey data. The administrative data uniquely identified members through several fields, including: account number, Member ID, phone number, name, and address. The credit reports identified clients based on social security number, name, and address. Our surveys were administered through a survey firm and were each appended with a seven-digit unique ID upon completion. We decided to use the seven-digit ID as our central unique identifier across all datasets. To do this, we included this ID in with the credit report input files (the list of customers to be included in the credit report pull), and through our agreement with the credit bureau asked that the IDs be sent back with each report. For the administrative data, we merged these unique IDs with the credit union's member IDs, keeping a separate ID dataset that had the match for each account number and each survey ID. This process helps to simplify the analysis because even when the data are appended into a panel dataset, each observation is uniquely identified through one ID and the time variable. Each data source was pulled and merged on a monthly basis. At the end of the study, we were able to strip the identifying information from all pieces of data and keep our single unique ID across all datasets, enabling us to post the dataset publicly and share it with other researchers.

servers. Below are a few additional things to consider when thinking about securing the project's data:

- **Do your data constitute Personally Identifiable Information (PII)?** If your data constitute PII, you will need to take extra measures to protect the data. This may consist of: storing the data in an encrypted location; removing PII from the dataset and replacing it with a unique identifier that only you can link to the person's identity (as was discussed in the previous section); and filing your study formally with an Institutional Review Board (IRB) and reporting when any identifying information is accidentally lost, disclosed, or stolen. You can read more about what constitutes PII and how to protect PII in the box to the right. Often, IT systems may already build in a set of identifiers that do not constitute PII, like an internal member ID; in these cases, you may be able to receive the data without any identifying information.
- **If your data do constitute PII, are you allowed to receive them?** Financial institutions should have their legal team confirm that they can share identifying information about their clients with a third party (in this case, an evaluator) with a written agreement. At the same time, if you will be using credit report data you should check your agreement with the corresponding credit bureau; some bureaus will not release data if it can be linked to a client's name, contact information, and social security number.
- **Obtain confidentiality agreements from surveyors, data entry operators, and project staff.** Anyone who handles PII should sign a confidentiality agreement before being allowed to view to PII.
- **Store, transmit, and use PII separately from the rest of your data.** If you are using a paper questionnaire to collect survey data this means separating PII from survey data as soon as possible. We usually collect PII (name, phone number, account number) on the first page of a survey and detach the first sheet from the rest of the survey at completion. Each printed questionnaire has a unique ID, which is printed on each page of the survey, to facilitate matching PII and survey data during the cleaning and analysis phase.

WHAT IS PERSONALLY IDENTIFIABLE INFORMATION (PII)?

The National Institute of Health (NIH) defines PII as information that is personal in nature and which may be used to identify a person. This is intentionally broad, and in the financial context covers a lot of ground. The EU Data Protection Directive (95/46/EC) defines "personal data" as information relating to an identified or identifiable natural person. An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. The EU Directive considers even standalone financial account numbers to be PII, and the standard that many IRBs in the United States adopt is similar.

Human Subjects

Under US Federal Policy, all studies using human beings as research subjects must follow certain ethical research guidelines.[7] As we discussed in the Preparing to Launch section, you will most likely need to have your study reviewed and approved by an accredited Institutional Review Board (IRB) before you begin collecting any data for your research. An Institutional Review Board (also referred to as an independent review board, independent ethics committee, ethical review board, a privacy board, and human subjects review board) is a group designated by an institution (such as a university or

[7] US Policy for the Protection of Human Subjects

non-profit) to approve, monitor, and review research involving human subjects to assure appropriate steps are taken to protect the rights and welfare of those subjects.

HIPAA COMPLIANCE

The Health Insurance Portability and Accountability Act (HIPAA) of 1996 **outlines specific requirements about how Protected Health Information (PHI) should be transferred, stored, and de-identified. Even if researchers are not working with health data, many organizations use HIPAA standards for their data security. PHI can include demographic information, the health condition of individuals, or additional information that can identify individuals, even indirectly. Researchers working with PHI will not necessarily need to comply with HIPAA. We suggest checking in with the partners' legal or compliance team and/or your institutional review board to determine if additional safeguards or controls are needed for the evaluation. For more information about HIPAA rules and regulations, visit the US Department of Health & Human Services** website.

DO YOU NEED TO APPLY FOR IRB REVIEW?

Studies that receive US federal funding must be reviewed by an IRB registered with the Office for Human Research Protections. Additionally, most universities require IRB review and approval for any research involving human subjects that is conducted by their staff, faculty, or students. While this means that most research will require some IRB oversight, there are cases where research may be exempt from continuing IRB review, research may qualify for an expedited review, or informed consent may be waived. Most universities have their own IRBs, and some research organizations do, as well. Before beginning study enrollment or data collection talk to your institution's IRB or human subjects coordinator to make sure you are not running afoul of human subjects protocols. Noncompliance can jeopardize a researcher's reputation, the publication of papers (as some journals require it), and funding opportunities.

REQUIREMENTS

Guidelines will vary from institution to institution, so it is important to consult your IRB and learn all of the requirements for requesting approval. Typically, however, most IRBs will require you to provide:

- The purpose of your study
- A review of your study's data sources and collection methods
- A copy of the informed consent that will be provided to study participants
- Copies of survey instruments
- Data security protocols
- Any educational or marketing materials used as part of your study
- Letter of support from your partner organization

Your IRB will be reviewing your study proposal and supplemental materials to determine if: (1) your study will make a valuable contribution to your field; (2) any risks (psychological, physical, or social) to the research participants are reasonable in relation to anticipated benefits to subjects, and to the importance of the knowledge that would be generated as a result; (3) subjects' data and privacy will be adequately protected; and (4) the consent forms adequately describe the study to participants including the time involved, benefits to participation, right to refuse to participate, and a way to contact someone from your research team about the study.

INFORMED CONSENT

Many IRBs require that researchers collect informed consent from study subjects. When consent is tied to obtaining data on study participants, the consent protocol and language have serious implications for the project. While each IRB will have strict guidelines about the content of informed consent scripts and the way they are administered, here are a few things that deserve consideration when you are designing your informed consent protocol:

- **Be careful about the language in the consent script.** Consent scripts often sound like legal disclosures, in that they are long and use a lot of technical jargon. Using bulleted lists instead of paragraphs for verbal scripts can make the disclosures more understandable to study participants. See an example consent script on page 54 for further guidance on language.
- **Pilot different consent protocols.** The perception of the intervention by potential study participants might be affected by the consent script, or vice versa, and it is worth asking whether the structure of the consent protocol affects consent or intervention take-up. For example, is the consent acceptance rate affected by whether the consent comes before or after the product offer? Where possible, test different methods of obtaining consent during the pilot phase.

Working with Administrative Data and Credit Report Data

Most RCTs on financial products and services use some combination of administrative data and other data sources, and many of them use credit report data. While the data analysis varies project-to-project, the concerns and limitations of collecting administrative and credit report data tend to be similar across different RCTs. In the following section we go through some questions to consider while refining your data collection process.

ADMINISTRATIVE DATA FROM FINANCIAL INSTITUTIONS

Administrative data are data produced and kept by an institution on their members or clients for primarily operational, rather than research, purposes. This typically includes basic demographic information and financial records, such as transactions histories and checking, savings, and loan balances. All financial institutions are required by law to keep data on their customers and accounts; for this reason, administrative data can be very reliable. It tends to be more consistent and unbiased than survey data.

Each institution usually has a data manager(s) who is responsible for creating queries that produce these reports. An early meeting with this person can help establish what data the research team has access to and the different types of reports that can be created. Make sure to ask your partner institution for a sample of each type administrative report they produce, to see which best meets your needs. If your partner doesn’t have a sample administrative data report ready, send them your data “wish list,” a fake spreadsheet of the variables of interest in the format you expect to receive from them, to get the ball rolling.

As helpful as administrative data can be, there are many issues that can arise during the process of collection and generation. To mitigate some of these issues, take time to learn how the administrative data you receive are generated. You may want to ask if the IT staff has a codebook or data user manual that describes the variables (also sometimes referred to as "attributes") in the reports and how they were created. If they do not have something comprehensive to reference, ask clarifying questions about how the partner institution collects the data. Understanding how the data are collected and generated prepares the research team to troubleshoot errors in queries and inconsistencies across reports. Some questions to ask IT staff or the data manager include:

- **What query generates the data?** If possible, it can be helpful to see the specific code that generates the relevant query. This allows the research team to troubleshoot and preserves a record of the query in the case it needs to be re-run or altered at a later date. A common problem arises when a query excludes observations based on the values of other variables, in which case you may find that people are absent from your sample once they close their account or make below a certain number of transactions per month, for example. Depending on how you run your analysis, you may need to ensure that the query you create does not condition on any single variable value.
- **Where is the data from?** Is it reported by the clients themselves, pulled from the client's credit report, or produced by the financial institution? This can help gauge the quality of the administrative data and the magnitude of any discrepancies that might arise.
- **When are the data fields updated and/or extracted?** This might affect the unit of time that is used in the analysis and might create some issues with inconsistencies month-to-month.
- **What are common errors that could occur in the dataset?** Ask the point of contact that works with administrative data, or see if the research team can check a randomly drawn sub-sample of records to see where there might be discrepancies. For example, if the data are inputted by the tellers by hand, then checking and savings transactions may be classified incorrectly at times. If so, there might be a way to identify a correction on the part of the financial institution, or in your analysis. You might also want to verify how the database distinguishes between missing data (null values) and zero values. During one of our evaluations we realized that the high incidence of zero values in our dataset was actually a function of underreporting on some variables.

On the page 57, we provide a checklist your research team might want to go through with the implementing organization regarding administrative data. This checklist is divided by data type: transaction data, account data, demographic data, account balance data, application data, and debt management plan data.

CREDIT REPORT DATA

Although credit report data are fairly expensive, it is one of the most accurate sources of financial data; much of the monitoring and quality checks that the administrative data may require can be forgone for credit reports. Credit reports can be obtained through agreements with any of the three major credit bureaus: TransUnion, Experian, and Equifax. Another way to obtain these might be through a third-party provider or directly through the financial institution the research team is working with. Many credit unions and banks have their own separate agreements with credit bureaus that they use to conduct soft credit pulls. The research team may be able to go through a pre-existing agreement, which is typically less costly and requires less paperwork.

Sample Consent Script

We used the following script to obtain verbal consent from customers during an evaluation of a savings product in New York. Surveyors approached customers in our partner's lobby during store hours and asked for consent to conduct a survey and for permission to conduct soft credit pulls. In addition to receiving the verbal script, customers received a paper copy of the informed consent to take home.

"*Hi. My name is _______. Today we're inviting people to participate in a research study on financial habits led by IPA, a non-profit research organization. Do you have 5-10 minutes to take a survey about your savings?*

Thank you! Before we begin, I just want to give you some standard information:

1. *There are no risks involved in taking this survey.*
2. *Your participation is completely voluntary.*
3. *You can skip a question or stop the survey at any point.*
4. *Your identifying information will be removed before the data we collect is shared with other researchers or results are made public.*
5. *Your answers will have no effect on your relationship with ACME Financial Institution.*
6. *They will be used to help financial institutions like ACME design better products and services.*
7. *We may invite you to take 1-2 more voluntary surveys within the next year using the contact information you provide.*
8. *We may also pull your data from credit reporting agencies for research purposes, over the next 24 months. Any credit report data would be obtained from 'soft pulls' of your credit report that have no effect on your credit score.*

If you have any questions or concerns about the study you can contact Jane Doe at IPA. I'm going to give you a printed copy of what I've just said with Jane's contact information.

<u>Do you consent to participate in this study?</u>"

Soft credit pulls on study participants typically include, at a minimum, a credit score based on the scoring model selected by researchers, and then any attributes that researchers elect to purchase. During a contract negotiation, credit bureaus may provide the research team with a codebook of all possible attributes to select, and then the research team can pick up to a certain amount on a flat-rate basis; additional attributes typically cost more. Other contracts require the research team to pay per attribute. Examples of possible attributes to include in a report are the client's credit limit, the main reasons for a score (e.g., a recent derogatory record or missed payment), the client's aggregate balance for all tradelines, or their aggregate balance under all auto loans.

- **Check the format in which credit reports will arrive.** This can be the most difficult part of working with credit reports. Ideally, the reports will arrive in a structured dataset format, like .CSV or .XML. Otherwise, the data may arrive in .RTF or .PDF files, which requires paying someone for data entry, parsing the text in-house to extract the data, or creating a script that scrapes these files for the required information.
- **Note that credit report data comes at two levels of detail—individual-level data (such as score) and trade-level data (such as balances on a specific credit card).** Determine which of the two the team will be receiving, and how this level matches to other data sources, if any.
- **Some bureaus and third party organizations will provide only aggregated tradeline data for each individual.** For example, an aggregated tradeline might include data on the total amount of student loans held by a consumer, but no data on each individual loan.
- **Consider who will be missing or unscored from your datasets.** This might include determining how many unscored individuals the research team expects, or whether the intervention will affect whether a client receives a score or not. Some credit reports will provide records for individuals that were cut (received a score below the minimum for whichever scoring system they are using—350 under both Vantage 3.0 and FICO 2008, for example), while others may not. To list who may be missing, it can be helpful to request a list of possible record types from the organization providing the credit reports, to see the range of record categories that might be returned. See the box on the next page for a sample of records returned with a one of our credit agreements.
- **Note that credit reports often have a very high variance across scores and tradeline balances, and some indicators (such as score) may be especially slow to move.** While credit reports are an excellent source of reliable and high-quality data, they tend to be less likely to pick up large effects as a result of changes and interventions
- **Although payday loans are not reported to credit bureaus, once these loans go into collections they can appear on a credit report.** Because of this, in some cases it may be possible to identify some clients that have previously used a payday loan; however, this sample will not capture those who successfully repay their payday loans.

CASE STUDY: CHECKING DATA SOURCES BEFOREHAND

We ran a project with a partner in New York where we noticed during the analysis that some of the demographic data changed month-to-month. We later realized that it changed depending on who had conducted the transaction. For example, an observation from a husband and wife with a joint account would alternate between M and F for the gender, depending on who had made the deposit or withdrawal.

