CHAPTER ONE

Population-Based Survey Experiments

A HYBRID METHODOLOGY FOR THE SOCIAL SCIENCES

Approaches to scientific knowledge are a bit like rabid sports rivals; often they cannot say anything nice about their own team without simultaneously disparaging the other side. At some level, they know these intense rivalries would not exist if the other team were not a worthy contender, but the positive aspects of the other side are seldom acknowledged.

Likewise, empirical social scientists tend to develop expertise either in large-scale observational methods such as survey research, or in laboratory-based experimental approaches. They then spend the rest of their careers defending their choice of that particular approach in virtually everything they publish. Each time we submit a journal article, we rehearse all of the advantages of our own methodological choice, briefly mention its weaknesses, and make the case in no uncertain terms that what we have spent our time on is worthwhile. Go team! Competition among methodological approaches is certainly implied, even if it is not explicitly stated. We do our best to defend our own ingroup by stressing the importance of internal validity if we have produced an experiment, or external validity if we have completed an observational study.

Fortunately, this caricature is gradually becoming less accurate, both in terms of its characterization of researchers—an increasing number of whom are trained in multiple methods—and in terms of how methodologists are focusing their attention. Although there are still many survey researchers working on improving their particular method, and many experimentalists focused on developing innovative experimental techniques, there are also methodologists paying specific attention to the problem of integrating results from experimental and observational studies. For the most part, these approaches involve applying complex statistical models to estimates of convenience sample-based experimental treatment effects in order to estimate what they might be in the population as a whole.[1] The goal of population-based experiments is to address this problem through research design rather than analyses, combining the best aspects of both approaches, capitalizing on their strengths and eliminating many of their

[1] See, for example, Sekhon (2009).

weaknesses. The purpose of this volume is to introduce scholars and students in the social sciences to the possibilities of this approach.

Defined in the most rudimentary terms, a population-based survey experiment is an experiment that is administered to a representative population sample. Another common term for this approach is simply "survey-experiment," but this abbreviated form can be misleading because it is not always clear what the term "survey" is meant to convey. The use of survey methods does not distinguish this approach from other combinations of survey and experimental methods. After all, many experiments already involve survey methods at least in administering pre-test and post-test questionnaires, but that is not what is meant here. Population-based survey experiments are not defined by their use of survey interview techniques—whether written or oral—nor by their location in a setting other than a laboratory. Instead, a population-based experiment[2] uses survey *sampling* methods to produce a collection of experimental subjects that is representative of the target population of interest for a particular theory, whether that population is a country, a state, an ethnic group, or some other subgroup. The population represented by the sample should be representative of the population to which the researcher intends to extend his or her findings.

In population-based survey experiments, experimental subjects are randomly assigned to conditions by the researcher, and treatments are administered as in any other experiment. But the participants are not generally required to show up in a laboratory in order to participate. Theoretically I suppose they could,[3] but population-based experiments are infinitely more practical when the representative samples are not required to show up in a single location.

To clarify further, for purposes of this volume, when I use the term "experiment" in the context of population-based survey experiments, I am referring to studies in which the researcher controls the random assignment of participants to variations of the independent variable in order to observe their effects on a dependent variable. Importantly, the term "experiment" is often used far more broadly than this particular definition. For example, many classic "experiments" such as Galileo's observation of gravitational acceleration do not involve random assignment to conditions. And in the social sciences, Milgram's famous demonstration of obedience to authority initially lacked any second group or source of comparison, although he later added these to his design.

[2] I use the terms "population-based survey experiment" and "population-based experiment" interchangeably throughout.

[3] The efforts closest to attempting this are studies of deliberative democracy that bring random samples of citizens together for face to face discussion (see Warren and Pearse 2008; Ackerman and Fishkin, 2004).

So while there are many important experiments that do not meet this definition, I exclude these types of studies from my definition of population-based survey experiments for two reasons. First, in order to be able to make clear statements about the contribution of population-based experiments to internal and external validity, I must limit discussion to experiments for which these two ends are indeed primary goals. Establishing causality and generalizing to a defined target population are not always the goals of research, but they are central to the majority of social scientific work. In addition, the type of experimentation I circumscribe is where population-based survey experiments have the most to offer. Other kinds of experimental studies undoubtedly could benefit from more diverse subject populations as well, but given that experiments that fall outside of this definition are focused on other purposes, this methodological development is less important to these types of studies. However, when scholars want to be certain that a given relationship involves cause and effect, and that their theory may be generalized beyond a narrow pool of subjects, then this is precisely the context in which population-based survey experiments can make their biggest contribution.

