10 THINGS TO KNOW ABOUT SURVEY DESIGN FOR EXPERIMENTS

The design of baseline and endline surveys differ in key ways 2

Baseline surveys 2

Covariates 2

Pre-treatment measurement of outcomes 3

Endline surveys 3

There are benefits and drawbacks of surveys as measurement tools 4

Develop your survey before or in tandem with your pre-analysis plan 5

Use standard measures in order to make out-of-sample comparisons 6

Behavioral measures are almost always better 6

Gathering behavioral data doesn’t have to be expensive. Here is how to develop low-cost measures: 7

There are survey methods that measure sensitive behaviors and attitudes in risky environments while protecting respondents 7

If social desirability bias and/or risk to respondents are not concerns: then use attitudinal measures with these qualities 8

How do you construct questions that accomplish these goals? 8

Question ordering & instrument length matter 9

Make sure response rates do not differ as a function of treatment assignment 9

Pilot! 10

# The design of baseline and endline surveys differ in key ways

Surveys are the most frequently-used tool for collecting experimental data in social science research, and the design of these surveys can have a profound effect on the conclusions we draw about the treatments we study. Therefore, the stakes are high when designing surveys around your experimental projects.

At a minimum, all that is needed to estimate a treatment effect after an experiment has been conducted is a measure of the outcome, collected after the treatment has been delivered, with a sufficient number of observations across the treatment and control groups. If time and budget allow, baseline surveys, conducted before the implementation of an experiment, serve important functions as well and can improve analysis greatly.

Baseline surveys should produce data that describe the experimental subject population as they were before the treatment is delivered. This is accomplished through the measurement of covariates, which are observed *pre-treatment* characteristics of experimental subjects. Using covariate data you can: 1) describe the subject population, 2) improve the precision with which you estimate treatment effects, 3) report balance, and 4) estimate heterogeneous treatment effects.

Covariates improve the precision with which you can estimate treatment effects by reducing variance in three ways; covariates can be used to rescale your dependent variable, as controls when using regression to estimate treatment effects, and to construct blocks in order to conduct blocked random assignment[[1]](#footnote-1). In order for covariate data to be used to reduce variance in our estimates of treatment effects, they need to be unaffected by treatment assignment, i.e. collected sometime before treatment is delivered. See the guide on [covariates](http://egap.org/methods-guides/10-things-know-about-covariate-adjustment) for more about how to use covariate data.

The greater the predictive power of included covariates, the greater increase in the power of your design and the precision with which you can estimate effects. If you believe covariates will likely predict outcomes in your experiment, then that is grounds to include them in your survey. For example, if the intervention involves providing a service at a cost to treated users, income will likely explain some variation in outcomes and is therefore a useful covariate to measure at the baseline stage.

Because pre-treatment covariates can improve precision, conducting a baseline becomes more important when the sample size is limited.

Covariates also allow you to conduct sub-group analyses. Heterogeneous effects are not causal, and so interpretation is limited. Still, understanding how treatment effects vary by subejcts’ attributes can provide you with important clues about mechanisms. The implication for design is to include covariates for which you would like to report heterogeneous effects.

Covariate data will also be used to show “balance”, or the extent to which the treatment and control populations resemble each other. Although random assignment alone ensures that outcomes across the treatment and control group are *in expectation* the same, it is standard practice to show that random assignment resulted in two groups that are “balanced” on covariates of interest. If, for example, the treatment group included 25% more men than the control group, we might worry that random assignment failed in some way. Collecting pre-treatment covariate data allows us to evaluate and report balance.

### Pre-treatment measurement of outcomes

The baseline provides an opportunity to measure the outcome before the experiment was conducted, later allowing you to use change scores as your outcome and the difference-in-differences estimator. The difference-in-differences estimator will improve precision only when a covariate strongly predicts outcomes[[2]](#footnote-2)

Endline surveys, conducted after the treatment is delivered, are primarily used to measure outcomes. Including questions about implementation can improve analysis and interpretation greatly.

Surveys conducted after treatments are delivered are one way to understand if there were compliance issues or other implementation issues that may have consequences for analysis. Survey data can help to determine the scope of non-compliance with the assigned treatment, and the underlying causes. This is an opportunity to ask subjects directly about reasons for noncompliance. Subjects may have understood the treatment differently than the researchers, and survey data can be used to both show this and speculate why this may have happened.