SAMPLE CREDIT RECORDS

Accepted file: credit file matched on information provided

Scorecut: record does not fall into score range; record provided but consumer unscored

Regular household drop: multiple member in the same household, randomly selected

SSN Dedup: one or more records containing the same Social Security number (SSN)

Unscored records: record could not be scored by the scoring model used

No hit: record could not be found in the database

Not processed: record dropped because it is outside of the required geographical area

- **Some clients may have duplicate and split credit files.** If the research team is collecting multiple addresses, then it is a good idea to pull a credit report on each person, for each address that they list--especially if the credit report pricing is done on a flat-rate basis. More likely than not, one address will return a credit file and the other will return a "no hit" record. These duplicates can be tricky to reconcile during the data cleaning process, but there are multiple and split files present at all of the credit bureaus. Most often this is a result of a data reporting issue where, for example, a consumer's information is reported by one credit grantor with a name, address, and social security number, and another credit grantor reports their information with a name, a different address, and possibly no social security number. The difference in reporting data can cause multiple files to be created.

Survey Data

There are a number of questions to consider if your project requires you to use survey data. You will need to determine the channel through which the survey data will be collected (e.g., phone, in-person, computer-based), whether or not to use surveyors (sometimes also known as enumerators), and whether to use paper or electronic questionnaires. You will also need to design the survey questions themselves to ensure that you are collecting the data you need.

SURVEYORS

Surveys are typically administered by one of the following.

- **Financial institution staff, such as tellers or member service representatives.** These staff might be well suited to approach customers but they may be less strict when it comes to following survey protocols or clarifying survey questions. Additionally, it is best to minimize the impact of the intervention on the workflow of frontline staff—adding a survey might be taxing for frontline staff and could be met with resistance. One of our partners encouraged their tellers to complete transactions quickly and to keep customer waiting times low. In this case, asking tellers to conduct a survey would have led to low survey completion rates. Instead, we hired a survey company to stand in the lobby and survey customers as they entered and exited the branch, with tellers referring customers to the surveyors.
- **Surveyors hired directly by your research team.** Direct hires provide you with the greatest amount of control over who is selected and how they are trained and managed. This will likely give you the best quality data. However, this also requires a large investment from your research staff to oversee hiring and logistics.
- **Surveyors hired through a survey or marketing firm will likely free you of some logistical and human resources headaches**. Survey firms will have the infrastructure required to handle surveyor recruitment, hiring, payroll, and tax form preparation. A firm may

Administrative Data Checklist

Transaction Data	❑ **Confirm that every transaction has a unique ID.** Since multiple identical transactions may be processed on the same date (e.g., a person might buy two $100 money orders), it is often hard to check for duplicates without a transaction ID. ❑ **Confirm that you can identify the processing order of transactions.** This will make it possible to reconstruct how account balances vary over time. Without a processing order, you may only be able to detect macro changes on a month-to-month basis. ❑ **Check whether transactions have a group ID or receipt number.** In some cases, when a client processes multiple transactions in a single teller session these are bundled together with a distinct ID number. This creates problems if researchers want to differentiate between activities in different accounts—if they are bundled together under the same teller session, it can be difficult or impossible to take them apart. ❑ **Check for double-entry accounting.** Every account credit should be matched to a corresponding debit; however, this is important for the financial institution's accounting and less so from a research perspective. If your reports include double entries, you will need to be able to identify these transactions so you can drop them from the data or account for this in the analysis. ❑ **Check whether you can identify fees. Are they bundled into the transaction, or processed separately?** Is there a code by which you can identify them and drop them from the datasets? Sometimes fees are included as separate line items, but sometimes they're bundled into other payments or withdrawals. ❑ **Check how positive and negative values are stored.** Systems often list the absolute value of the transaction and the transaction code (e.g., deposit or withdrawal) to indicate whether the amount should be positive or negative. An institution might not produce this information for all transactions, however, so it is impossible to tell the difference between a deposit and a withdrawal. ❑ **Confirm that you can match each transaction to an account and client.** Make sure that the research team knows how to uniquely identify who conducted each transaction.
Demographic data	❑ **Check whether joint accounts are permitted, and if so, how they are recorded.** If joint accounts are lumped together, it may be difficult or impossible to track individual-level financial changes. Similarly, if you are randomizing at an individual level, joint accounts processed together (where you can't tell the difference between who conducted each transaction) might have to be excluded.
Account balance data	❑ **Check when balances are pulled.** For example, are balances recorded on the date the dataset is generated, or can you reconstruct them retroactively? Ideally, you'll want the balances from specific times of the month, such as the 1st and 15th. ❑ **Confirm that you can match each account to the corresponding client.**

Administrative Data Checklist Continued

Special considerations for Debt Management Plan (DMP) clients	□ **How are fees and refunds processed?** Fees might be bundled into debt payments, or they might be listed as payments on separate line items, so if your analysis focuses on debt repayment then ensure that you can differentiate these. Also, bounced checks are sometimes not processed until the following month and appear as negative balances. □ **Is it possible to gather data on clients who are self-administered?** Self-administered refers to clients who elect to exit their DMP and make payments to their creditors on their own, usually to avoid paying the monthly DMP fee or because of newly realized potential from the budgeting and planning support during the set-up process. In these cases, DMP providers will likely lose the ability to track credit outcomes of the clients. □ **Are debt balances ever revised?** If debts are reconsolidated and revised on the same account per client, then debt balances might unexpectedly jump up. □ **Can clients restart the DMP?** If so, how is this recorded? Can you get records of both the original, and the latest, client start date? □ **Are you able to track the client's DMP status at the start of the month, and monitor the reason for non-payment?** What triggers a change in status? Do two missed payments result in a status change for non-payment? □ **Can clients' scheduled payments change over time?** It's not uncommon for DMP providers to adjust the payments in the face of adverse events—for example, after losing a job the client may reduce payments by 50 percent. Make sure that this is not identified as an underpayment in your data.

have different quality standards and emphases than academic researchers, however, so you may have to spend more time monitoring surveyors.

COLLECTION METHODS

Data can be collected via paper, or through computer assisted interviewing (CAI). Questionnaires are typically administered in-person, on the phone, via the Internet, or other remote data collection. What you choose depends on the needs, budget, and time frame of your evaluation and survey.

- **Paper surveys require comparatively little technical training to construct and deploy.** Researchers can design the survey using familiar tools like Microsoft Word or Excel (although formatting the questionnaire in these programs can be time-consuming). Paper surveys can become expensive, however, especially if you reprint the surveys each time they are updated and they go through many iterations. Entering paper data into an electronic format is also costly. If entry is done in-house (i.e., by surveyors or members of the research team), designing the data entry interface can be time-consuming and the managerial requirements can be significant. You will likely not be able to run quality checks on incoming survey data if you use paper forms. If data entry is done by a third party data-entry firm, researchers will need to create specialized data entry templates and monitor the quality of the data entry. Plan ahead if you intend to revise and deploy new survey versions. For example, you might send one-third of the households in your sample Version 1.0, wait for the initial set of responses, revise the survey accordingly, and send Version 2.0 to the rest of your sample.
- **Electronic or Computer-Assisted Interviewing (CAI) allows researchers to control data quality.** By restricting value fields and programming skips and logic checks, CAI can help surveys run more quickly and diminish the possibility of errors in survey logic. CAI surveys will require a significant up-front investment however, and depending on the software and the research team's technical proficiency, may also require hiring a survey programmer. In addition to the time needed to program, debug, and test the survey, a considerable amount of time is required to procure equipment, pilot, and install software on devices. Additionally, during data collection, the research team should be prepared to spend time correcting technical and equipment problems.

DATA SECURITY AND OUTSOURCED DATA ENTRY

Double-check the project's data sharing agreements when selecting a third-party data entry firm. Some forms of data, such as social security numbers, cannot legally be sent out of the US for data entry. If you are using a firm where data entry will be outsourced to workers in a different country, make sure that your agreements allow you to send PII overseas.

QUESTIONNAIRE DEVELOPMENT

A good questionnaire is key to gathering accurate and reliable primary data, so you should be as thoughtful as possible when creating and selecting your questions. Please refer to the discussion of Survey Piloting in the Pilot section for a more detailed discussion of the types of hurdles you might encounter when developing and testing a survey. We've compiled a list of survey development resources (see Questionnaire Bank) to help with designing your questionnaire. In addition, there are a few things from our experience that are worth highlighting.

- **Where possible, use published measures.** Pre-piloted survey questions help cut down on some of the iterations in question development. Since they have been tested previously, they're likely to be more accurate in their targeting and are less susceptible to being misinterpreted by respondents. Using validated measures also allows researchers to easily compare their intervention's outcomes to those of other evaluations that have used the same survey questions. The survey question banks included in the text box on page 62 include surveys with commonly-asked household finance questions.
- **Explicitly relate each question in your survey to your Theory of Change while designing the survey.** Keep analysis in mind as you consider what to include and how to structure your survey. While it might be tempting to add questions to your survey simply because you have a captive audience and it seems like a good opportunity to gather as much information as possible, each question adds to the length and complexity of the survey which in turn affects the focus of respondents and the precision of the data. Keep surveys short by restricting the content to measures dictated by the Theory of Change.

PHONE SURVEYS

Phone surveys have the benefit of being ubiquitous in the US. Additionally, the implementing financial institution may already be equipped with the infrastructure required to conduct a round of outbound calls to their customer base (e.g., contracted a call firm, customer phone numbers, and appropriate permissions to contact customers). If the evaluation requires setting up a phone survey from scratch, however, researchers will need to become familiar with telecommunications laws; the US Federal Communications Commission (FCC) has strict rules about calls or texts sent to mobile phones. For example, telephone solicitation calls to homes are prohibited before 8am or after 9pm. If the text or call is considered a "commercial text", senders will need to obtain informed consent from customers before initiating SMS contact.

Researchers might also include behavioral games and observational questions in their survey plan:

- **Behavioral games are measures of economic preferences designed to measure risk (loss aversion, probability, ambiguity), time-preferences (self-control, hyperbolic discounting) and social attitudes (altruism, pro-social behavior, and trust).** These games are becoming more ubiquitous in consumer surveys as they are 'context-free' and can be useful for testing hypotheses about heterogeneous treatment effects. The framing of the question and surveyor neutrality in the administration of the questions are key to reducing any social connotations, so the research team will need to take special care when developing the questionnaire and during surveyor training, especially in cases where there is interactivity between subjects.
- **Observational data is gathered by visiting the program site or the respondent and taking note of key program or respondent characteristics.**[8] For an individual, this might include characteristics of their dwelling, such as the number of rooms, the quality of public services available, etc. For a financial institution, this might include the number of people who

[8] Yoong, Joane; Mihaly, Kata; Bauhoff, Sebastian; Lila, Rabinovich; Hung, Angela. 2013. *A toolkit for the evaluation of financial capability programs in low, and middle-income countries.* Financial Literacy and Education Russia Trust Fund. Washington DC; World Bank. http://bit.ly/FinCapToolkit

CAI Checklist

❑ **Encrypt the hard drives of laptops, tablets, and phones used to collect data.**

❑ **Create two different log-ins, administrator (researcher) and surveyor to control access to the device's programs and applications.** Set the administrator account to have full rights. Set the surveyor account to have limited access—the bare minimum required to run and upload surveys.

❑ **Install an anti-virus program and run it at least once per day.** There are many free options, but we recommend Avast!

❑ **Uninstall or disable unnecessary programs.** Make sure though, that the survey does not require any of the programs you've disabled, for example audio or video functions. We recommend AppLocker.

❑ **Use a sync program to regularly sync important data to an external device, like Box Sync or Dropbox, and make sure to encrypt and password protect any and all sensitive data stored on the cloud.**

❑ **Have surveyors sign Device Liability Forms.** These typically state that the surveyor promises not to loan the device to anyone, or use it in a manner that could damage it or that is not in line with the data collection protocols.

❑ **Develop a system for surveyors to check-in and check-out the device.**

❑ **Lay out a plan to charge the devices daily.**

❑ **Create a password for the device's lock screen.** Make sure the lock settings allow the surveyor to leave the screen unlocked during interviews, especially if they don't know the device passcode.

❑ **Leave extra time during training for surveyors to learn how to use the devices.**

spend time at a branch location and for how long. Observational data can be a useful supplement to survey or qualitative data, particularly when you have reason to believe that respondents may have an incentive to misstate their behavior or experiences on surveys. However, observational data can also be time consuming to collect, and requires additional training of surveyors to ensure that observations are consistent and objective across surveyors.

QUESTIONNAIRE BANK

Survey of Consumer Finances, Federal Reserve Board (http://www.federalreserve.gov/econresdata/scf/files/2010_scfoutline.pdf)

National Survey of Unbanked and Underbanked Households, Federal Deposit Insurance Corporation (https://www.fdic.gov/householdsurvey/2013appendix.pdf)

American Life Panel, RAND (https://alpdata.rand.org/index.php?page=data)

National Financial Capability Study, FINRA (http://www.usfinancialcapability.org/downloads/NFCS_2012_State_by_State_Qre.pdf)

Measuring Financial Literacy: Questionnaire and Guidance Notes for Conducting an Internationally Comparable Survey of Financial Literacy, OECD (http://www.oecd.org/finance/financial-education/49319977.pdf)

A Survey of Consumer Views on Debt, Consumer Financial Protection Bureau (http://files.consumerfinance.gov/f/201412_cfpb_survey-of-consumer-views-on-debt.pdf)

Data Collection Checklist

- □ Identify the variables of interest in each dataset
- □ Record each variable’s content and structure in a Codebook
- □ Map out your data sources and how they will connect in a data schema
- □ Identify the unique anonymous keys that will link the datasets
- □ Describe the data transfer protocol and reporting time frame in a Data Sharing Plan
- □ Create a Data Security Protocol that details how the data will be collected, handled, and protected at each stage of the evaluation
- □ Gather and analyze sample reports to get a sense of issues that may come up during data collection, transmission, and analysis

*The **Pilot** section covers the reasons for and best practices of running a pre-RCT pilot study.*

Pilot

Regardless of the intervention being tested, there are typically three potential goals to accomplish via piloting for the purposes of an RCT: to test the implementation of randomization and data collection protocols, to test the implementation of a product, service, or nudge in a new context, or to test a brand-new product or service (product piloting). In some situations, the product has been sufficiently tested in other contexts or is already currently in use. In this case, the pilot is solely intended to test the logistics of the evaluation, which may include the randomization, the marketing of the treatment, the delivery of a nudge, and the data collection.