Strictly speaking, population-based survey experiments are more experiment than survey. By design, population-based experiments are experimental studies drawing on the power of random assignment to establish unbiased causal inferences. They are also administered to randomly selected, representative samples of the target population of interest, just as a survey would be. However, population-based experiments need not (and often have not) relied on nationally representative population samples. The population of interest might be members of a particular ethnic group, parents of children under the age of 18, those who watch television news, or some other group, but the key is that convenience samples are abandoned in favor of samples representing the target population of interest.

The advantage of population-based survey experiments is that theories can be tested on samples that are representative of the populations to which they are said to apply. The downside of this trade-off is that most researchers have little experience in administering experimental treatments outside of a laboratory setting, so new techniques and considerations come into play, as described at length in this volume.

Why Now?

In one sense, population-based survey experiments are not new at all; simplified versions of them have been around at least since the early years of survey research in the United States. However, technological developments in survey research, combined with the development of innovative

techniques in experimental design, have made highly complex and methodologically sophisticated population-based experiments increasingly accessible to social scientists across many disciplines. Unfortunately, aside from a few journal articles that have been contributed by early adopters of this technique,[4] there has been no book to date addressing this topic in a comprehensive and accessible fashion.

Population-based experiments are neither fish nor fowl. As a result, the guidelines available in textbooks for each of these individual methods—for example, the considerations related to internal and external validity, the design advice, and so forth—do not address concerns specific to population-based experiments. The purpose of this volume is to fill this niche, and thus to encourage wider and more informed use of this technique across the social sciences.

Why is the population-based experimental approach emerging as a distinct methodological option only now? Two technological innovations have brought about the emergence of this method. The first was the development of technology for computer-assisted telephone interviewing (CATI). Until the development of CATI, there were rigid constraints on experimental designs executed in the context of large population samples. The classic "split-ballot" experiment allowed for variation of a single facet, whereas current technologies allow for multiple variations of multiple factors. It has become unnecessary to produce many different versions of a paper questionnaire because the software simply does this for you, with the appropriate variation of the experimental stimulus automatically popping up on the interviewer's computer screen. This advance has allowed researchers to execute extremely complex experimental designs on large and diverse subject pools via telephone surveys.

In addition, the development of the Internet has further expanded the possibilities for population-based experiments. Although Internet-based interviewing of representative population samples is still in its infancy at this point, it is already possible to provide pictorial stimuli as well as video footage to random samples of respondents. The ability to exploit such dynamic data collection instruments has expanded the methodological repertoire and the inferential range of social scientists in many fields. Although population-based survey experiments were done by telephone or face to face long before Internet-based interviewing emerged, the Internet has greatly increased their potential.

The many advances in interviewing technology present social science with the potential to introduce some of its most important hypotheses to virtual laboratories scattered nationwide. Whether they are evaluating theoretical hypotheses, examining the robustness of laboratory findings, or

[4] See Piazza, Sniderman, and Tetlock (1989).

testing empirical hypotheses of other varieties, scientists' abilities to experiment on large and diverse subject pools now enable them to address important social and behavioral phenomena with greater effectiveness and efficiency.

Who Uses Population-Based Experiments?

Population-based experiments can and have been used by social scientists in sociology, political science, psychology, economics, cognitive science, law, public health, communication, and public policy, to name just a few of the major fields that find this approach appealing. But the list does not end there. Population-based experiments have been utilized in more than twenty disciplines including psychiatry, anthropology, business, demography, African American studies, medicine, computer science, Middle Eastern studies, education, history, and even aviation studies. So long as the perceptions, behaviors, or attitudes of human beings are of interest, and the researcher's goal is to test a causal proposition of some kind, population-based survey experiments are likely to be valuable. But they are particularly so when the study is one that would benefit from combining the internal validity of experiments with the external validity of representative population samples.

