Note:

I wrote this guide to be about instrument/questionnaire design and aimed it at a wider audience that includes not only late stage graduate students and professors, but also practitioners and early-stage graduate students. My thinking is that it’s the latter group is most likely to look to a guide like this, versus some of the others on the site that go into more technical topics.

That said, if you want me to scrap some of the more basic sections and include an in-depth treatment of sampling, I can!

10 THINGS TO KNOW ABOUT SURVEY DESIGN FOR EXPERIMENTS

Surveys can be a source of outcome measures or pre-treatment covariates 3

There are benefits and drawbacks of surveys as measurement tools 3

Develop your survey before or in tandem with your pre-analysis plan 4

Collect as much covariate data as possible because predictive covariates = more precise estimates of treatment effects 5

Collecting covariate data 5

Behavioral measures are almost always better 6

Gathering behavioral data doesn’t have to be expensive. Here is how to develop low-cost measures 6

There are survey methods that accurately 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 7

How do you construct questions that accomplish these goals? 8

Use standard measures in order to assuage external validity concerns 8

You can always coarsen data, but you can’t make coarse data more granular 8

Question ordering matters 9

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

# Surveys can be a source of outcome measures or pre-treatment covariates

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.

When it comes to experimental research, surveys can serve two important, yet different, functions. Surveys can provide outcome data used to measure differences across outcomes in the control and treatment groups or they can provide pre-treatment covariate data that can improve the precision of treatment effect estimates.

A baseline survey, conducted before the implementation of an experiment, should include covariates, and, if possible, a pre-treatment measure of the outcome. Covariate data can be used for increasing the precision with which treatment effects are eventually estimated (see below).

An endline survey, conducted after the implementation of an experiment, is primarily used to measure outcomes. Describing the population is important here as well, but covariate data collected after implementation are less useful for improving precision.

# There are benefits and drawbacks of surveys as measurement tools

There are tradeoffs in the collection and use of survey data for the analysis of experiments. The below table details the benefits and drawbacks of each data source, but in general surveys are more flexible and less reliable than pre-existing data. This isn’t always true: there is a wide variation in the quality of pre-existing data, but assuming you only consider high-quality data from other sources, you face the following trade-offs.

|  |  |  |
| --- | --- | --- |
|  | **Pre-existing or adminstrative data** | **Collecting your own survey data** |
| **Benefits** | Accuracy, unbiasedness, unaffected by treatment and/or Hawthorne effects | Create tailored measures, survey measures can also get at mechanisms |
| **Drawbacks** | Not readily available for the exact question you want to answer | Concerns over truthful reporting, developing reliable and valid measures, and sampling |

# 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 (footnote Casey et al. paper here). Pre-registering a design and analysis plan, therefore, is a solution that prevents “fishing”: data mining and specification searching[[1]](#footnote-1).

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 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 [LINK TO POWER GUIDE]. 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[[2]](#footnote-2).

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).

# Collect as much covariate data as possible because predictive covariates = more precise estimates of treatment effects

A covariate is an observed *pre-treatment* characteristic of an experimental subject. 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[[3]](#footnote-3).

The larger 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. Simply stated: the more covariate data you have, the better.

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 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.

## Collecting covariate data

In order for covariate data to be used to reduce variance in our estimates of treatment effects, they needs to be unaffected by treatment assignment. The implication for survey design is that covariate data that could be plausibly affected by treatment needs to be either collected before the experiment begins. This means using a baseline for field experiments, and in the case of survey experiments, placing covariate modules before the embedded experiment.

Not all covariate information needs to be collected from respondents during the course of the interview. Enumerators can observe subject-level covariates (like sex), household level characteristics (like the material of the roof, a common element of lived poverty indices), and village-level characteristics (like the availability of water and electricity).

There are many standard measures of covariates. Using an existing measure increases confidence, external validity, and allows for easy comparisons across studies.

# 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[[4]](#footnote-4). 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 accurately 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. [ADD CITATIONS]

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.
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. Provide reference frames, particularly for open-ended questions. 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 where everyone means the same thing, and you have missed important information. See below for a discussion of scales.
6. Keep in mind the concerns with social desirability discussed above; don’t lead the respondent towards a certain response. Priming a response will bias your data.

# Use standard measures in order to assuage external validity concerns

Review instruments posted on the EGAP pre-registration archive and regional public opinion surveys; 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 carry to other contexts. Using common measures is one step towards being able to compare effect sizes and underlying populations across contexts.

