On Survey Experiments

Timothy R. McDade
trmcdade@gmail.com

Department of Political Science
Duke University

June 8, 2020

1 Describe the key types of survey experiments used to estimate *causal* parameters.

Survey experiments are experimental interventions incorporated into survey methodology where the treatment often takes the form of varying a survey question between treatment groups. To establish causation, any study must show co-variation between the independent and dependent variables, temporal precedence of the independent variable, and no causal explanation of the treatment effect via a third variable. Methods such as regression and time-series analysis can establish the first two. Survey experiments accomplish the third by randomly assigning participants to an intervention, at its simplest to a treatment or control group, allowing researchers to compare groups that are identical in all variables except the treatment and thereby estimate a causal relationship between the treatment and the outcome.

Sampling procedure is important to the validity of causal conclusions drawn from a survey experiment. In population-based survey experiments, random sampling of the overall population of interest creates a sample that is like the population in all variables, minimizing sample selection error. Alternatively, survey experiments can also be based on convenience samples, which are non-random samples often governed by data availability. Such studies can suffer from nonrepresentativeness of covariates compared to the population of interest, but researchers can use weighting and other statistical procedures to approximate the treatment effect of the overall population. The appropriateness of each sampling method depends on the scope of the research and whether the researchers want to estimate the Sample Average Treatment Effect (SATE), Population Average Treatment Effect (PATE), or Super-Population Average Treatment Effect (SPATE).

Survey experiments come in several types. Measurement experiments vary the wording of elements of the survey (e.g. questions or instructions) to elicit information about the relationship between the treatment and the outcome. List experiments attempt to circumvent social desirability bias by indirectly measuring participant endorsement of the socially undesirable item as the difference in mean affirmative number of responses to a question whose possible responses are varied by treatment and control groups. Conjoint experiments analyze how people make decisions in multidimensional spaces and vary elements of the survey theorized to affect the respondent's answer. Factorial experiments test for the effects of the interactions between two treatment variables by assessing treatment-outcome effect in each combination of the variables. Vignette experiments ask questions about the respondent's reaction to a short prompt (usually text, but other forms are possible, such as resumes), which varies randomly across factors of interest. In each type, the type of treatment varies but random selection and random treatment assignment remain to estimate causal effects. Survey experiments can be considered a type of field experiment if the experiment's scope is a non-laboratory setting that seeks to replicate respondents' actions in the real world (Gerber and Green (2008)).

2 What specific problems are they designed to solve?

Survey experiments seek to simultaneously achieve some of the internal validity of laboratory experiments and the external validity of field, observational, and survey studies. Laboratory experiments provide the control required for random treatment assignment and its associated internal validity, but lack external validity because of concerns that participant behavior could differ in the laboratory from a more natural environment. Survey experiments retain this control and internal validity but improve external validity through attributes of the selection process. Random selection of participants from the population of interest instead of from subject pools, measuring phenomena at the same level as their occurrence (Tingley (2014)), and reaching respondents in a setting more similar to the normal setting of the simulated interaction all increase confidence that experimental results accurately portray what happens in the real world. These strengths make possible more complex experimental designs (Mutz (2011)).

In addition to better internal validity, survey experiments can improve upon field experiments' difficulty probing the mechanism underlying subject behavior by asking why subjects respond how they do. Evaluation of participant reactions immediately post-treatment contributes to construct validity by closely controlling the likelihood that the experiment measures what it intends to. When used with representative samples, survey experiments can provide firmly grounded inferences (Krupnikov and Findley (2016)). They can be cheaper and more easily replicable than other types of field experiments (Gerber and Green (2008)).

3 What are practical challenges inherent in their use?

Survey experiments face several practical challenges, the most important of which is sampling procedure. To best identify the causal effect, the researcher should carefully recruit a sample that is representative of the population of interest. Actually doing so is costly and usually outsourced to survey companies. These companies can pull the sample from standing subject panels, which can be problematic: survey panels can contain repeated participants, who can exhibit different behavior than naive participants. This introduces the possibility for unobserved confounders like responsiveness as a function of past survey experience.

Measurement is another major challenge to survey experiments. The scope and unit of analysis can pose challenges to construct validity: what is the definition of a participant (household, individual, etc.), and how does that participant interact with the treatment? Survey experiment treatments are limited to tasks that a participant can reasonably complete within a survey environment. Any variance in framing or implementation of the survey can introduce bias. The self-reporting inherent in surveys is also susceptible to bias (e.g. social desirability bias). Survey experiments are often designed to establish some pattern in participant preferences about a topic, but people don't always act in the real world how they their survey responses indicate they would. This introduces the risk that the researcher is actually measuring what respondents *tell a researcher that they would do*. As a result, their interpretation is limited by their success in approximating real-world behavior (Tingley (2014)). Survey experiments are unable to generalize beyond the population of interest or the time of enumeration, and thus can only illuminate macro models that are micro-founded or focus on micro-level causal mechanisms (Jensen, Mukherjee and Bernhard (2014)).