My personal interest in population-based experiments stems in part from my experiences as an avid user of this method in my own research. In graduate school I was nominally trained in both survey and experimental methods, but these were conceived of as alternative rather than synthesizable approaches. The extent to which experiments were integrated with survey work was limited to tests of alternative question wording, the kind of study that was focused on minor methodological advances rather than substantively focused survey or experimental research. Given that I was not particularly interested in survey measurement issues, this did not seem like an exciting approach to me at the time. But just a few years later, I became aware of the potential this method offered for examining substantive research hypotheses and began incorporating it regularly into my own research.

Beginning in 2001, Arthur (Skip) Lupia and I served as the original principal investigators involved in Time-sharing Experiments for the Social Sciences (TESS), a large-scale infrastructure project supported by the National Science Foundation which had as its mission to promote methodological innovation through the use of population-based survey experiments. Our inspiration for this program came from its intellectual forerunner, The Multi-Investigator Study, which was spearheaded by Paul Sniderman of Stanford University. Paul originally gathered a group of

scholars within the field of political science to share time on a single telephone survey. Each team of investigators was allotted a small amount of time on the survey, and all shared the core demographic questions. The theme that tied these studies together was methodological rather than substantive. Individually, the studies would make contributions to their respective fields and subfields. But collectively, by all using experimental designs, they would demonstrate novel ways to establish causality within the context of diverse population samples.

Skip Lupia and I were fortunate to be two of the young scholars who were invited to put experiments on the Multi-Investigator Study. This platform gave us an opportunity to test our hypotheses in a new experimental context and advanced our research agendas substantially. This relatively simple, but powerful idea demonstrated the tremendous benefits of combining separately conceived and jointly implemented original studies. There were efficiencies of both time and money in this combined effort that meant that more researchers could engage in original data collection. TESS took this idea a step further by establishing an ongoing cross-disciplinary platform for research employing population-based survey experiments.

Our desire to provide this opportunity to social science writ large was the origin of the plan for Time-sharing Experiments for the Social Sciences. Population-based survey experiments could be found here and there across the social sciences even before its inception in 2001, but with TESS, we took the spirit and success of the Multi-Investigator Studies and extended them to serve a greater number of researchers across a larger number of disciplines on an ongoing basis.

The advantages of TESS are straightforward from the perspective of users: it requires a minimum investment of investigators' time to propose a study, provides a quick turnaround time, and is free of charge as a result of generous support from the National Science Foundation. Under these circumstances, few social scientists find reason to complain. In addition, there are broader benefits that accrue from the population-based experimental approach, which I outline in greater length later in this book.

As of 2009, TESS is under the able leadership of psychologist Penny Visser of the University of Chicago, and sociologist Jeremy Freese of Northwestern University. It continues to offer graduate students and faculty from all over the world opportunities to run population-based experiments free of charge. Based on a simple streamlined online application process, proposals are reviewed within their respective disciplines, and once accepted they are placed on a data collection platform for execution on the population of interest. For details, interested researchers should visit the website, ExperimentCentral.org, where the short application (maximum of five double-spaced pages!) and review process are explained.

Indeed, the bulk of the examples I draw on in this book come from TESS-sponsored studies. By adding greater flexibility to the instruments, and establishing a streamlined review process for proposals, we were able to serve an even greater number of scholars at a lower cost per experiment. Further, by expanding TESS outside of political science, we drew on the creativity and ingenuity of a much larger pool of scholars and a much broader range of research subjects. It is this insight that added so much to our own ideas about the breadth of potential applications for population-based experiments.

In its original incarnation, TESS ran many telephone-administered population-based survey experiments, studies in which researchers administered experimental treatments aurally. However, in more recent years, Internet-based experiments have become increasingly popular for a number of reasons. Most importantly, it became possible to acquire representative samples of the U.S. population through a company that recruited via random digit dialing, but then put equipment into the homes of selected participants who were not already online, thus facilitating random probability samples that were accessible via Internet.[5] Demographic and other background information was collected in advance, thus making the required interview time per study quite short.[6]

In addition, Internet-based interviews open up possibilities for graphics, photographs, and video as part of the experimental treatments. Scholars interested in utilizing less obtrusive treatments, particularly when studying sensitive topics, found this advantage particularly desirable. Moreover, because online survey research companies have ongoing relationships with their panel participants rather than one-time encounters, it was also possible to provide monetary incentives that respondents knew they would, in fact, receive. Both telephone and Internet platforms allow the computer to assist in the interviewing procedure so that even highly complex experimental designs can be administered seamlessly.