# You can always coarsen data, but you can’t make coarse data more granular

As frequently as possible, replace dichotomous response categories with scales. Using ordinal or interval scales will allow you to estimate effect sizes and compare respondents in a more precise way.

Respondents will have trouble locating their attitudes on a scale with too many response categories and scales without a midpoint. Research has shown that the maximum number of response categories a respondent can easily grasp is 7. Scales with an odd number of response categories will have a mid-point (5 or 7).

Unipolar scales are more easily understood by respondents. For example, it’s better to use a scale that ranges from “extremely responsive” to “not at all responsive,” rather than a scale that ranges from “extremely responsive” to “extremely unresponsive.” That said, you need to have all possible responses included.

Your PAP should lay out how you plan to rescale your data if you plan to dichotomize some variables for analysis.

# Question ordering matters

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.

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 studying an intervention that aimed to raise incomes and was administered at the village level. You might return to survey treated villages and find that economic success has allowed people to move into the city. Simply interviewing a random subsample of village residents would yield data that was biased away from the true treatment effect—those with the highest potential outcomes (the biggest income gains) have dropped out of the sample.

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]), is to track a subsample of individuals from the hard-to-reach group. Choose a subset of missing respondents and invest in tracking them. At the analysis stage you have the option of weighting the data from this subsample in order to account for attrition, or using it to create bounds in a more informed way.

Lin comments:

The focus on outcomes and covariates (and the emphasis on precision as the main reason to collect covariate data) feels a bit too narrow to me. Covariate adjustment often doesn't improve precision that much, but there can be other important uses for survey data, e.g.:

- Describing the population

- Measuring the contrast between the treatment condition and the control condition (e.g., if the control group is denied access to a specific social program but can find similar services elsewhere, we want to know if the treatment group really got a larger dose of services, and by how much). Orley Ashenfelter has a nice little speech briefly making this point ("I think one of the most critical lessons learned from the program evaluation literature is the necessity of first showing that a program exists"):

<http://legacy.iza.org/en/webcontent/prize/history/prize2003/Ashenfelter.pdf>

The data collected to measure the treatment-control contrast may also be used to develop estimates of the net cost of the treatment (e.g., see Larry Orr's book Social Experiments, pp. 180-182).

- Related to the above is measuring to what extent people understood the treatment condition. For example, David Card, Phil Robins, and I used survey data for a section called "Do people understand the treatment?" on pp. 11-15 of this 1997 working paper:

<http://www.srdc.org/uploads/how_Important_entry_effects.pdf>

- Surveys may be useful for asking about reasons for noncompliance with the assigned treatment (though of course people's stated reasons aren't necessarily the only reasons). My coauthors and I wrote a 1998 report on an experiment where the treatment group was offered a huge earnings supplement if they left welfare for full-time work. Only one-third of the treatment group took up the supplement. Ch. 2 of the report uses survey data to speculate about why.

<http://www.srdc.org/uploads/when_fin_inc_encourage_work.pdf>

Some of these (and other) uses of surveys are mentioned in Glennerster & Takavarasha's book Running Randomized Evaluations (ch. 5) and Orr's book Social Experiments (ch. 5).

The table mentions accuracy and unbiasedness as benefits of pre-existing or admin data, but that isn't always the case. Government admin data on employment and earnings (e.g., from Unemployment Insurance, Social Security, or the IRS) may systematically understate informal employment, which may be better captured on surveys. Bob Kornfeld & Howard Bloom have a 1999 paper comparing survey and admin data in the JTPA experiment:

<http://www.jstor.org/stable/10.1086/209917>

In the section mentioning pre-analysis plans, I'd also recommend linking to Ben Olken's article:

<http://doi.org/10.1257/jep.29.3.61>

Finally, the Development Impact blog has a "curated list" of posts on measurement and survey design. I haven't read most of these, but here's the link:

<http://blogs.worldbank.org/impactevaluations/curated-list-our-postings-measurement-and-survey-design>

1. 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-1)
2. 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-2)
3. Gerber, Alan S., and Donald P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton, 2012. [↑](#footnote-ref-3)
4. Young, Lauren. The psychology of political risk in autocracy. *Working paper GIVE UNIVERSITY OR THINK TANK,* September 2015. [↑](#footnote-ref-4)