In other situations, the product is new or recently developed, and the pilot will test both the logistics of offering the product during the evaluation and the evaluation itself. Sometimes multiple pilots are necessary; in cases where, for example, the partner organization has not conducted any prior studies or experiments, the product has just been developed, and the research team has a nascent relationship with the partner, it may be helpful to do multiple sequential pilots that build on each other and test each component separately. However, there are ways of rolling the different measurements of interest into the same pilot and adjusting the full-scale implementation from the findings of a single pilot.

In this section, we discuss considerations for piloting the feasibility of your evaluation, the elements of the intervention, and the data collection and outcomes. All three sections point to similar prescriptions: early contingency and action plans that draw from potential outcomes found in the pilot help pave the way for moving into a successful RCT launch.

Product Piloting

During a product pilot, institutions test the offering, implementation, and marketing of a new product before widely offering it to their customers. Institutions and researchers should plan a product pilot any time the institution (1) offers a new, never-before-tested product or (2) introduces a product to a new market. Results of the pilot will help determine whether there is sufficient customer demand to justify scale-up, highlight the strengths and weaknesses in implementation, and identify potential tweaks in the design and delivery of the product that better suit customer or organizational needs. While organizations may have existing protocols when it comes to testing and rolling out new products, there a few that have worked well in piloting new products.

BEFORE THE PILOT

- **Identify and clearly state the value proposition.** How will the product fit with the institution's financial and operational strategy? How will the new product complement the overall portfolio of products and services? Institutions might expect data from industry reports or the impact on their financial bottom-line to comprise a significant portion of the value proposition, but we have found that incorporating academic literature (e.g., behavioral economics literature or evidence from analogous randomized evaluations) can also pique the interest of key decision makers.
- **Identify milestones at which to take stock of the pilot's successes and failures.** Project teams should establish benchmarks ahead of time. For example, a milestone could occur on a quarterly basis or after the first thousand offers, at which the team formally reconvenes and assesses progress so far. Use these pauses to formally reflect on take-up, staff and customer feedback, and changes to the pilot timeline or work plan.
- **Set SMART (specific, measurable, attainable, realistic and time-bound) goals.**[9] How will you determine if the pilot is a success? What results will determine whether or not the organization continues offering the product? These metrics should also be used to inform decisions about whether to continue, change, or scrap the product offering. Reasonable metrics for determining success—as well as a plan in case of failing to meet those metrics —should be set before launching the pilot, to avoid situations where future decisions are made based on path dependency and sunk cost fallacies, i.e., continuing to pursue the project regardless of whether it is working simply because you've already invested time and resources in it.
- **Determine where you will pilot the product.** To maintain control and oversight of the pilot consider testing the product at a subset of the financial institution's locations or branches. This will also facilitate working out the kinks in implementation and troubleshooting as things come up.
- **Decide which staff will offer the product.** To ease the early implementation of a new product and avoid disrupting regular operations, consider tasking a select number of staff to offer the product. For example, if you are piloting a new underwriting criteria for a loan, it might be easier to forward applications to one loan officer who can be tasked with making approval

[9] Doran, George T. "There's a SMART way to write management's goals and objectives." Management review 70.11 (1981): 35-36.

decisions, instead of training multiple staff members. On the other hand, if you need to know how well the product will ultimately work at scale, it may make more sense to train multiple people.

- **Determine how long will you pilot the product.** The nature of the data you need to collect from the pilot will dictate how long you pilot for. For example, loan delinquency or changes in account balances will take longer to measure than product take-up or referrals.

DURING THE PILOT

- **Collect quantitative data on customer demand and product usage.** To get a sense for whether the product is meeting customer demand, work with the financial institution to measure the take-up and usage of the product. Formally tracking take-up (those enrolled out of offers made) may mean adding a new field to intake forms, updating enrollment forms, or creating a new product code in the institution's system.
- **Collect qualitative data from staff and users.** Data from focus groups, interviews, or surveys can be a nice supplement to (but not a replacement for!) quantitative data when gauging the demand for and perceptions of a new product. If possible, speak to both new and existing clients, and gather qualitative data throughout the life cycle of the pilot.
- **Keep track of the actual costs of implementation.** Researchers and project teams should identify ways to keep track of the budgeted and actual spending throughout the pilot, to get a sense for the actual costs at scale.

AFTER THE PILOT

You have offered your product, collected quantitative and qualitative data, and now it's time to make decisions about how to move forward. If you have determined that the product merits further piloting, or scale-up, think about how the product, or certain processes, might change for the next iteration. Sit down with staff to identify major pain points, and brainstorm solutions or alternatives that might streamline and simplify the process.

RCT Feasibility

As mentioned briefly in the Partnership Development section, there are interventions that are good to test with RCTs, and there are interventions that are better tested with other methods. One of the goals of any pilot study (except for product piloting) should be to determine whether an RCT is feasible for the intervention in question and test how the full study might run. To help you

CASE STUDY: COLLECTING QUALITATIVE DATA

While discussing the prototype of a new loan product with credit union members before the pilot, we learned that the name of a new loan product triggered a negative reaction from members. With this in mind, the project team revised the marketing strategy to make sure the product was perceived favorably in the eyes of its customers.

After several months of piloting the product, the project team decided to revise the terms of the loan to better suit the needs of prospective borrowers. Feedback from staff suggested the initial loan terms excluded a significant portion of the credit union's members. Additionally, based on the performance of the loan, the executive team decided that the credit union could afford to increase the size of the lending portfolio without introducing too much additional risk.

determine whether the full RCT should continue from the pilot, consider the logistics of the randomization, marketing, survey, consent (if applicable), data collection, and any potential sources of bias.

RANDOMIZATION

Whether the organization already offers the product or not, they almost certainly do not randomize offers or access to the product or service, so it is important to test the logistics of the randomization of the product offers or nudges and how this interacts with the partner's current systems. Throughout the pilot, measure and take note of any biases that may arise during the randomization or interferences with the randomization.

MARKETING AND NUDGES

If the pilot involves testing the execution of a nudge or message, consider whether the partner will actually be able to implement this change and the staff compliance and monitoring. The pilot should reveal whether the partner will be able to incorporate this behavioral change into their workflow, and should be modeled after the full-scale RCT as much as possible. This is inextricably linked to staff effort—if proven to be effective, the research team will not be monitoring the intervention indefinitely, and so it is helpful to determine how compliant the staff will be going forward and how well the intervention fits into the organization's current culture.

THREATS TO DESIGN

The pilot is an especially good time to begin to anticipate potential sources of noncompliance, attrition, and the accuracy of treatment targeting. Identify how people might enter or exit the treatment group, especially in ways that were not considered during the evaluation design. In financial contexts its often difficult, or even illegal, to completely prevent access to the treatment, so use the pilot to monitor how participants in the treatment group might avoid receiving or taking their treatment assignment, and how participants in the control group might find ways to enter the treatment sample. Particularly severe or largely inevitable sources of noncompliance, attrition, or spillover are compelling reasons to change the context for the full-scale study or change the evaluation design to mitigate some of the effects of these biases.

SURVEY PILOTING

All surveys should be piloted before being deployed to ensure data are useful and accurate. Even well-designed, previously used questions need to be piloted when being used in new contexts, especially when they have been translated to different languages. Survey piloting tends to be iterative, but there are two main phases: (1) early piloting, when the questions are still being designed, and (2) piloting of the actual instrument. The objectives of piloting a survey are to:

- Determine the length of the survey
- Make sure skip patterns are correct
- Identify potential problems with wording or content
- Verify that respondents are interpreting the questions as intended, and that questions are being interpreted in a consistent way
- Identify questions that do not provide any additional information
- Determine that respondents' interest and attention is sustained throughout the interview
- Determine the "flow" of the interview, i.e., do questions lead naturally from one to the next? Are certain questions influencing answers to future questions?
- Identify potential stumbling blocks

for the interviewer, i.e., do certain questions need to be repeated? Are there misunderstandings with certain words?

  - Identify common responses to open-ended questions so they can be pre-coded

These objectives can help to set the goals and pace of the pilot. Other things to consider include:

- **If the questions are brand new, consider using focus groups at first.** Focus groups are especially important if there are questions that address subjects the research team has little to no experience measuring, or if there are no preexisting survey questions that have been used to test this. Focus groups are used to gauge respondents' understanding of each question, and to determine if there is a reason for them to lie about a response or mislead in a preventable way.
- **In general, the survey should be piloted to at least 15 individuals if the survey contains previously-used questions, and to at least 30 individuals if it is new.** Typically we try to not use participants from our actual sample during this process, but they should be as similar to and representative of the population that will eventually be used in the study as possible.

CONSENT PILOTING

Make sure to test the consent script during the pilot. There is a lot of overlap with what to look for in piloting your survey and piloting your consent script; in short, ideally, the process of obtaining consent will be integrated seamlessly with your intervention. It might be difficult or impossible to get around some of the required elements in your consent script depending on the risk your evaluation may pose to participants, but even small changes in content, timing, and the format of the consent script can have an impact on the consent rate and take-up of the intervention. Some general guidelines for creating an effective consent script are located in the Data Collection section.

In addition to enrollment rates, make sure to keep in mind your study's retention rates during consent piloting. Your consent script will not only give subjects an idea of what data will be collected and what their participation entails, but it will also "sell" your study to subjects. A script that does not accurately or fully explain the time requirements of your study might result in low retention rates for the length of your study, even if your study has initially high enrollment rates.

DATA COLLECTION

Another goal for the pilot should be to collect outcome data. As is discussed in the Research and Evaluation Design section, power calculations can be conducted or updated using data collected during the pilot. Additionally, data collection at this early stage can help gauge the variance in outcomes and refine outcome measures. While piloting the general data collection, the survey, and the consent script, make sure to take note of participant questions that arise consistently. Participants will inevitability ask questions about the study, product, and process, and the implementing staff may not know how to answer without some guidance from the research team. Researchers can address these questions by updating FAQs, creating and updating marketing materials, or clarifying common questions through a different survey or consent design.

Timeline

Timing the pilot correctly is key for preserving momentum while moving into an RCT, but doing so involves anticipating the possible outcomes of the pilot period and being thoughtful about a good start date for the full scale study.

- **Ideally, the timing of the pilot will leave room post-pilot for alterations to the research and evaluation design.** Timing the pilot correctly is tricky; too soon and the partner and research team lose momentum moving into the RCT launch, too late and there is not sufficient time allotted for important and sometimes necessary changes to the implementation and design. Because of this, it can be helpful to consider all the ways that the pilot may "fail" before setting the timeline of the pilot. In any case, we typically leave about two weeks between a well-planned pilot and the RCT launch, but more time may be needed for new products, if the goal of the pilot is to determine whether a given intervention lends itself to RCT-testing, or if the timing of the RCT is contingent on another external factor outside of the research team's control.
- **For any evaluation, it is important to know when to delay or pause a full-scale RCT launch.** As mentioned previously, outlining each possible pilot "failure" and determining the appropriate course of action from each is important, but it can be difficult to separate which part of the implementation is causing problems. Piloting component-by-component can provide a good assessment and picture of which part of the full evaluation may create issues. Testing the data collection piece separately from the messaging of the project, for example, will give the research team separate data on the efficacy of each and be crucial in making a decision like pausing the launch.

Pilot Checklist

- □ Product Pilot
    - □ Identify the value proposition for the financial institution
    - □ Set benchmarks to evaluate pilot successes and failures
    - □ Set up systems to collect quantitative data on customer demand and product usage
    - □ Collect qualitative data from staff and users
    - □ Record the actual costs of offering the new product
- □ Evaluation Pilot
    - □ Test the logistics of the randomization, product offers, or nudges
    - □ Assess the potential for treatment noncompliance, attrition, and spillovers
    - □ Discuss pilot results with the research team and partner staff. Where applicable make alterations to the research design and re-pilot
- □ Survey Data Collection
    - □ Pilot the survey instrument with a sample of respondents
    - □ Prepare a survey manual

Ongoing Management

*The **Ongoing Management** section provides suggestions for successfully managing your project throughout its life.*

Your evaluation plan is finalized. The IRB is signed off. Your pilot is complete. Your project materials are set to go. You are now ready to launch your RCT! In this section, we provide some tips on how to monitor your survey, intervention, and data, as well as how to maintain staff engagement throughout the life of the RCT.

Monitoring

To ensure that a randomized study is internally valid, researchers must have evidence that the treatment is being administered correctly (i.e., participants are receiving the treatment to which they were assigned). Monitoring the data collection is a method of obtaining evidence of this; it helps the research team ensure that surveyors are correctly collecting the data they need, data are coming in regularly and accurately, and that the intervention itself is working as planned. Monitoring is especially important where the treatment is administered repeatedly such as text message reminders ahead of loan payment due dates. The frequency and type of monitoring depends on the nature of the intervention, your partner organization's ability and willingness to comply with study protocol, and the nature of the study. Below we give some tips on how to monitor each aspect of the project's implementation.

DATA MONITORING

- **Conduct high-frequency checks.** High-frequency checks provide information about the quality of the data, surveyor performance, the CAI survey program, and the data flow. High-frequency checks check trends across all surveys, rather than within each survey. You should write the code for your high-frequency check before data collection begins in a programming software like Stata or Python, as the launch of the survey and baseline data collection period are often extremely busy. The check file should be run as often as possible—ideally, on a daily basis.
- **Check the merge early and often.** It's tempting to collect all data before trying to combine different datasets together. If you do, however, you may find that your data does not fit together as expected because of problems like missing data, duplicate observations, incorrect queries, or high attrition. Try to automate this process by writing program files that recreate the same merge at consistent intervals over time and run basic tests to ensure that the data consists of what is expected. See the pullout on the next page for some tips on what to check often.

CORRECTLY PROMPTING RESPONDENTS

Sometimes when a respondent provides an unclear answer to a question, a surveyor may be tempted to either guess the appropriate answer or lead the respondent to a certain answer. For example, let's say that one of our survey questions is: **"How would you describe your overall financial situation? Would you describe it as excellent, good, neutral, fair, or poor?"**

Our respondent answers: "It's okay right now. I've been worse, but I could be better too."

Wrong: A surveyor would be incorrectly probing the respondent if they respond with something like, "So, it sounds like it's good!" Or, "Eh, that sounds neutral to me, would you agree?" In both of these cases, the surveyor has introduced their own bias into the response of the participant.