[5] The survey company originally known as Intersurvey of Menlo Park, CA, but now doing business as Knowledge Networks, was the first and, to my knowledge, remains the only survey firm offering this possibility. In addition, other companies have used elaborate matching techniques to create demographically representative opt-in samples of people who were already online.

[6] TESS relied on three different survey organizations while Lupia and I served as co-PIs. For telephone surveys, the Indiana University Center for Survey Research collected national survey data for a large number of TESS participants. Knowledge Networks of Menlo Park, CA, provided access to their Internet panel sample solicited initially through random digit dialing, and then interviewed regularly via Internet using personal computers or WebTV equipment (put in the homes of those without Internet access). YouGovPolimetrix of Palo Alto, CA, also executed a few TESS studies using novel instrumentation combined with a matched sample of opt-in participants. For a discussion of the quality of these kinds of samples, see Chang and Krosnick (2009) and Yeager et al. (2009).

Drawing on the Advantages of Both Experiments and Surveys

Although most social scientists recognize the tremendous benefits of experimentation, the traditional laboratory context is not suitable for all important research questions, and experiments have always been more popular in some social science fields than in others. To a large extent, the emphasis on experimental versus survey methods reflects a field's emphasis on internal versus external validity, with fields such as psychology more oriented toward the former, and fields such as political science and sociology more oriented toward the latter.

Regardless of field or the emphasis of one's methodological training to date, population-based survey experiments challenge us to expand our methodological repertoire, and to reconsider the "truisms" about more traditional methods as well. For some researchers, survey methods are their primary means of data collection. There are, however, often substantial obstacles to drawing strong causal inferences from conventional survey data. Over the years, many have hoped that advances in statistical methods would allow scholars to use survey data to control for all plausible rival interpretations of a potentially causal relationship. But despite massive and impressive advances in statistical methods over the years, few people are as optimistic today that statistics can solve all of our causal inference problems. For survey researchers, population-based experiments provide a means of establishing causality that is unmatched by any large-scale survey data collection effort, no matter how extensive.

Experimentalists come to population-based experiments with a different monkey on their backs. Having solved the problem of causation in many areas of research by relying primarily, if not exclusively, on experiments, fields like psychology are commonly accused of ignoring external validity. Can we really just assume that the way that college sophomores work is the way all people work? Psychologists have made some good arguments for the generalizability of findings in areas such as basic human perceptual processes. But they are less convincing when it comes to many other areas of knowledge where generational differences or life experiences come into play.[7]

[7] There are some cases in which effects can be effectively extrapolated from a convenience sample to the general population. For example, as Druckman and Kam (forthcoming) note, when the treatment effect is entirely homogeneous across people, extrapolation is obviously valid. But this is a very big "if" because any dimension along which the subject population differs from the general population could potentially invalidate extrapolation. Moreover, establishing homogeneity of effects is not something we can do easily or with any certainty.

Druckman and Kam also suggest some ways to tease out the average effect in the general population from a convenience sample in cases where the treatment effect is not homogeneous. The problem with extrapolation in this case is that one must assume 1) that all relevant moderating variables are known, that is, all possible sources of heterogeneity of effects; 2) that the general population averages for all variables that moderate the impact of treatment are also known; and 3) that measurement of the moderating variable is error-free.

In an ideal world, researchers would not be identified by method, and we would all be well-versed in a variety of approaches. But given that we are clearly not there yet, what is new and challenging about population-based experiments will vary for researchers from different fields. For this reason, I will risk redundancy at times in reviewing some basics so as not to assume too much from any particular reader. My goal in this book is not to provide a resource for the basics of experimental design, nor to discuss the fine points of survey research methods, mainly because there are much better sources on both of these topics. Instead, my goal is to stimulate the use of this approach by providing a sense of its potential and by situating population-based survey experiments in relation to experimental research and survey research. In order to accomplish these goals, it is sometimes necessary to review some basics, and I ask for the reader's forbearance in those cases.