4 Describe the inferential challenges involved in their use.

According to Imai, King and Stuart (2008), the major inferential challenges in the use of survey experiments derive from the sampling procedure, which is difficult to make truly random. Survey experiments usually randomly select a large number of respondents from a known population of interest. Random selection reduces sample selection error Δ_S, which refers to differences between study participants and the general population of interest, towards zero. The authors decompose Δ_S into Δ_{S_X}, error due to observable characteristics X, and Δ_{S_U}, error due to unobservable characteristics U.[1] A larger size study should decrease Δ_S because its variance decreases with study size: $lim_{n\to\infty}V[\Delta_S]=0$. Randomly sampling units from a population, as in survey experiments, reduces sample selection error on average across experiments ($E[\Delta_{S_X}]=E[\Delta_{S_U}]=0$) but not necessarily in any one experiment. The possibility that $\Delta_S\neq 0$ remains unless the sample is changed to a census of the population or the quantity of interest is changed from PATE to SATE.

For valid inference, the distribution of observables in the treatment sample needs to match that of the control sample; the same with unobservables is impossible because we cannot observe them. To achieve this, survey experiments randomly assign participants to treatment and control groups. This random assignment reduces towards zero the treatment imbalance Δ_T, defined as the difference between the SATE and the difference in means of the observed outcome variable between the treated and control groups. The authors additively decompose Δ_T into its parts due to observed Δ_{T_X} and unobserved Δ_{T_U} variables.[2] Random treatment assignment reduces the components of estimation error arising from observed and unobserved variables on average, but not exactly in one sample, i.e. $E[\Delta_{T_X}]=E[\Delta_{T_U}]=0$ but $\Delta_{T_X}\approx 0$ and $\Delta_{T_U}\approx 0$.

Although blocking can contribute greatly to reducing treatment imbalance, survey experiments usually have limited or no blocking because researchers rarely know observable characteristics of the participants before the treatment is assigned. If any characteristics of the participants are known before treatment assignment, the researcher can block for them and randomize assignment within blocks. Blocking for *any* variables in estimation of the SPATE (or of the PATE, with large n) reduces Δ_{T_X} towards zero, with more complete blocking corresponding to lower Δ_{T_X}, but blocking only reduces Δ_{T_U} if $u\in U$ is correlated with a blocked observable $x\in X$.[3] In survey experiments, complete blocking is unlikely unless researchers know all observable respondent characteristics before the treatment is assigned; therefore, usually $\Delta_{T_U}\neq 0$ and $\Delta_{T_X}\neq 0$. Imai, King and Stuart conclude that because it will decrease the estimation variance of the causal effect even further than randomization alone, ”blocking is almost always preferable when feasible.”

5 Pick a classic survey experimental paper and reflect on it in light of King et al.

Hainmueller and Hiscox (2010) use a survey experiment to estimate the Average Treatment Effect of skill level of natives on the native's approval (disapproval) of immigrants of the same (differing) skill level. The authors expect, in line with the labor market competition model, that natives with a certain skill level will disapprove of immigrants of the same skill level and approve of immigrants with a differing skill level out of rational self-interest in decreased competition for employment. The authors also expect that, in line with the fiscal burden model, rich natives in states with a high fiscal burden will oppose low-skill immigration more than poor natives because they pay a higher percentage of the taxes funding welfare programs that low-skill immigrants might be likely to receive, and that this gap should be larger in states with higher immigrant

[1] The authors assume no interactive effects between observed and unobserved variables.

[2] To check balance, the authors recommend comparing the values of a statistic that is a characteristic of the sample and whose value the sample size does not affect, such as the difference in means of each covariate, across the treatment and control groups.

[3] One potential downside of blocking is fewer degrees of freedom and lower statistical power in small samples.

access to public services. The authors find disconfirmatory evidence of both sets of hypotheses.

The first weakness of this study is the methodology used to select the random sample from the population of interest. In a randomly sampled experiment, sample selection error should also equal zero in expectation but not necessarily in any one sample. The relatively large sample size of the study helps decrease the variance of the selection errors: $lim_{n \to \infty} V[\Delta_{S_X}] = 0$ and $lim_{n \to \infty} V[\Delta_{S_U}] = 0$. The panel from which the stratified random sample was selected was itself collected via random digit dialing (RDD) and is intended to be representative of the United States over-18 population.[4] But it might not approximate well the population of interest, native citizens. The replication data does not include columns for nativity or citizenship, and the supplementary documentation does not mention screening survey respondents by either variable.