Right: An effective way to elicit an answer to a survey question without pushing the respondent toward any one answer is to simply repeat the survey options. A good response to the answer above might be, "I see, so, would you say your overall financial situation is good, neutral, fair, or poor?" This narrows down the possible choices according to the information that the respondent provided without pushing the respondent into any one answer. Repeating the answer choices in this way typically elicits a clear response, and if not, doing it a second time helps as well.

SURVEY MONITORING

Sometimes, for one reason or another, surveyors may falsify survey data or try to cut corners while administering the survey. Survey accompaniments, random spot checks, and back checks are quality control measures that help ensure high-quality survey data collection.

- **Accompany surveyors.** Especially at the beginning of any survey period, aim to spend time with your surveyors as they administer surveys. Accompany each surveyor for the duration of the entire survey. Make sure they understand the purpose of each question and ask the questions in a clear way that respondents understand. Pay particular attention to how the surveyor may prompt or probe respondents. Participants might indicate that they misunderstand questions by giving answers that don't address the question. Surveyors should be prepared to answer these appropriately without

Data Monitoring Checklist

High-frequency checks	□ The unique IDs are actually unique, and any other variables that should be unique are actually unique. □ The variables that you are merging on (e.g., survey ID, name, address, and account number) uniquely identify each observation in each dataset. □ Certain variables that should not have missing values do not in fact have missing values. □ Double-check skip patterns, survey logic, and hard checks (if you are using CAI). Hard checks are value restrictions or other restrictions that prohibit out-of-range responses and do not let surveyors continue without correcting. Skip patterns and survey logic both refer to the flow of your surveyor; you should develop a method for ensuring that the appropriate questions are being skipped, and that the skip instructions refer correctly to the questions they should. □ Check interview dates. Interview start and end date should be the same, the interview date should not be before the start of data collection, and that interview dates should be close to the system date (when they were uploaded or entered). □ Check the percentage of “don’t know” and “refusal” values for each variable. □ Check for unusually short or long survey durations, average durations, surveyor productivity, check the number of program interruptions by surveyor, if applicable.
Merge checks	□ Do the same clients appear in every dataset (demographics, accounts, or transactions.)? If a client is missing from one dataset, why? □ Is the data logically consistent between sets? For example, do all of the transactions in an account sum to the month-end balance? □ Do the same variables appear at each point in time? Do they move over time in expected ways? □ What are the expected bounds of each variable? What are the outliers, and what can help to explain them? □ Does the randomization appear to be balanced? Is it generating the expected proportions of participants in each study arm?

leading respondents to certain answers.

Survey accompaniments also serve as an opportunity to learn more about the study population and the measurement tool. As an example, during one set of survey accompaniments researchers noticed patterns in customers' responses to a product offer question. The question was originally designed to include a binary yes-or-no response but researchers decided to add more detailed responses to the survey question to more formally capture the reason for product refusal.

- **Conduct random spot checks.** One way to prevent shirking or discourage sloppy survey work is to make frequent and unannounced visits to the survey location. During our own spot checks, we have identified surveyors who were absent from their survey station, late, or falsifying survey data. Spot checks are also a good opportunity to check in with surveyors and solicit feedback about the survey process.
- **Conduct back checks.** Also known as a field audit, a back check involves revisiting some respondents to ask them a few questions from the survey and matching their answers with the originally collected responses. Back checks hold surveyors accountable by comparing original responses and also test the robustness of the survey instrument. See page 73 for a brief overview of running back checks, with suggestions on how many respondents to contact and what type of questions to check.

MONITORING COMPLIANCE WITH RESEARCH PROTOCOLS

- **Check administrative data or the intervention's meta-data to verify treatment compliance.** You may be able to verify treatment administration via administrative records. For example, for a text message intervention, researchers might use a report of outbound SMS's to verify that each customer received the correct message.
- **Audit the treatment assignment of your respondents.** Make sure that the implementing partner's administrative or process records of customers' treatment assignment matches the research team's randomization.
- **Observe a sample of the intervention in-person.** High-touch interventions may benefit from periodic in-person monitoring. For example, researchers may consider listening to or recording outbound calls to check that the correct product offer scripts were administered. Privacy laws and concerns surrounding the disclosure of financial information may make on-site observation difficult or impossible. If you cannot observe the intervention directly, you might be able follow up with clients afterwards. For example, if the intervention involves a product offer, you could call and ask, "Have you heard of "XYZ Program? Where did you hear about it?"
- **When checking treatment compliance, select a random sample of people and verify they received the correct treatment.** Where possible, we suggest stratifying your monitoring sample by branch or teller, for example, to detect systematic differences in the delivery of the treatment. If the compliance rate is low, re-sample and re-test treatment administration. If you find early on that compliance is near perfect, you may be able to increase the intervals between regular monitoring visits. If you find systematic discrepancies with treatment implementation, you will need to follow up with the research team and implementing partner. Low treatment compliance could mean that some aspect of the implementation is impractical or

How to Conduct Back Checks

During a back check, a member of the research team (someone other than the original surveyor) revisits a respondent to administer a selection of questions from the original survey. The research team then compares the back check response with the respondent's original response. The selection of back check respondents should be random. Back check questionnaires should be short (aim for less than five minutes) to avoid respondent fatigue and annoyance.

TYPES OF QUESTIONS

There are three types of questions to include when designing your back check questionnaire:

Type 1: Respondent and interview information. The answers to type 1 questions should never change, regardless of the interviewer, location, or time of day. A high error rate (our rule of thumb is 10 percent or more) might be an indicator that the interview did not occur. Examples include: date of birth, race, and, typically, gender. If applicable, confirm the respondent received the correct compensation for participating in the survey.

Type 2: Survey and surveyor performance. These questions are intended to assess how well the surveyors administered the survey, and how well respondents understood the survey. The responses to these questions are unlikely to change, but they are questions where surveyors may be tempted to cut corners, potentially due to complexity of the question. Additionally, type 2 questions ensure you are all on the same page about the meaning of certain questions. During one survey, we discovered that some members of the survey team misunderstood the question and administered it incorrectly without realizing it. Examples include: Categorization questions (i.e., the surveyor categorizes the respondent's answer), questions with a lot of examples, and questions that prompt skip patterns.

Type 3: Key outcomes and metrics. These questions test the stability of the measure. Responses are likely to be key outcomes and may or may not change over time. If answers do change over time, the research team might be interested in understanding the trends. If you notice high variance in type 3 questions, it is important to adjust for this in the analysis of these questions. Examples include: self-reported income, savings, debt, and behavioral questions.

We recommend you use a variety of questions from Type 1, Type 2, and Type 3, and strive to keep the entire back check under five minutes per participant. Collaborate with the rest of the research team to determine which questions will best paint a clear picture of your data collection. Ideally, you will have multiple versions of the back check questionnaire, but this is even more important for bigger surveys. Make a list of roughly 50 questions to be back checked, for example, and then allocate 10-20 per back check questionnaire. If the question order could affect responses, keep the order the same in the back check survey as it is in the original survey.

LOGISTICS

How many back checks should you conduct? IPA suggests administering back checks on 10 percent of surveys, but the exact number will depend on the length of the survey (longer surveys may require a smaller portion of back checks) and the stage of the survey period. You might

How to Conduct Back Checks Continued

want to back check more intensely during the first weeks of your survey. Timelines and budgets should also be taken into account when determining your back check protocol.

When should you conduct back checks? Ideally, back checks should be done within one to two days, and no more than one week, of the initial survey.

How? While back checks are among the gold standard of ensuring the data quality, we do not want to understate the difficulty of conducting them in the US. One option is to conduct back checks over the phone. It is difficult to connect with respondents on the phone. We typically budget for three attempts per respondent which, within the time frame of a week, may be hard to pull off. Consider mailing surveys to respondents or conducting back checks via two-way text SMS**.**

ANALYSIS PLAN

It can be tricky to know whether changed responses are simply a product of the respondent not being sure in their answers, actually changing their mind, or a surveyor error. Create your analysis framework in advance and determine what you would consider to be an error. Establish a range of acceptable deviation for every relevant back check question. There will be some variables for which there is no range of acceptable deviation (e.g., ethnicity). Work with the research team to determine what these ranges are and what you anticipate. If you don't have an intuition for the quality of your questions or what these discrepancies might be before you begin, look at the distribution of discrepancies and use the standard deviation to set your rule.

Analysis for each back check question will depend on which type it falls under:

Type 1. If the error rate for these questions is more than 10 percent, there may be systemic issues in your survey or survey administration. Examine error rates by surveyor and question to narrow your focus to the poorest performing surveyors and questions.

Type 2. Examine each of the Type 2 variables separately. The error rate guideline (10 percent) is the same for these types of questions.

Type 3. Examine overall error rates for these questions and perform stability checks on the variables to look for significant differences between your back check and original data.

ADDRESSING DISCREPANCIES

How should you respond to high discrepancy rates? If a particular surveyor is responsible for high discrepancy rates, audit additional surveys completed by this surveyor. If they have greater than 20 percent discrepancies in the additional audits, audit all surveys completed by the surveyor and re-do the ones with greater than 20 percent discrepancies. Put the surveyor on probation or give them a strict warning. If a surveyor has more than 40 percent discrepancies, fire the surveyor responsible and re-do all surveys with more than 20 percent discrepancies. If high error rates are caused by particular questions rather than particular surveyors, meet with the research team to consider rewording the question or adding additional rounds of surveying. Whatever conditions you decide on for addressing a surveyor's performance, make sure they are incorporated into the surveyor's hiring contract.

CASE STUDY: TREATMENT AUDITS

An evaluation of a savings account in New York required manual assignment of customers to their treatment group. After randomly assigning customers to one of four treatment groups in Stata, research team members manually input each customer's assignment in our implementing partner's customer relationship management (CRM) database. Since the assignment was done manually, human error and a few slips of the fingers meant that a small number of customers were assigned to the wrong treatment group. To audit the treatment assignments, we downloaded a list of customer assignments from the CRM regularly (daily-weekly, depending on enrollment volume) during the launch phase and cross-checked the list with the randomization records. We then sent a list of incorrectly assigned customers to the CRM administrator, who subsequently re-assigned customers.

infeasible, or that there are problems with staff buy-in.
Low treatment compliance can be addressed in various ways in the analysis phase, but if it is detected during data collection then consider adding an extra marketing push or brainstorm with the key contacts at the implementing partner to see their recommendations for increasing compliance with the product or service.

- **Periodically cross check power calculations and design assumptions with the 'real-time' data.** During the enrollment and data collection period, verify that the numbers used for the initial power calculations still hold. For example, if your sample size is sensitive to product take-up and the project's actual take-up is lower than originally projected, you can recalibrate your enrollment plan or update outcome variables. In some cases, you can address low sample size this by re-training surveyors and incenting them in some way to subscribe more participants. Anticipating this during the survey and data collection phase can help give the project team enough time to readjust the design and the data collection strategy, if necessary.

MONITORING THE PRODUCT OR SERVICE

During an RCT we ran on promoting access to saving accounts, we learned that technical glitches had prevented customers from making withdrawals from their accounts for several days. While this was clearly a major concern for our partner, it was also a concern for us, as the resulting loss in trust could reduce uptake of the savings account. Understanding how well the implementation of the intervention is going is important. This means being aware of elements outside of the control of the research and implementing partner staff that can affect the research.

One method of checking the logistics of your implementation is via random spot checks. This involves visiting surveyors and your implementing partner unannounced to monitor a few offers or openings in person. Another method, if visiting in-person is unfeasible, is to set up consistent check-in calls with the people administering your intervention (in the above example, it would have been the frontline tellers). We ultimately suspended enrollment to correct these errors and then restarted once systems were in place. This is an effective way of conserving your budget and making sure that your data quality is as high as it can be.

Maintaining Staff Engagement

One of the most important tasks of ongoing project management is maintaining staff engagement. The project will, as a whole, be more successful if your partner maintains a high and consistent level of interest in the evaluation. Issues will inevitably arise throughout implementation, and it's great to have them on your team to help you sort through these.

Share non-sensitive data with partners throughout the evaluation. Most implementing partners will administer an intervention or transmit data to researchers for some time before results are available. To maintain partner interest and buy-in for the length of the evaluation, consider sharing preliminary data at regular intervals throughout the project's duration. We typically do not recommend sending entire datasets—partners are generally more interested in summaries and overviews than they are with the specifics of your dataset. In addition, make sure to be careful with sending any data with Personally Identifiable Information (PII). If your partner does want to see a more detailed dataset, you still may only want to include the aggregate data, even if the data are de-identified, to avoid any breaches in data security. Here are a few reports that you can share with partners before the final analysis stage:

- **Survey Reports.** If your evaluation includes a survey of the financial institution's market, consider sharing a question-by-question summary of the baseline or midline data.
- **Quarterly Reports.** Most financial institutions have quarterly board meetings. Briefing management and preparing up-to-date reports ahead of these meetings can help secure buy-in from the institution's senior management or executives.
- **Data Dashboards**. Build an infrastructure that can automate the monitoring process for partners by feeding real-time data into a dashboard that displays key metrics. For example, when surveying, SurveyCTO can integrate Google Fusion Tables to stream data into graphs and create charts that are updated as soon as data are submitted from devices.

CONSIDERATIONS FOR SHARING PRELIMINARY DATA WITH PARTNERS

Be careful that the data you share does not affect the implementation of the research. It's not hard to imagine a partner who sees that treatment A is outperforming treatment B, and then diverts resources or attention away from one treatment to another before the final results are in. To prevent this, avoid sharing metrics on the different treatments too early, or make sure that the people with whom you are sharing these metrics understand that results may not be statistically significant. In this case, we especially recommend sharing aggregate data as opposed to individual-level data.

Solicit feedback about which data points are interesting to your partner. Researchers tend to focus on individual-level data but financial institutions may be more interested in aggregated data, such as total loan volume, total deposits and withdrawals, and annual growth. Use the metrics that your partner expressed interest in during the evaluation design phase and try to incorporate them in your regular communication.

Project Documentation

Detailed project documentation is key to planning and successfully managing an RCT from start to finish. A project work plan, project manual, and project log will help you stay on top of tasks and keep track of the decisions you make during your project. The best way to maintain these is to create them well before the start of the actual intervention, so they serve to track key decisions and changes made during the preparation stage which may impact the study and analysis later on.

Some of these documents might feel so similar that it's hard to remember the differences between them and the purpose of each. The call-out box on the following page outlines differences in types of project document.