In the endline, you can learn about spillover by asking subjects in the control condition about their knowledge and access to the treatment. Interviews with treated subjects are useful for understanding spillover as well, because survey data can be used to understand the networks through which the treatment could have “spilled-over” into the control group.

Describing the population is important here as well, but covariate data collected after implementation are less useful for improving precision. Ordinarily, covariates collected after treatment assignment are considered suspect, as they could conceivably be affected by treatment.

|  |  |
| --- | --- |
| **Baseline checklist** | **Endline Checklist** |
| *Will your data allow you to:*   * Describe the population * Adjust treatment effect estimates (are the covariates included likely to be prognostic of outcomes?) * Estimate heterogeneous treatment effects * Design a blocked randomization procedure * Describe balance across treatment and control conditions | *Will your data allow you to:*   * Estimate effects * Assess if spillover occurred, of if there was interference * Measure non-compliance (and the reasons for non-compliance if it occurred) * Look for causal mechanisms (how are effects transmitted?) |

# There are benefits and drawbacks of surveys as measurement tools

Occasionally researchers face a choice between using administrative data or collecting original data in order to measure outcomes and estimate treatment effects. Governments or organizations implementing social programs typically collect their own data, and its possible that this data fits your needs, and will allow you to estimate treatment effects without implementing your own survey.

Existing data can be used in place of original data when:

1. It covers the experimental subject population
2. It includes outcome measures that are consistent with the outcomes of interest specified in the pre-analysis plan
3. It has been collected in a reliable way:
   1. the group that collected the data implemented a known sampling procedure
   2. the incentive was to collect unbiased data, rather than prove the program had a positive effect, i.e. respondents were not intimidated or manipulated

Using administrative data saves money and time, to that end it is important to understand the tradeoffs involved in the collection and use of survey data for the analysis of experiments. In some cases, administrative data is more reliable than original data. For example, court records might be more reliable than interviews with plaintiffs and defendants.

Collecting original data usually means more options. As a researcher you are able to collect tailored measures, hone in on a population of interest, and you have more opportunity to explore mechanisms.

# Develop your survey before or in tandem with your pre-analysis plan

It is now standard practice in experimental social science research to register a **pre-analysis plan** (PAP**),** which lays out in advance exactly what effects will be estimated and how. The key concern is that if you choose your tests *after* you see the data, you can get whatever results you like, or at least you can accentuate the tests that bolster a pet hypothesis.[[3]](#footnote-3) Pre-registering a design and analysis plan, therefore, is a solution that prevents “fishing”: data mining and specification searching.

If you plan on developing a PAP, there are good reasons, beyond the normative value in increasing the level of transparency in your work, to develop your survey(s) at the same time. Early development of survey instruments is an opportunity to increase confidence in your power calculations, and to enhance precision by including measures of all possible predictive covariates.

Power is the overall probability that your design will detect an effect if an effect exists. Calculating [power](http://egap.org/methods-guides/10-things-you-need-know-about-statistical-power) involves making assumptions about the range of possible effect sizes, the range of possible values of predictive covariates, how many subjects you will include in your experiment, and how many subjects will answer your survey. Developing your instrument and sampling procedure can significantly improve the accuracy of these assumptions by setting the range of possible values for covariates and outcome measures.

The best PAPs include a mapping between survey measures (questions) and each equation of interest that will eventually be estimated. Completing this exercise as you develop your survey ensures you have included necessary survey-based covariates and alternative outcome measures[[4]](#footnote-4).

A good PAP will also include how you plan to code open ended questions, how you plan to coarsen data, and how you will deal with missingness (including power calculations under different levels of missingness).

# Use standard measures in order to make out-of-sample comparisons

Review instruments posted on the EGAP pre-registration archive and regional public opinion surveys ([Latinobarómetro](http://www.latinobarometro.org/lat.jsp), [Afrobarometer](http://afrobarometer.org/)); most likely someone has already measured what you want to measure. The critique of experiments often rests on the “portability” of results; we aren’t sure if results would replicate in other contexts. Using common measures is one step towards being able to compare effect sizes and underlying populations across contexts.