RDD aims for random panel membership among households that have telephones with listed and unlisted numbers, but the survey invitation response rate of 24.6% might obscure variables contributing to panelist acceptance of the survey invitation. For example, RDD could under-represent people with telephones but no internet access, and the sample could under-represent respondents without time to complete an online survey (e.g. those who work multiple jobs). With the proliferation of mobile phones and decrease in land lines during the period of panel recruitment (which began in 1999), area codes could be an imprecise indicator of respondent location or cell phone respondents could exhibit systematic non-response bias. To eliminate this concern, the authors could collect the respondent's taxpaying location from the survey company.

Weighting to correct for these possible shortcomings decreases bias due to known variables but not unobservable variables such as the cell phone location concern, except inasmuch as the unknown variables are related to the known weights. This weakness could decrease confidence in the assessment of the fiscal burden model and the effectiveness of geographical weighting, both of which rely on knowing participant location. One further concern arises from the author-provided supplemental material, which indicates that panel members are expected to complete an average of four surveys a month. The resulting possibility that panelists could exhibit different behavioral patterns than naive members of society potentially introduces respondent survey experience as an unobserved confounder.

The second weakness of the study is the ambiguity of estimated effect: the researchers don't explicitly state if they're estimating the SATE, PATE or SPATE. The scope of the research question implies the authors intend to estimate the SPATE. Random sampling from a panel that is randomly sampled from the overall population of interest suggests likewise (with the above caveats about identifying ”native citizens”). Since treatment is randomized and the noncompletion rate is less than 1% for the items used in the analysis, the analyzed effect is likely an ATE, not ATT. The only other clue to this is the authors' mention of a two-stage probability design, which implies that they are identifying the SATE as an approximation of the SPATE. However, this ATE is conditional on education: therefore, the estimated effect is the SPATE conditioned on education. One major limit on the interpretation of this experiment is its inability to identify this effect over time: the researchers only measure a snapshot of variation across participants, obscuring any temporal variations in the effect due to macroeconomic circumstances, societal unrest, elections, or other macro variables.

Concerns about treatment imbalance should be few. The authors stratified by respondent education level and randomized treatment allocation within each stratum. This random assignment should result in treatment imbalances equaling zero in expectation and approximating zero in any particular sample, and incomplete blocking on education helps reduce Δ_{T_X} even further, but not as much as if blocking were done on more variables. Because the sample size is large, the variance of Δ_{T_U} and Δ_{T_X} is small, increasing the precision of

[4] I reached out to the authors and they provided replication data and details on the sampling procedures.

estimates. The authors' report of extensive balance checks that confirmed balance, made possible because respondent characteristics were known beforehand to the survey firm, further mitigates concerns about treatment imbalance. Although the authors do not report results for an experiment with a pure control treatment (i.e. approval of 'immigrants'), footnotes 17 and 18 describe substantively identical results achieved in the same experiment conducted with a pure control group.

This paper falls prey to neither warning in Pepinsky (2014) about survey experimental research in IPE. Its focus on methodological individualism is appropriate because it tests a theory about individual preferences. It empirically tests assumptions about individual behavior made by the labor competition and fiscal burden models, but makes no claims about establishing a micro-founded theory.

The design of this survey experiment approximates but likely does not achieve zero estimation error. A large sample size helps limit the variance of estimation error and makes low values likely. Random selection from the population of interest limits sample selection error, but threats to inference remain through concerns that sample selection error arises in this particular experiment (not in expectation) from sample misrepresentativeness of the population of interest. Random treatment assignment and blocking both reduce treatment imbalance in known and unknown variables to close to zero. This experiment provides plausibly reliable causal inference and allows confidence in its conclusions.

References

Gerber, Alan S and Donald P Green. 2008. Field experiments and natural experiments. In The Oxford handbook of political science.

Hainmueller, Jens and Michael J Hiscox. 2010. "Attitudes toward highly skilled and low-skilled immigration: Evidence from a survey experiment." American political science review 104(1):61–84.

Imai, Kosuke, Gary King and Elizabeth A Stuart. 2008. "Misunderstandings between experimentalists and observationalists about causal inference." Journal of the royal statistical society: series A (statistics in society) 171(2):481–502.

Jensen, Nathan M, Bumba Mukherjee and William T Bernhard. 2014. "Introduction: survey and experimental research in international political economy." International Interactions 40(3):287–304.

Krupnikov, Y and B Findley. 2016. "Survey Experiments: Managing the Methodological Costs and Benefits." Rae Atkeson, L. and Alvarez, RM (edt.): The Oxford Handbook of Polling and Survey Methods. Oxford University Press .

Mutz, Diana C. 2011. Population-based survey experiments. Princeton University Press.

Pepinsky, Thomas B. 2014. "Surveys, experiments, and the landscape of international political economy." International Interactions 40(3):431–442.

Tingley, Dustin. 2014. "Survey research in international political economy: Motivations, designs, methods." International Interactions 40(3):443–451.