PROJECT WORK PLAN AND CALENDAR

With the large number of tasks and activities that go on simultaneously throughout your study, it is imperative to stay organized and aware of tasks that have been completed, tasks in progress, and tasks that are upcoming. Before you begin your activities in the field, we strongly recommend that you create a detailed and comprehensive schedule of all start and completion dates for each task, along with a list of who is responsible for each task in the narrative work plan. This could be done a number of ways; at IPA, we typically use a Gantt chart. For an example, see the Appendix.

PROJECT LOG

Research staff are subject to high turnover, and sometimes projects go on for many years, so a project log is essential for keeping track of the rationale for ongoing decisions that contribute to the outcome of your entire RCT. The project log serves as an up-to-date transcription or diary that tracks decisions made about the research, such as changes to the design or challenges that arise.

Research is an inherently iterative process and it is likely that the implementation of your intervention will vary from the original research design. As a rule of thumb, any decision made that could affect the research should be documented. Documenting the actual (versus planned) details, activities, and decisions about evaluation design, implementation, and data collection ensures that anyone trying to understand the project in the future can do so. Many of these details also must be reported to the IRB, so keeping track of them in one place will also help to ensure that no important changes to the intervention are accidentally overlooked and unreported. Please see the Appendix for a Project Log template. Below, however, are some suggestions for what we recommend including in a project log. Each change to the study should also include who the key decision-makers were.

- **Changes to the study design or intervention**: at a minimum your project log should detail (1) why these changes were made, (2) exactly what the subsequent changes were, both in design and workflow, and (3) how this change affects the data collection and outcomes.
- **Changes to the initial project timeline: changes to project timelines are incredibly common.** Your project log should at least include (1) what necessitated the changes, (2) what contracts and budgets were updated or created as a result, and (3) any subsequent changes in data collection and outcomes.

DIFFERENCES BETWEEN PROJECT DOCUMENTS

Evaluation Plan

This outlines the entire evaluation, from the pre-launch materials and the pilot through the analysis stage. Important specifics to address in the evaluation plan include the details of the research design (including the treatment arms and study outcomes) as well as justifications for the different components of the design.

Work Plan/Project Calendar

The work plan specifically addresses each task that needs to be completed for the evaluation, along with who will be responsible for each activity.

Pre-Analysis Plan

A pre-analysis plan is essentially a detailed version of the analysis section included in the evaluation plan. The pre-analysis plan differs in that it typically includes models, specifications, and regression equations to be used in the analysis stage. Different outcome variables and their coding can be described here, as well as the power calculations for each.

Project Log

In contrast to the above three documents, the project log is useful for keeping track of the day-to-day decisions and changes that occur and meeting notes.

Project Manual

The project manual can be thought of as a high-level and retrospective project log. The project manual should include details of the original evaluation plan, the work plan, the pre-analysis plan, and how the implementation actually happened, using the major occurrences detailed in the project log. It is the piece of project documentation that should remain ‘timeless,’ even after your evaluation ends—someone unrelated to the study should be able to pick up your project manual and understand the justification behind the evaluation, the major outcome variables, the basic implementation, any changes that happened in the field, and the form and fit of the datasets.

- **Changes to organizational or political environment:** staff turnover might also be common in the implementing organization. Make sure to keep an updated contact list and contact information for outgoing and incoming staff.
- **Consistency of intervention implementation:** how did the implementation change week-to-week? Did certain tellers or staff members administer the intervention in different or unexpected ways? Essentially, any discrepancy between anticipated implementation and observed implementation should be tracked in the project log.
- **Data questions and decisions: the project log provides detailed background information for some of the observed results in the dataset that might not be initially obvious.** On a recent project, some examples of data notes in our project log included why a batch of surveys was re-coded, explanations for why various N’s didn’t sum to the expected totals, information about why subjects attrited from different data sources, and the rationale for when certain pieces of data were received.

- **Check-in meetings and phone calls**: many decisions will be made through these check-in meetings, and the project log is a great place to keep track of the dates of these meetings and the key decisions made.

PROJECT MANUAL

The project manual is the single document that captures all the study's key information. Research projects are subject to staff turnover and divisions of labor, so the person managing the evaluation in the field at the start may not necessarily be the person who is there during the analysis stage. Detailed project documentation can avoid the risk of erroneous assumptions and repeated mistakes. Additionally, a project manual is an archive of the logistics of your study and can be a helpful guide to you or others when developing future studies.

The project manual typically builds on the content the evaluation plan, but should also include the administrative and logistical details of the study. For more information, please see the Project Manual Template in the Appendix.

Ongoing Management Checklist

□ **Regularly update the Project Log to document the entire evaluation from start to finish**

□ **Check administrative data or the intervention's meta-data to verify treatment compliance**

□ **Audit the treatment assignment of your respondents**

□ **Share data with your partner throughout the evaluation**

□ **Track the evaluation's schedule of events, along with a list of who is responsible for each task, in the Project Manual**

□ **If surveying, conduct back checks, spot checks, survey accompaniments and/or high frequency checks to ensure the quality of your survey data**

__Wrapping up__ your project includes data analysis and dissemination of results, as well as tying up loose ends with your implementing partner and ensuring your datasets are in line with research transparency guidelines.

Wrapping Up

Congratulations! You've finally come to the end of your data collection and you're ready to start analyzing your results. As you wrap up your project, this section provides tips on how to tie up any loose ends with your implementing partner, make sure your data are well-documented and ready for publication, and think about your audience for the dissemination of your results.

Partner Wrap-up

Your partner interactions shouldn't end with the last data pull. In addition to circulating the final paper, presentation, or project brief with your partner, we suggest making time for the following activities:

- **Organize a meeting with your partner and research team to identify the project's most significant successes and complications.** Take this time to both examine how well the intervention was implemented and talk about the evaluation (randomization and study protocols). Your Project Log will contain the running list of things that came up throughout the evaluation, but once things have wrapped up researchers and partner staff should be in a good position to identify the roadblocks that didn't amount to much in the end, and those that ended up being a real headache. This is also good time to ask your partner to "rate" the research process. What part of the evaluation took longer than expected? Did staff receive the right amount of training to be able to successfully comply with the study protocols? How would they have set-up the study differently?

- **Don't forget about frontline, IT, and support staff!** Ask your implementing partner about the best way to share results or next steps with the rest of the organization. It would be a shame to leave frontline staff to wonder what came out of their efforts to comply with your strict randomization protocol. And, if you cannot do so directly, ask management to thank frontline and IT staff on your behalf, for their involvement throughout the project.
- **Help your partner develop the systems required to continue collecting and evaluating data.** Due to resource constraints it is unlikely that research staff will be able to provide partners with summary reports with the same frequency as during the evaluation. However, to continue promoting evidence-based decision-making at your partner organization, think of ways to build their capacity to monitor and evaluate their products internally. Ask your partner if they found any reports particularly helpful or useful and try to identify a way for them to continue generating the information after the evaluation ends. You could create an Excel file with macros that auto generate graphs and charts, train IT staff to create dashboards and summary reports, or create a resource list that they can use to continue monitoring and evaluating the product after your engagement with the project ends.

Administrative Wrap-up

As you wrap up data collection and conduct your data analysis, we recommend updating the Project Manual. As a refresher, the Project Manual captures all of the key information about your research study. This includes research motivation, context, and design, data collection instructions and timeline, links to all relevant documentation (surveys, codebooks, project logs), and a list of research and implementing partner staff who have worked on the project.

During data collection wrap up make sure your Project Manual includes:

- Up-to-date contact information for all key stakeholders
- The location of all raw and cleaned data files
- Documentation of how different data files connect to one another. This is especially crucial if you have used different programming files (.do files if you are using Stata) throughout the project to clean and compile data on an ongoing basis
- Any known issues with the data and how you dealt with them

We are reiterating how important it is to keep this information current and easily accessible because we have seen too many projects where data or files have been stored in hard-to-find places, raw data are misplaced, or key decisions regarding the data, or even the implementation itself, are forgotten. In short, running an accurate and replicable analysis of the data later on becomes virtually impossible when many of these pieces are missing.

If you are working with co-authors, or there are multiple research assistants working on analysis either concurrently or sequentially, decisions regarding data flow and organization become doubly important. It is always worth the extra time to try to stick with a consistent file structure for your data, cleaning, and analysis files. A great resource for establishing consistent file structures, naming styles, and general workflow is J. Scott Long's *The Workflow of Data Analysis Using Stata*.[10]

In addition to finalizing the Project Manual and making sure your data files are easy to read, a few other administrative items will need your attention as you finish up your project:

[10] Long, J. Scott. *The Workflow of Data Analysis Using Stata*. Stata Press books (2009).

- **Close the human subjects review.** Human subjects research projects should be closed (or notified that continuing review of the study is no longer needed) when all data have been collected and identifying information is no longer needed. Once a study has been closed, no new data can be collected, study participants cannot be contacted, and researchers cannot access PII. To close the project's human subject's file, you may need to complete a project closure report.
- **Erase and destroy PII and sensitive data.** Researchers should shred paper files and erase electronic files using a secure file eraser. If you are transferring or disposing of computers or portable storage, request help from an IT professional to make sure all disks have been cleared.
- **Make sure you are compliant with any records retention policies.** Federal regulations and donor, funder, university, and research organization specific policies may require that certain types of reports and records be retained for a specified period of time. Research records, such as consent forms and signed disclosures, and legal documentation, such as NDAs or MOUs, may need to be retained or destroyed within a certain time frame. Check with your legal or compliance team to determine the retention period of any documents and make sure they are handed off to the appropriate person if needed.

Research Transparency

In May of 2015, the retraction of the *Science* article by LaCour and Green following the discovery that Michael LaCour may have fabricated the data led to a firestorm of media attention and many concerns about the validity of the peer review process and of social science results. Thankfully, such outright fraud is rare (as far as we know), but the case did serve to highlight the importance of making experimental design and data available for use by other researchers: it was the attempted replication of the study that led to a deeper and more thorough analysis and uncovered the study's inconsistencies.

Even in cases where fraud is not present, open sharing of data is an extremely valuable tool for social science researchers. It enables verification of any code used to run analysis, reproduction of results, and checks of robustness to different statistical specifications and sample populations—in short, it allows us to test our confidence in the study's results.

At IPA, we take research transparency very seriously and require that data for all our research projects be shared in our public repository within three years of the completion of the study. The open sharing of data and code goes a long way toward making research replication possible and enabling validation of our research results. Additionally, if researchers intend to publish in an academic journal, many journals now require that submitters make their data and analysis files publicly available. On our website, our Research Transparency Initiative has already compiled a number of resources to help researchers prepare their data for publication. The good news is, if you have been following the guidelines for data management listed in this toolkit, you're already in great shape!

As an example, the American Economic Association lists the following requirements for publication of papers on experimental results:

- Instructions for research implementation, presented in such a way that, “together with the design summary, conveys the protocol clearly enough that the design could be replicated by a reasonably skilled experimentalist”

- Information about eligibility or selection of research participants
- “Any computer programs, configuration files, or scripts used to run the experiment and/or to analyze the data”
- The raw data from the experiment

If all of these are laid out appropriately in your project manual and filed so that they are easy to find, you will have a much easier time preparing for submission, and other researchers will have an easier time replicating your results.

Disseminating Your Results

There has been a huge amount published about how to increase the impact of research and make it accessible to policymakers, journalists, and the general public. Some of our favorite resources for this are listed in the Additional Resources guide following this section. Beyond all the blog posts and *Freakonomics* appearances that you may have lined up after the wrap-up of your project, it is important to remember that a key player in disseminating your research is your implementing partner. Not only do they often have a separate marketing and communications budget that can help with dissemination, but they also often have the ear of policymakers in their communities who researchers might not have access to. Some things to keep in mind:

- **Give your partner a sneak peek at the results.** At IPA, we strive to uphold the impartiality of our research and publish results that are not driven by what either the implementing partner or our research team wants to find. For this reason, we set up agreements before we begin to conduct research to ensure that we will not be restricted in the publication of our results, even if those results are counter to what the implementing partner might want to disseminate (see Legal Agreements). But that doesn’t mean that you can’t give your partner the courtesy of sharing results before they’re published. If you’ve been maintaining good communications with your partner throughout the study, the results shouldn’t be a surprise anyway, but if you can give them explicit access to results before they go public, your partner will have the chance to think through how they want to respond to the results, and you will have made sure that your relationship with the partner is intact for any future dissemination or research.
- **Train them in the results.** When you work with partners who already have an academic background, it may be possible to simply hand them your journal article with the published results, but for most partners, you’re going to need to create something to help them digest the findings. They’re going to be discussing the results with peers, at conferences, and throughout the organization, so make sure they’re interpreting what you’ve come up with in a way that you’re comfortable with. Taking the time to sit down with them to make sure they understand the research and creating talking points with them that they can use will help ensure that you maintain control over the interpretation of the research.
- **Be sensitive to how you name them in your research.** While many organizations want their names included in the written research results, not all do. Make sure to ask how they wish to be represented and what identifying information you can include in your publications.

- **A change of policy within the implementing partner organization is a policy "win" for your research.** You don't have to solve world hunger for your research to have policy impact. If your implementing partner changes how they are implementing a program or service because of your research, then you have had impact. Even if this current impact is small, it may grow through ripple effects at other organizations.

Conclusion

We hope that this toolkit has given you some guidance for running your own randomized controlled trials. Many other resources exist on running RCTs, and our goal has been to complement these existing resources, not substitute for them. We therefore recommend that you keep reading and learning more. We have included links to many of our favorite resources throughout this guide as well as in the Appendix. We also hope that this guide remains a living resource that users can update with further information as it becomes available. Finally, we invite you to send us your ideas for content changes or additions to usfi@poverty-action.org. Thank you.

Wrapping Up Checklist

- □ Finalize the Project Manual
- □ Close the project's human subject's review
- □ Store and/or destroy PII/sensitive data according to your human subjects protocol
- □ Share the study results with your partner
- □ Prepare your data cleaning and analysis files for publication

Glossary

Administrative data
Data collected by an organization or government agency typically for purposes other than research. In financial institutions, this is typically transaction or account data.

Attribute
A credit report variable.

Attrition
The exiting of participants from the study sample before the evaluation finishes. Account closings, death, data loss, and moving away are all possible examples of attrition.

Back check
Also known as a field audit, a back check involves contacting a set percent of survey respondents to re-administer a random sample of the survey questions to determine (1) whether surveyors are actually administering the survey, and (2) the accuracy of the survey instrument.

Baseline
Also known as a pre-test, the baseline is a survey or other form of data collection that occurs pre-intervention.