# Behavioral measures are almost always better

As social scientists, we ultimately care about attitudes and beliefs because we expect they drive behavior. The mapping of attitudes to behaviors, however, is often less coherent than we might expect. For example, a respondent might be willing to express dissatisfaction with an authoritarian regime in the context of a survey, but entirely unwilling to join a protest. Therefore, making claims about the link between attitudes and behaviors involves some assumption about the costs of behaviors, which are unknown to the researcher. Within the context of a survey, however, there are more opportunities than you might expect to inexpensively and directly measure respondent’s behaviors.

For example, Lauren Young used a clever behavioral measure when estimating the effect of fear on participation in dissent. The key outcome in her experiment is the willingness of participants to publicly signal support for the opposition in an authoritarian regime. We might believe that simply asking respondents if they support the regime would not yield truthful responses: it is relatively low cost to say “no,” or alternatively, the respondent might feel pressure from the researcher to say “yes.” In order to more accurately measure behavioral change, she gave subjects an opportunity to accept a pro-opposition wristband[[5]](#footnote-5). This measure is more convincing because wearing the wristband is public and potentially costly. The behavioral measure is a more accurate indicator of political behavior in the world.

## Gathering behavioral data doesn’t have to be expensive. Here is how to develop low-cost measures:

1. Brainstorm a set of actions that subjects would do if the treatment had had an effect or would not do in the case that the treatment did not have an effect. It also works to think about behavior on a continuum—i.e., what would people be *more likely* to do if affected, and *less likely* to do if not? Local context matters a lot here; rely on your enumerators, local survey staff, or implementation partners to help you think through a set of possibilities. One nice way to think about this is to challenge yourself to think about “hints.” In the above example, we might think about a group of opposition activists wearing wristbands publicly as a “hint” that people are more likely to take risks. Keep in mind your eventual audience: what behaviors are frequently tracked in the literature you hope to speak to?
2. Isolate the set of behaviors that are feasible to measure. This will most likely be the set of behaviors that can be immediately observed by the enumerator and involve minimal materials. What is the least expensive or costly action that would be associated with the behavioral change you want to detect?
3. Ideally, pre-test the measures either with the rest of your survey, or in smaller focus groups. Learning why respondents did or did not behave a certain way will increase confidence in your results.

# There are survey methods that measure sensitive behaviors and attitudes in risky environments while protecting respondents

The estimation of treatment effects depends on accurate measures of attitudes and beliefs, which in turn depend on honest responses to survey questions. Respondents, in some cases, will answer survey questions in the way they believe others in their community would like them to, or in the way the enumerator or researcher would like them to. People are more likely to do this when they could face punishment for their true beliefs—from social sanctioning to physical harm. In some instances, respondents may not feel comfortable giving a response at all. If this is the case, your measures will be biased in the direction respondents believe to be “socially desirable.”

Social desirability bias or even worse, risk of harm, makes it difficult to elicit truthful responses in sensitive questions on a survey. A class of statistical methods for survey methodology address this problem. Each method-- list experiments, endorsement experiments, and the randomized response technique—works by introducing random noise that conceals individual responses. This means that, on the subject side individual responses are hidden from the researcher, and on the researcher side you can only study the responses of groups (not individuals).

In the case that you use one of these methods, you may also want to include direct survey measures of the same concept or directly measure for at least a sample of the subset to be able to capture the level of reporting bias across direct and list/endorsement/randomized response measures.

LINKS TO LIST EXPERIMENT RESOURCES:

http://imai.princeton.edu/projects/sensitive.html

# If social desirability bias and/or risk to respondents are not concerns: then use attitudinal measures with these qualities

High-quality measures share two main characteristics; they are valid and reliable. A valid measure measures the underlying concept it is intended to capture. A reliable measure produces the same measurements when a given quantity is assessed repeatedly.

Experiments should be “replicable” in the sense that other scientists could run the same experiment in the same context and get the same results. Having a survey with valid and reliable measures is essential to making this possible.