Buy-in
Support for or investment in a project or idea.

Codebook

A document that lays out the structure and content of a dataset. This typically includes variable names, labels, changes from raw to clean data, and definitions of anything not self-explanatory in the dataset.

Compliance

Participants that are assigned to treatment and take the treatment (e.g., are assigned to receive information encouraging them to open a loan, and then they actually open the loan), and those assigned to not receive the treatment do not take the treatment (e.g., are assigned to not receive the loan information and do not open the loan).

Computer Assisted Interviewing (CAI)

A method for interviewing in which the surveyor uses a computer or other form of technology to administer the survey (e.g., a tablet or cell phone).

Confidentiality agreement

An agreement that determines whether information shared between two parties may be shared with others and, if so, under what conditions.

Cost-reimbursable contract

Contract in which expenses incurred will be reimbursed, usually up to a certain amount, upon receipt of an invoice. This is as opposed to a Fixed cost contract.

Data Security Protocol

A comprehensive plan to secure your data.

Data Sharing Plan

Outlines the data collection and transfer protocols between the implementing partner and researchers.

Debt Management Plan (DMP)

An agreement between a creditor and a person in debt that results in consolidation of the debt or a repayment plan.

.do file

A programming file used to run computer code for cleaning and analyzing data in Stata.

Effect size

The difference between the average outcome observed in the treatment group and the average outcome observed in the control group.

Endline

The endline is a survey or other form of data collection that occurs post-intervention (versus Baseline, which occurs before the intervention).

Evaluation plan

The evaluation plan outlines timelines, roles and responsibilities, intervention details, and data collection procedures, and acts as the project's central planning document.

External validity

How broadly the results of the study can be generalized (versus internal validity).

Field audit

See Back check.

Field scan

Also known as a market scan, a field scan is a comprehensive review of what similar programs already exist, how they have been implemented, and what has been learned from them.

Fixed-cost contract

A contract in which a certain amount of money will be paid to cover all work and expenses. This is as opposed to a cost-reimbursable contract.

Frontline staff

Staff members at the implementing partner organization who directly interact with clients. This includes bank tellers, member services representatives, financial coaches and counselors.

High-frequency checks

A review of the survey data conducted often (high-frequency) and measures trends across all surveys to ensure that the data being collected is of sufficiently high quality.

High-Touch RCT/Evaluation

Typically refers to an RCT in which multiple people are involved in the implementation of the intervention. For example, an RCT in which staff members from multiple branches of a financial institution are required to offer a product in accordance with a randomization protocol.

Implementing Partner/Organization

The company or group with whom the research team is partnering to conduct a randomized evaluation.

Intellectual property (IP)

Refers to ideas, inventions, literary and artistic works, designs, symbols, names, and images which are developed and/or reduced to practice by an individual or organization. IP may be protected by copyrights, patents, or trademarks, or simply defined as such by memorandum of understanding existing between two organizations or individuals.

Intervention

The intervention is the treatment randomized between experimental groups. An intervention in the finance context might be a loan offer, an encouragement to pay down debt, or financial education.

Information Technology (IT)

The group within an organization that is responsible for the maintenance, security, and control of the organizations technological and computer systems and data.

Institutional Review Board (IRB)

An Institutional Review Board (also referred to as an independent review board, independent ethics committees, ethical review boards, and human subjects review boards) is a group designated by an institution (such as a university or non-profit) to approve, monitor, and review research involving human subjects to assure appropriate steps are taken to protect the rights and welfare of those subjects.

Internal Validity

Refers to how well a study was conducted as well as to how confidently we can conclude that a change in our dependent variable was produced solely by our independent variable and not extraneous ones.

Literature Review

An examination of existing studies, articles, and evaluations that have been conducted related to the question or issue of interest.

Low-Touch RCT/Evaluation

Typically refers to an RCT in which very few people need to be involved in the intervention. For example, an RCT which only involves the randomization of placement of information on a web page.

Member

What credit unions typically use to refer to their clients.

Memorandum of Understanding (MOU)

A preliminary agreement to work together. Typically include a statement of the proposed project to be undertaken jointly by both parties, a scope of work defining the obligations (including any reporting requirements) of each organization, and statements regarding the confidentiality and ownership of information and other intellectual property generated during the course of the partnership.

Midline

A survey or other form of data collection that occurs during the intervention, between the baseline and the endline.

Needs Assessment

A systematic approach to identifying an unmet need of a specific population.

Noncompliance

When people assigned to the control group are able to circumvent the randomization and gain access to the intervention, or when people assigned to the treatment group decide not to take up the intervention.

Non-disclosure Agreement (NDA)

A document used to protect information that must be shared between two organizations in order to determine whether or not they will work together.

Nudge

As defined by Richard Thaler and Cass Sunstein, a nudge is a way of presenting choices in such a way that freedom of choice is retained but individuals are encouraged (often subconsciously) to make choices that are in their best interests.

Participant

A subject in the evaluation.

Partner Organization

See Implementing partner/organization.

Personally Identifiable Information (PII)

As defined by the NIH, PII is information that is personal in nature and which may be used to identify a person.

Pilot

An initial test—often with a small group of people—of a program, service, experimental design, or survey, for the purposes of garnering information about how well the thing to be piloted "works." Questions to be answered in a pilot might include whether or not program is favorably received, whether survey questions return the expected data, and whether the experimental design can be implemented or whether there are unforeseen logistical challenges.

Power

The probability of detecting a treatment effect of a specific size.

Practitioner

Defined here as individuals or institutions that focus on delivering programs or services. See also Implementing partner/organization.

Pre-analysis plan

A document that is typically drafted before the launch of an evaluation, and includes models, specifications, and regression equations to be used in the analysis stage. Different outcome variables and their coding can be described here, as well as the power calculations for each.

Principal Investigator (PI)

The lead researcher for a project.

Process Evaluation
An evaluation that determines whether existing programs are being implemented successfully.

Project Log
A document useful for keeping track of the day-to-day decisions and changes that occur and meeting notes. See the Appendix for Project Log template.

Project Manual
Project Manual captures all of the key information about your research study. This includes research motivation, context, and design, data collection instructions and timeline, links to all relevant documentation (surveys, codebooks, project logs), and a list of research and implementing partner staff who have worked on the project. See the Appendix for Project Manual template.

Sample Size
The number of observations in a sample.

Scale-up
The act of increasing the number of participants or users of a product, service, or program.

SMS
Short Message Service, also known as a text message.

Soft Credit Pull
A credit pull that does not have an effect an individual's credit score

Spillover
Also known as an externality, a spillover refers to impacts on people who are not the direct beneficiaries of a program or service. For example, if someone receives financial education and then teaches what she learns to her neighbor, the neighbor's increase in financial knowledge is a spillover effect of the financial education program.

Spot Checks
Frequent and unannounced visits to the survey location to check the quality of the survey implementation.

Stata
A statistical software program used to analyze data.

Survey Accompaniments
A form of checking data quality in which a research sits with a surveyor throughout the duration of a survey to ensure that the questions are being asked as they were intended to be asked and study subjects' answers are not being influenced inappropriately by the surveyor.

SurveyCTO
A software useful for programming surveys into computers or other electronic devices.

Surveyor
Someone who is hired to administer surveys for the purposes of data collection.

Take-up
The proportion of people who accept an offer of a program, product, or service.

Theory of Change
A logic model that traces the causal pathway from the intervention to the end goal.

Touchpoints
The number of times that two groups or individuals share an interaction.

Treatment Arm(s)

The number of groups of participants that each receive a unique combination of interventions.

Uptake

See Take-up.

Work Plan

A detailed and comprehensive schedule of all start and completion dates for each task, along with a list of who is responsible for each task.

Additional Resources

General Resources

Evaluating Social Programs: Executive Education at J-PAL
R. Glennerster, A. Banerjee, and E. Duflo
http://bit.ly/EvalSocialPrograms

Field Experiments: Design, Analysis, and Interpretation
D. Green and A. Gerber
http://bit.ly/GreenGerber

Impact Evaluation in Practice
P. Gertler, S. Martinez, P. Premand, L. Rawlings, and C. Vermeersh
http://bit.ly/ImpactEvaluationPractice

Impact Evaluation Toolkit
World Bank
http://bit.ly/WorldBankToolkit

An Introduction to the Use of Randomized Control Trials to Evaluate Development Interventions
H. White
http://bit.ly/DevInterventions

Rigorous evaluation of financial capability strategies: Why, when and how
Consumer Financial Protection Bureau
http://bit.ly/CFPBFinCapEval

Running Randomized Evaluations
R. Glennerster and K. Takavarasha
http://bit.ly/RunningRE

Use of Randomization in the Evaluation of Development Effectiveness
E. Duflo and M. Kremer
http://bit.ly/DufloKremer

Partnership Development & Research Design Resources

Beyond baseline and follow-up: the case for more t in experiments
D. McKenzie
http://bit.ly/McKenzie2011

The core analytics of randomized experiments for social research
H. Bloom
http://bit.ly/Bloom2006

Designing experiments to measure spillover effects
S. Baird, J. Aislinn Bohren, C. McIntosh, B. Ozler
http://bit.ly/MeasureSpillover

Empowering low-income and economically vulnerable consumers
Consumer Financial Protection Bureau
http://bit.ly/CFPBEmpoweringConsumers

The essential role of pair matching in cluster-randomized experiments, with application the Mexican Universal Health Insurance Evaluation
K. Imai, G.King, and C.Nall
http://bit.ly/ImaiKingNall

Guidelines for Conducting a Stakeholder Analysis
K. Schmeer
http://bit.ly/SchmeerStakeholder

Handbook on Planning, Monitoring and Evaluating for Results
UNDP
http://bit.ly/UNDPMonitoring

In pursuit of balance: randomization in practice in development field experiments
M. Bruhn and D. McKenzie
http://bit.ly/BruhnMcKenzie

Understanding statistical power in the context of applied research
T. Baguley
http://bit.ly/BaguelyPower

Making effects manifest in randomized experiments
J. Bowers
http://bit.ly/Bowers2010

Minimum detectable effects a simple way to report the statistical power of experimental designs
H. Bloom
http://bit.ly/Bloom1995

The Programme Manager's Planning, Monitoring and Evaluation Toolkit
UNFPA
http://bit.ly/UNFPAMEToolkit

Tools for Development: A Handbook for Those Engaged in Development Activity
UK Department for International Development (DFID)
http://bit.ly/DFIDToolkit

Data Collection

Data Classification and Examples
Harvard Security
http://bit.ly/HarvardDataClassification

GitHub Guides
GitHub
https://guides.github.com/

GitHub Training
GitHub
https://training.github.com/

Guidelines for Effective Data Management Plans
ICSPR
http://bit.ly/ICSPRDataMgmtPlans

HIPAA Guidance Document
MIT Committee on the Use of Humans as Experimental Subjects
http://bit.ly/MITHIPAA

Is it human subjects research?
U.S. HHS Human Subjects Regulation Decision
http://bit.ly/HumanSubjects

Is the research eligible for exemption?
U.S. HHS Human Subjects Regulation Decision
http://bit.ly/HumanSubjectsEligibility

Manual of Best Practices in Transparent Social Science Research
Garret Christensen and Courtney Soderberg
http://bit.ly/GitHubTransparency

Planning for Electronic Data Collection
SurveyCTO
http://bit.ly/SurveyCTOPlanning

Survey Data Collection for Impact Evaluation
The World Bank
http://bit.ly/WBSurveyData

Pilot, Preparing to Launch, & Ongoing Management

Beyond Design, Behavioral Science for the Pilot and Scale of Product Innovations
Alissa Fishbane and Allison Daminger
http://bit.ly/CFIBeyondDesign

Driving Positive Innovations to Scale in the Financial Services Sector
Allison Daminger, Katy Davis, Piyush Tantia and Josh Wright
http://bit.ly/Ideas42ScaleFinancialInn

Savings and Credit Toolkit: Product Launch
Corporation For Enterprise Development
http://bit.ly/CFEDToolkit

bcstats: a Stata program for analyzing back check (field audit) data Innovations for Poverty Action
https://github.com/PovertyAction/bcstats

Wrapping Up

Bridge the Gap between Research and Policy, One Panel Discussion (and 145 Studies) at a Time
David Evans
http://bit.ly/WBBridge

Data + Design
Trina Chiasson, Dyanna Gregory
http://bit.ly/ChiassonGregory

An Economist's Guide to Visualizing Data
Jonathan A. Schwabish
http://bit.ly/SchwabishDataViz

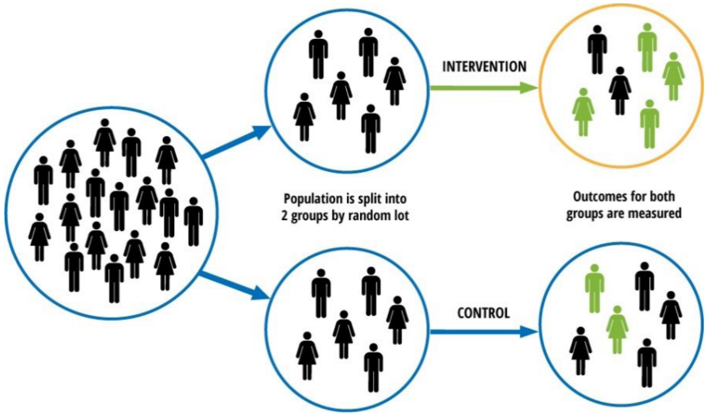
Presenting to policy vs. academic audiences: some thoughts.
Markus Goldstein
http://bit.ly/WBPresenting

PDFs or Not? That isn't the Right Question.
Hobbs, J. David
http://bit.ly/HobbsPDFs

What do White House Policy Makers want from Researchers? Important survey findings
Duncan Green
http://bit.ly/OxfamGreen

What is a Randomized Controlled Trial?

A Handout for Practitioners


INTERVENTION
Population is split into 2 groups by random lot
Outcomes for both groups are measured
CONTROL

IPA uses the most rigorous methodology available to evaluate what works in fighting poverty: the randomized controlled trial (RCT). Also known as A/B tests, randomized trials, or randomized evaluations, RCTs are considered the gold standard in evaluation techniques.