## How do you construct questions that accomplish these goals?

1. Use the simplest possible form of each question, using the most widely-understood words. Avoid jargon or technical terms, and be straightforward and brief.
2. Be specific, such that if the question were to be lifted from the section and asked without context you would get the same response.
3. It helps to begin the question by providing a context. For example, you can prime a time period (“Thinking of the last year: has your income been better, the same, or worse?”), or a place (“Thinking of people in this village: have people earned more or less this year as compared to last year?”)
4. Avoid measuring multiple things at once. For example, the following question measures attitudes about both the president and government concurrently, making it difficult to draw a clear conclusion from the data: “Do you think the president and the government are doing a good job in terms of protecting basic freedoms?”
5. When constructing response categories, be as comprehensive as possible. Include all possible responses. You don’t want to record a lot of “don’t know” responses and miss important information. See below for a discussion of scales.
6. Keep in mind the concerns with social desirability discussed above; the wording of the question shouldn’t lead the respondent towards a certain response.

# Question ordering & instrument length matter

Responding to a survey is costly for the respondent. They are volunteering their time and being asked to think in ways that may be new or unfamiliar to them, often for long periods of time. Keeping this in mind, keep your questionnaire as short and as engaging as possible. If you have a large enough sample, you can shorten a questionnaire by randomly assigning some of the respondents to different forms.

Split your instrument into sections by topic. Give your questionnaire a logical flow as you move from one section to another, starting with general questions and moving to more complex topics.

Start your questionnaire with an introduction that makes the respondent feel informed and safe. Ask sensitive questions towards the end of the survey; respondents will (hopefully) feel more comfortable as the interview goes on, and if the respondent ends the interview you still have most of your data. Be aware that priming in earlier sections will affect the responses to later questions.

# Make sure response rates do not differ as a function of treatment assignment

If response rates are related to treatment status, your survey data can yield biased estimates. For example, let’s say you are conducting a study of an intervention that involves testing study participants for health problems and informing those in the treatment group of their diagnosis. This is a sensitive intervention and may cause people in the treatment group to become more apprehensive about answering a survey than those in the control group. Simply analyzing the data without accounting for differential response rates would yield data that was biased away from the true treatment effect—we lack data on those with the highest potential outcomes, those subjects who have been exposed to the strongest version of the treatment.

One way to deal with this in the design of your *survey* (not the design of the treatment or in analysis alone [e.g., using bounded treatment effects[[6]](#footnote-6)]), is to track a subsample of individuals from the hard-to-reach group. Choose a subset of missing respondents and invest in tracking and reaching them. At the analysis stage you have the option of weighting the data from this subsample in order to account for attrition.

# Pilot!

It is absolutely essential to pre-test your survey instrument after you have designed it. The purpose of a pre-test (or pilot) is to make sure your instrument works in the real world. This means that, by the time you implement your survey in the field, you are confident that:

1. Respondents will understand the questions
2. Respondents will be comfortable responding
3. The questions are not leading
4. The questions produce enough variation in the data
5. That you have included all relevant response categories (and you won’t get a lot of non-response)
6. The instrument is not too long

The [survey implementation guide](http://egap.org/content/10-things-know-about-survey-implementation) explains about how to set up and carry out a pre-test.

1. Gerber, Alan S., and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton, 2012. [↑](#footnote-ref-1)
2. Gerber, Alan S., and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton, 2012. See equation 4.2. [↑](#footnote-ref-2)
3. Link to fishing article:

   <http://www.columbia.edu/~mh2245/papers1/PA_2012b.pdf>

   Link to Casey et al:

   <http://qje.oxfordjournals.org/content/127/4/1755.full.pdf+html> [↑](#footnote-ref-3)
4. If employing multiple survey measures for the same outcome; pre-commit to how you will analyze them e.g. in an index/with accounting for multiple comparisons etc. to avoid cherry-picking measures post-hoc. [↑](#footnote-ref-4)
5. Young, Lauren. The psychology of political risk in autocracy. Working paper, Columbia University, September 2015. [↑](#footnote-ref-5)
6. This approach involves estimating the upper and lower “bounds”, which are the largest and smallest ATEs we would obtain if the missing information were filled in with the highest and lowest outcomes that appear in the data we have. [↑](#footnote-ref-6)