At their most basic level, RCTs are a way of comparing people who receive a product or service (the "treatment" group) with those who do not (the "control" group). Most people are familiar with RCTs from having read about drug trials. Let's say you had a new pill that was supposed to cure the common cold, and you wanted to see if it worked. In order to test out your pill, you would start by finding a group of people who all had colds. You would then give the pill to half of them, while the other half of them received a placebo. You would then monitor them to see if the group that received the pill got better faster than the group that didn't.

The same methodology can be applied to evaluating social programs. For example, one of our partners in Philadelphia wanted to learn if sending text message reminders would help their Debt Management Plan (DMP) clients to make their debt payments on time. They selected half of their DMP clients to receive the reminders, while the other half didn't, and tracked the percentage of on-time payments in each group.

By randomly assigning people to these two groups, we are able to ensure that the groups are identical. This means that, on average, both groups are the same on all observable characteristics (e.g., same gender composition, same average income, same average age), and on all unobservable characteristics (e.g., internal motivation or other factors that cannot be measured). Therefore, any measurable differences in outcomes after the treatment group has received the product or service can be attributed to the treatment itself, rather than to something inherent to the recipients, or to some other external factor.

How to Randomize

The first step in RCT design is to identify the study population. In the example above of providing reminder messages to debt management plan clients, the population registered in the study included every client enrolled in a Debt Management Plan (DMP) with the partner organization. In evaluations that seek to provide a new service, the study population might be, for example, people living under the poverty line in Chicago who express interest in learning about the service.

Once the study population is determined, the treatment and control groups will need to be randomly selected. Randomization really can be as simple as flipping a coin or drawing names out of a hat. In some of our evaluations, IPA has actually held public lotteries so that, if certain members of a community receive access to a service and others do not, the process of selection is transparent, with no appearance of favoritism. In others, we use a computer program to assign names to one group or the other.

However, it often happens that for logistical or political reasons, this kind of basic randomization is not feasible. The following describes several common obstacles to randomization and how we typically approach them.

- **Phase-in design:** Sometimes donors or partner organizations are unwilling or unable to exclude some clients from receiving their service. One option for testing an intervention of this kind is to phase in the program in stages. The first group of beneficiaries would receive the program in year one, the second group in year two, and so on. In this way, everyone in the community eventually gains access to the program, but in the initial year(s) of the evaluation, the second and third groups serve as the control.
- **Cluster randomization:** Not all programs can be provided at the individual level. In these cases, it might be best to randomize at the community or branch level. The downside to this kind of randomization is that it dramatically increases the number of people that need to be included in the study (the sample size), so may not always be feasible. Sample size is discussed in more detail below.
- **Encouragement design:** It is often not possible (or ethical) to force someone to participate or deny participation in a new program. For example, when we evaluated a new savings CD with a partner in New York, not all of the people to whom we offered the CD were interested in opening the account. In this case, the "treatment" that the person receives is not the product or service itself, but merely targeted advertising (such as a discount) designed to encourage them to enroll. Although some of the people in the "encouraged" group may decide not to enroll, and some of the people who don't receive the encouragement will enroll, all that is required is that the encouragement increase the likelihood that the participants will follow through with what they are being encouraged to do (i.e., that the "encouraged" group be more likely to open the CD than the "not encouraged" group). By randomizing encouragement and carefully tracking outcomes for those who do and do not receive the encouragement, it is possible to obtain reliable estimates of the impacts of both the encouragement and of the product or service itself.

How Big? Sample Size Considerations

Let's say I decide to find out what the impact of eating only donuts is on weight. I recruit two people for my study, flip a coin and tell one to eat nothing but donuts for the next month, while the other eats normally. At the end of my study, I find that the person who has eaten nothing but donuts weighs 10 pounds less than the person who ate normally. You would be right if you thought my study design was a little suspect—that one donut-eater could have a very fast metabolism, or the "normal" eater could be eating lots of foods that are unhealthier than donuts.

But if I redid the study with 300 people in each group, and still found that, on average, the donut-eaters lost more weight, then you might be more likely to believe my results.

This is because, when there are only a handful of people in the study, any changes observed might be due to the individual characteristics of those people, or just to chance (for example, even though we randomly assigned people to the two groups, we still could accidentally end up having people with, say, better metabolisms in one group than in the other). But when there are a lot of people, and their average outcome (in this case, weight) still changes, it is much less likely that the change observed is simply a result of chance.

You might have heard researchers talk about statistically significant results. Statistical significance is simply a way of measuring the probability that the result we observe (the donut-eaters lost 10 more pounds than the non donut-eaters) is due to chance. If a result is statistically significant, then it's unlikely that the result is due entirely to chance—that is, you can believe that it's real.

In general, the larger the sample size of the study—that is, the greater the number of people included—the more likely you are to find statistically significant results, assuming they are there to be found. This is known as the power of the study.

OTHER TYPES OF EVALUATION

There are many other ways to evaluate social programs. As an example, imagine that you want to know the impact of over-the-phone financial counseling on credit scores. What are some of the ways you might choose to explore this?

Pre-post tests: In this example, you would first collect the credit scores of the people enrolled in the phone counseling both before and after they received it. Then you would compare the two. What might be some of the problems with this approach? Well, assuming that the participants do, in fact, have higher credit scores after the counseling, you can't tell if this was due to the counseling itself, or if there were other factors that might have played a role. Maybe the people enrolled in the counseling also had a credit-building loan at the same time. Or maybe their credit was already going up, and just happened to correspond with their participation in the counseling

Simple difference: What if we compare the people who received the phone counseling with some other group of people who did not? This approach is known as the simple difference approach, and in some ways it looks a lot like an RCT—you compare two groups, one of which received your program, and one of which did not. The difference is that, with the simple difference approach, the people are not randomly assigned to the groups. The people who received the phone counseling may have signed themselves up, and thus might be different in some way from the people who did not choose to receive counseling. They might have chosen to sign up, for example, because they had worse credit than their peers and wanted help improving it. In this case, the counseling could have helped these people improve their credit, but if we just compare their credit scores with those of others who did not get the counseling, it might appear as if there were no difference because our "treatment" group started off worse. Alternatively, the people who signed up might be more motivated to improve their scores. Higher post-counseling credit scores could simply be an effect of their higher internal motivation, not of the counseling.

There are a number of other non-experimental and quasi-experimental evaluation techniques. Understanding these requires some training in statistics. In general, however, they require making assumptions about (1) what differences might have existed between the two groups prior to your treatment and (2) what other events might have taken place at the same time as your treatment that might also have had an impact.

Going back to the donut example, let's suppose that it is a scientifically established fact that eating donuts causes people to lose 10 pounds. I run my study with 100 people in each group, control and treatment, but I don't find any statistically significant results. How is this possible?[1]

Randomized trials can be compared to microscopes. The more powerful a microscope is, the smaller the objects it can see. This is true of RCTs, too. When RCTs aren't very powerful—that is, when they don't have a large enough sample size—they can't "see" effects that are small. This means that, although the effect is there—although the donuts did in fact cause the treatment group to lose 10 pounds—the result is not statistically significant. Statistically speaking, that effect of 10 pounds looks no different from an effect of zero pounds. So, if the actual effect of donuts were to cause people to lose 50 pounds, I might have been able to detect that, but with my sample size of only 100 people in each group, the 10 pound difference can't be detected—even though it's there.

When researchers talk about doing power calculations, they mean doing calculations to figure out how many people need to be in the study in order to be able to "see" a result of a given size—that is, to be able to say that it is statistically distinguishable from an effect of zero. If it were only important to me to know if eating donuts caused people to lose 50 pounds or more, then I might be perfectly happy with my sample of 100 people per group. But if I really cared about finding out if eating donuts caused people to lose 10 pounds, then I would need to increase the sample size of my study—increase my power—so that my effect of 10 pounds would be visible.

It is important to note that, in the case of cluster-randomized trials, increasing the number of people in the study will not affect the power as much as increasing the number of clusters, which has much larger implications for overall sample size. The size of the effect that needs to be detectable by the study is determined in conversation with the partner organization. IPA researchers will then determine the sample size necessary for the study.

Implementation and Data Collection

Once the sample size has been set and the treatment and control groups have been selected, it's time to implement the intervention. It's important to have monitoring controls in place to ensure that people don't switch from one group to another, and that your staff offer the correct product or service to the correct person. This can necessitate changes to your systems, so it's important to make sure that everyone is on board with the goals of the RCT.

Many RCTs wait until the intervention is complete and then conduct an endline survey to measure the impacts of the intervention on the treatment group. However, increasingly, RCTs are relying on administrative data that may be collected both during and after the intervention. You should be aware that this may require additional work on the part of your staff to pull this data together and transmit it to the researchers.

[1] These numbers are made up and are meant as an example, not as an actual estimate of the sample size needed for this kind of study.

Appendix

I. Partnership Development Questionnaire

Below is a checklist of questions that we often ask of implementing partners to start gauging the answers to the questions we have discussed in this section. Depending on your context, you may only need some of them, or you may decide to ask some in a preliminary conversation and others further down the line. While many of these are designed to help you assess risk, others are there to help your research team design the optimal evaluation.

Partner Organization

- ❑ Please describe your organization and any partners involved in the implementation of this program.
    - ❑ What is the legal structure (non-profit, for-profit) of your organization?
    - ❑ What is your history in the community being served?
    - ❑ What is your history with research and evaluation?
- ❑ Who is/will be the main point of contact for this project? Are there other people who we will be working with?
- ❑ Who is in charge of data, i.e. who will be responsible for running queries and reports on client data? Is there anyone else in your organization who also works in this area and is capable of sending data if necessary?

Program Background

- ❑ How long has this product or service been offered?
    - ❑ How long has the product/service existed in its current state? Please describe any recent changes.
- ❑ Is it in pilot stage or being offered at full scale? If web-based, is it past the initial testing stage?
- ❑ Have there been barriers to successful implementation that have led you to modify this program?
- ❑ How has this program been marketed?
    - ❑ Please describe traditional marketing materials and customer referral incentive structures.
- ❑ What is the process for a consumer to take-up this product or service?
    - ❑ Which employee roles are involved with the offering?
    - ❑ Please describe the number of touchpoints between offering employee and consumer.
    - ❑ When is the product offered/cross-sold? What triggers product offering?
- ❑ What technology is currently involved in the offering of this product/service?
    - ❑ Please describe consumer-facing and/or back-end technology.
- ❑ What are the selection criteria for a consumer to be offered or enroll in this product or service?
    - ❑ Is there an income ceiling for potential customers?
    - ❑ Are bilingual services offered?
    - ❑ What are other disqualifying factors?
- ❑ What take-up have you seen thus far (how many customers/clients are buying the product/using the service)?
    - ❑ What is the rate of take-up and over what time period?
    - ❑ If you believe take-up can be increased, how do you justify this belief? (More specifically, what steps would be taken?)
    - ❑ How has take-up changed over time?
    - ❑ Is the customer base drawn from new or existing (loyal) customers?
    - ❑ If new, how do customers typically find out about program offerings?

Program Background Continued

- ❑ What is the take-up/rate of take-up amongst low-income consumers?
  - ❑ Based on your experience, do you think that overall take-up rates are representative of potential take-up rates for low-income consumers (e.g., is take-up evenly spread across income brackets or is this product/service/feature much more popular among certain income brackets)?
- ❑ How is the product used?
  - ❑ After enrollment, is there a pre-determined plan for subsequent appointments or touch-points, or is that determined as needed by user? (Please note: this question is most relevant to counseling services).
  - ❑ Is this product typically used on a repeat basis? (For example, with prepaid cards, how often do users reload?)What customer data do you collect?

Data Collection

- ❑ How much of an organizational priority is data collection? Please describe training, time investment, and data analysis. Note if there is an existing research unit/department.
- ❑ What data are you already collecting on program use?
  - ❑ Administrative data? Survey data?
- ❑ What is the frequency with which you collect administrative data?
- ❑ What program-specific data do you collect on this program?
- ❑ Do you collect data on your customers from other sources?
- ❑ Do you collect follow-up data post-product/service offering? For how long?

Evaluation Design

- ❑ In order to make meaningful determinations about impact, we would need to be working with ____ people. Do we think we will be able to meet this?
- ❑ Can enrollment be staggered? If so, how many people/month will need to be recruited to achieve adequate sample? How many months of enrollment?
- ❑ Will treatment depend on a client opting in (measuring intention to treat)? If so, what is anticipated take-up rate?
- ❑ What are the spillover risks?
  - ❑ How much does this population talk to each other?
  - ❑ What is the percentage of multi-family homes in the sample?
- ❑ What is the risk of consumers being assigned to multiple treatments (variations on the intervention, which need to be kept separate for research purposes)?
- ❑ Outcome Measures
  - ❑ What metrics would you think we should be using to evaluate the success of this intervention?
  - ❑ Of these metrics, what can be gleaned from the existing account/administrative data? (Again, what do you collect and how? How often?)
- ❑ In your experience, how long does it take to see changes in the outcomes we're interested in measuring? (Minimal period vs. optimal period?)
- ❑ How long do you think it will take to measure for effects plus enrollment of adequate sample if staggered?

Preparation for Evaluation

- ❑ Is there any additional development that needs to happen before you can offer both intervention being tested and the counterfactual option?
  - ❑ Software development?
  - ❑ Staff training?

Preparation for Evaluation Continued

- □ How can randomization and data collection be monitored by the research team?
    - □ Have evaluations with monitoring been done at your organization before? Who would be responsible for coordination around monitoring?
- □ Who do employees responsible for maintaining data integrity to evaluation report to? In order to collect data on any of the outcome metrics we have identified, are any new data systems required? What is needed to set this up?
    - □ Does this require a modification to your procedures (note employee roles)/database/data collection systems?
- □ If the research team has determined surveying will be necessary, what form of implementation makes the most sense to you based on your operations and customer base?
    - □ Carried out by research team (for example, independent surveys in branch lobby) or incorporated into partner operations (for example, included in an application form, mobile-based app, or kiosk)?
    - □ For endline survey, based on your knowledge of your customer base, what would you estimate as success rate for re-contacting clients?
- □ If it has been determined that it will be necessary to collect credit report data:
    - □ Would it make more sense for the research team to manage collecting credit data or do you have capacity to do so?
    - □ When asking customers for consent to collect credit data, do you think we should offer incentives? If so, what incentive do you think your customers would respond to (e.g., previously we've done $5 convenience store gift cards and iPad drawings)?
- □ Will you be able to directly provide the research team with de-identified data? If so, can we do this without soliciting consent from subjects?
    - □ What means of obtaining informed consent would be easiest with your customer base?

II. Pre-Analysis Plan Template

[PARTNER] and [RESEARCH TEAM]

Sample Evaluation Plan Prepared for [PROJECT NAME]

1. General Information

- Title of the project
- Researchers involved
  - Name
  - Title
  - Department
  - Institution
- External partner institutions
- Project staff
- Conflicts of interest

2. Introduction

- Project summary
- Aims, rationale, and background

3. Study Design

- Hypotheses
- Treatment effects and measurement
  - Main variables of interest
  - How outcomes are defined
  - Distinction between primary and secondary outcomes
- Preliminary studies, if applicable
- Details of study
  - Geographic regions
  - Research population
    - Demographic information on target populations
    - Clear rationale for inclusion or exclusion of certain populations
    - Procedures for recruitment and consent
    - Potential benefits and risks to subjects

Pre-Analysis Plan Continued

4. Study Design Continued

- Details of study continued
  - Sampling frame
  - Inclusion/exclusion criteria
  - Withdrawal criteria
  - Early termination criteria
  - Expected timeline
  - Treatment waves
- Intervention
  - Technical components of the intervention
  - Differences between treatment and control
    - Differences between distinct treatment arms
    - Definition of cluster and differences between cluster and unit of analysis
    - Flow chart diagramming treatment arms, sample sizes, timelines
- Data collection methods and procedures
  - Description of data collection method
  - Description of any other data sources used and source of data
- Randomization procedure
  - Detailed description of the mechanism for randomizing, including how the process will be safeguarded for tampering
  - Individual vs clustered randomization
  - Stratification variables
- Blinding
  - Describe blinding (who and how)
- Power calculations
  - Justification of effect size used in power calculations

5. Pre-specifying Analytical Decisions

- Type of model and justification for use
- Variables to be constructed—describe details of construction
- Accounting for multiple hypothesis testing
- Any indices, mathematical formulas, explanation, and rationale

6. Expected Issues

- Details of procedures in place to address issues: noncompliance and monitoring

7. Conclusion

III. Evaluation Plan Template

[PARTNER] and [RESEARCH TEAM]

Sample Evaluation Plan Prepared for [PROJECT NAME]

1. Executive Summary

- *What is the point of the study? What are the key study activities? How do they correspond with partner/grantor objectives?*
- *What intellectual gap is the research addressing? How will the research fill this gap?*
- *What data will be generated? What approaches will be evaluated? What outcomes will be examined?*

2. Intervention Design

A. List treatment arms

T0:

T1:

T2:

B. Evaluation Methodology

- *What are you evaluating? What kind of data will be used? What will you measure? What are the key research questions? How do your data points map to research questions?*

Objective	Metrics	Grouped by (If Applicable)	Data source
Decrease loan defaults	*Default rate*	*Prime/sub-prime borrowers*	*Partner admin. records*

C. Study Enrollment & Randomization Strategy

- *What are the projected start and end dates? How will participants be recruited? What is the target total enrollment? (If randomized by group) What is target enrollment by site?*
- *What is the level of randomization? Is the randomization stratified? Who will implement the randomization? When will it be implemented?*
- *Can people opt out or in, and if so, how? How will this be tracked?*
- *[Include diagram explaining enrollment process]*

Evaluation Plan Template Continued

D. Power Calculation

- *What are the key outcomes for the study?*
- *What is the target sample size per treatment arm, given the confidence level and power?*
- *What effect size will you be able to detect?*
- *If applicable, what data was used to make these assumptions?*

E. Data Collection Process

- *Over what time horizon will impact be measured? What activities will happen in that horizon?*
- *Who is responsible for transmitting data? How will you ensure the security of the data?*

3. Challenges, Risks, and Threats

- *Enrollment, sample size, and power*
- *Implementation delays*
- *Study attrition & Randomization integrity*
- *Partner buy-in*

4. Appendices

- *Partner Protocols/Policies*
- *Respondent Flow Diagram (Recruitment, enrollment, consent, etc.)*
- *Anticipated Timeline*
- *Study sites*
- *Marketing Materials*
- *Scripts*
- *Consent Forms*
- *Survey Instruments*

IV. Project Gantt Chart Template

[PARTNER] and [RESEARCH TEAM]

Sample Evaluation Plan Prepared for [PROJECT NAME]

	Year 1												Year 2												
Tasks	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Deliverables**
Field work preparation	■	■	■	■	■	■																			
Developing instruments	■	■																							Drafts of instruments
Survey of potential treatment and control markets		■	■																						Summary plan for field work
Hiring and training				■	■	■																			
Testing instruments and certification methods					■	■																			Summary of preliminary tests of certification methods
Project coordinator trip					■																				
Fielding of survey							■	■	■	■	■	■													Progress report of field supervisors
Data entry and cleaning													■	■	■	■									
Data analysis																	■	■	■	■	■	■			Preliminary report
Writing final report																						■	■	■	Final report

V. Project Log Template

[PARTNER] and [RESEARCH TEAM]
Sample Evaluation Plan Prepared for [PROJECT NAME]

Partnership Development

Project Management

1. Project history
2. Project timeline
3. Research team personnel roles and hierarchy

ROLE	NAME	CONTACT INFORMATION	NOTES
At [RESEARCH ORGANIZATION]			
Principal Investigators			
Research Staff			
At [PARTNER]			
Data transfer			
Project management			
At [FUNDER]			
Grant management			

(*continued on next page*)

Project Log Continued

Project Management Continued

1. Location of key documents

FOLDER NAME	CONTENTS	DROPBOX	SECURE SERVER
1 Admin	IRB approvals, grant agreements, financials		
2 Project management	Call notes, timelines, workplans		
3 Partner	Branch lists		
4 Study design	Evaluation plan, power calculations, study development		
5 Intervention	Staff training manuals		
6 Data Collection	Consent scripts, survey instruments		
7 Raw data	Raw administrative data, data cleaning		
8 Analysis	Do-files, datasets, graphs, analysis tables, codebooks		
9 Writeup	Quarterly reports to funder, final reports to funder, conference presentations		

2. Handover tasks: Status & next steps

Finance

1. Description of the funders; if there is more than one funder, which funders cover which project expenses
2. Reporting requirements
3. Project budgets
4. Project expenses to date

Intervention

1. Partner information
2. Background
3. Implementation
4. Monitoring

Research Design

1. Policy questions
2. Relevant Literature
3. Power calculations
4. Randomization
5. Sample selection
6. Data sources
7. Survey
8. Baseline

Project Log Continued

Human Subjects

1. IRBs
2. Applications, amendments, renewals, termination

Measurement and Questionnaires

1. Instruments
2. Administrative data
3. Survey
4. Pilot
5. Baseline
6. Endline

Data Collection

1. Sources
2. Data Schema
3. Decisions

Data Management

1. Security protocol
2. Data management tips
3. Analysis Plan
4. Analysis

Results and Outreach

1. Reports
2. Project presentations

VI. Project Manual Template

[PARTNER] and [RESEARCH TEAM]

Sample Evaluation Plan Prepared for [PROJECT NAME]

1. **Background**
    - Motivation
    - Local context and relevant indicators
2. **Description of intervention**
    - Research question
    - Research design
    - Target population
    - Randomization unit and method
3. **Data collection and intervention timeline (updated to reflect actual completion dates)**
4. **Structure of research team: surveyors, team leaders, and supervisors, and an evaluation of their performance**
5. **A “map” of your files, both hardcopy and softcopy**
6. **Relevant data files and a codebook to the data**

VII. Project Budget Template

[PARTNER] and [RESEARCH TEAM]
Budget Prepared for [GRANT]
Note: Add additional columns to the right for each month that the RCT is expected to run

Project Name:												
Partner Name(s):												
Project ID:												
Grant ID(s):												
Expense	**Total Estimated**	**Coding**	**Notes**	**Month 1**	**Month 2**	**Month 3**	**Month 4**	**Month 5**	**Month 6**	**Month 7**	**Month 8**	**Month 9**
Survey Expenses												
Printing	$0.00											
Postage	$0.00											
Incentives	$0.00											
Enumerator costs	$0.00											
Data entry	$0.00											
Other (specify)	$0.00											
Travel Expenses												
Airfare	$0.00											
Ground transportation	$0.00											
Vehicle rental	$0.00											
Personal vehicle expenses	$0.00											
Hotel/lodging	$0.00											
Meals or per diem	$0.00											
Travel communications	$0.00											
Meals or gifts for partner	$0.00											
Other (specify)	$0.00											

Project Budget Continued

Expense	Total Estimated	Coding	Notes	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9
Other Data Collection												
Admin data pulls	$0.00											
Credit report pulls	$0.00											
Other (specify)	$0.00											
Other												
Other materials	$0.00											
Software	$0.00											
Computers/peripherals	$0.00											
Total	**$0.00**											

VIII. Codebook Template

[PARTNER] and [RESEARCH TEAM] Codebook for [PROJECT]					
Variable Name	**variable1**	**variable2**	**variable3**	**variable4**	**variable5**
Classification					
Group					
Sub-Group					
Group Type					
Unique identifier					
Description					
Weight					
Weight Variable					
Format Type					
Decimal					
Interval					
Dataset Label					
Imputed?					
Unit of Analysis					
Name in raw data					
Variable label					
Question Information					
Question ID					
Question Text					

Codebook Continued

Variable Name	variable1	variable2	variable3	variable4	variable5
Valid Ranges					
Unit					
Min					
Max					
Key					
Notes					
Invalid Ranges					
Unit					
Min					
Max					
Key					
Notes					
Summary Statistics					
Total Responses					
Mean					
Standard deviation					
Notes					

IX. Data Sharing Plan Template

[PARTNER] and [RESEARCH TEAM]

Sample Evaluation Plan Prepared for [PROJECT NAME]

Introduction

This data sharing plan specifies the types of data that will be collected in the course of the evaluation, along with how and when this information will be gathered. This plan is subject to change as the evaluation evolves.

Data Structure

- What is the primary unique ID for the data?
- How is it assigned to participants?
- Are there secondary unique IDs?
- What type of capacity is required from the partner to maintain these keys and pull the data accordingly?
- What are the principal data types?

Data type # 1 (Repeat for each data type)

- What datasets does this consist of?
- How will this data be collected?
- What are the survey procedures?
- What survey channel will be used?
- How will respondents be contacted?
- How will data be transmitted, and on what time frame?
- If applicable, how will the data be entered, and who will enter it?

	Dataset #1	Dataset #2
Identifier		
Variables		
Sample		
Timing of collection		
Timing of transmission		
Lead		

Data Sharing Plan Continued

Data Collection and Reporting Time Frame
• What is the enrollment time frame? • What is the data transmission time frame, for each data type? • When will reports be produced? • When will balance checks be conducted? • When will the analysis occur, and the corresponding academic paper be drafted? • Who is responsible for coordinating these transmissions on both side?

X. Stakeholder Analysis

Stakeholder analysis can be a useful tool for thinking through who the different people are who will be involved in your project (the stakeholders), identifying their interests and incentives, and using this understanding to both leverage the support of those in favor of the project and manage the risks posed by those who are against it. We recommend doing a stakeholder analysis early in your project, but remembering to revisit this document throughout the life of the project as you learn new information about the context, the stakeholders change, or the project changes. Keeping this document up-to-date will help you keep in mind the relevant people and ensure that you are maintaining their engagement and monitoring for any possible risks.

DFID[1] defines three different types of stakeholders as follows:

(1) Key stakeholders: Those who can significantly influence or are important to the success of an activity

(2) Primary stakeholders: Those individuals or groups who are affected by an activity, either as beneficiaries (positively impacted) or disbeneficiaries (negatively impacted). Typically in an RCT of a financial product, this would include any current or potential clients of the financial institution

(3) Secondary stakeholders: All other individuals or institutions with a stake, interest, or intermediary role in the activity. This would include, for example, loan officers, member services representatives, etc.

DFID recommends conducting stakeholder analysis in the form of a workshop with representatives from different affected groups. In practice, we have not found this to be feasible. However, you can use informal interviews, focus groups, and training sessions to gather the information you need from the different individuals and groups you identify. We have found that it is well worth the time investment to sit down with—at the very least—managers, frontline staff, and a sample of clients at the financial institution (preferably at the pilot stage) to learn what they like and do not like about both the product/service and the evaluation, as it allows us to adapt the design appropriately.

The steps involved in conducting a stakeholder analysis are as follows:

(1) Identify the main stakeholders, listing key, primary, and secondary stakeholders.

(2) Identify their interest in the project. This includes the costs and benefits to the stakeholder. For example, managers could be supporting the RCT because they believe that positive results will allow them to leverage larger amounts of donor funding. Keep in mind that, as this is an evaluation, what we are identifying now are perceived costs and benefits—for example, a sub-prime borrower might believe that greater access to small-dollar credit is beneficial and therefore approve of the roll-out of a new loan program, even though the actual impact of the loan program is still unknown (that’s why we do the RCT!).

[1] “Tools for Development: A Handbook for Those Engaged in Development Activity.” UK Department for International Development (DFID), September 2002. http://commdev.org/tools-development-handbook-those-engaged-development-activity.

(3) Identify the level of influence and importance of each stakeholder. Influence is the power the stakeholder has to facilitate or impede the activity. Importance is the weight given to meeting that stakeholder's needs. For example, clients of a bank may have very little influence over the product design, since they are not in a position to make decisions for the bank. But the research team may accord them large amounts of importance, as they are the intended beneficiaries of the project. Influence and importance can be rated high/low, on a five-point scale, or however it makes sense to your team.

(4) Identify the risks posed by each and discuss mitigation strategies.

Stakeholder Analysis Example

This is not meant to be complete analysis, but rather to give you an idea of how the stakeholder analysis works.

Stakeholder	Interest	+/-	Importance	Influence	Risks
CEO of FI	Thinks that positive results from RCT will improve FI image Positive impact of product on bottom line Worried about staff time costs of RCT	+	High	High	RCT will show negative results and CEO will not want them released
Primary					
Borrowers – treatment	Greater access to credit	+	High	Low	Different underwriting standards may cause confusion and complaints
Borrowers – control	(If find out about treatment) may feel upset at not having access	-	High	Low	Increased complaints
Secondary					
Frontline staff	Feel they can provide a good service to their clients Paid more for originating loans Have to spend time on baseline survey	+	Medium	Medium	Staff don't implement research activities If clients decide they don't like the loans, staff will cease to offer them

Innovations for Poverty Action
poverty-action.org