Throughout the natural and social sciences, researchers employ experimental designs in order to combat the challenges posed by the fundamental problem of causal inference. To review the problem in a nutshell, in order for one variable to be said to "cause" another, three conditions generally must be met, the "holy trinity" of causality: (1) the two must co-vary, whether over time or across units of analysis; (2) the cause must precede the effect in time; and (3) the relationship between the cause and effect must not be explainable through some other third variable, which would render the association spurious. In practice, few scholars have problems establishing the first criterion, and the second is problematic only for studies based on cross-sectional observations in which a plausible causal argument can be made for reverse causation.

Thus the "third variable problem" is the key reason experiments are known as the gold standard for inferring causality. Experiments are the best possible way to address the problem of third variables and potentially spurious relationships. This "holy trinity" of causation is well known across the social sciences, but the third variable problem has distinguished itself because of the lack of solid approaches to resolving it. For temporal precedence, researchers can use time series designs, but there is no parallel solution to the third variable problem. In observational research, omitted variable bias plagues or at least menacingly threatens most causal arguments, and there is no simple solution short of an experimental design.[8]

[8] Over-time panel designs come the closest, but in addition to possible problems from attrition and conditioning, panels are expensive and thus few and far between.

Across the social sciences, experimental design strategies entail one of two approaches to the problem of causal inference. Researchers either (1) evaluate a unit of analysis before and after a given treatment relative to those evaluated before and after without treatment, and then draw inferences from these pre-post within-subject comparisons; or (2) use a between-group design in which different subjects are randomly assigned to groups that receive different experimental treatments, often including a control condition.[9]

What is notable is that either of these approaches, as well as far more complex experimental designs, is easily implemented in the context of surveys utilizing computer-assisted telephone interviewing platforms or Internet-based interviews. The ability to make strong causal inferences has little to do with the laboratory setting per se, and a lot to do with the ability to control the random assignment of people to different experimental treatments. By moving the possibilities for experimentation outside of the laboratory in this way, population-based experiments strengthen the internal validity of social science research and provide the potential to interest a much broader group of social scientists in the possibilities of experimentation. Of course, the fact that it *can* be done outside the lab is not a good reason in itself to do so. Thus, below I review some of the key advantages of population-based experiments, beginning with four advantages they have over traditional laboratory experiments, then ending with some of their more general benefits for the accumulation of useful social scientific knowledge.

Population Samples Rather Than Subject Pools

Most laboratory experiments rely on undergraduate subject pools created explicitly for the purpose of providing an ongoing supply of experimental subjects for researchers in one or multiple disciplines. With population-based survey experiments, scholars instead expose randomly-selected respondents to randomly-assigned treatments. A key advantage of this approach is that using survey sampling techniques, researchers can assign both *larger* and *more diverse* samples to experimental conditions of their choosing.

Anyone who has ever executed a laboratory experiment knows that it takes a lot of time, sometimes money, and often other forms of coaxing to encourage large numbers of people to show up in a laboratory at an appointed time. The easiest approach—the subject pool in which students are required to participate in some fashion—has its drawbacks in terms of sheer numbers, as well as the diversity of experimental subjects. In addition, participants in subject pools typically participate in more than one experimental study, thus making them less naïve than one might hope. By contrast, sample surveys regularly include two to three thousand people per study, and population-based survey experiments can do the same. The greater statistical power that comes with large samples makes it possible to identify more subtle differences between experimental groups.

Larger samples also make it easier to identify moderating relationships—that is, situations in which a given experimental effect is not present, or at least not equally so, across different subgroups of the population. Seldom do social science theories serve as true universals. For this reason, it is useful to know precisely what the boundaries are for a given theoretical generalization. For example, one recent population-based survey experiment established that the characteristics that make people more or less attractive to others appear not to be the same for men and women and for blacks and whites in America.[10]

In addition to sheer numbers of participants, population-based experiments also make possible broader, more diverse subject pools. Though not all social scientists require large and diverse subject populations to accomplish their research goals, many can benefit from them. Critics over the years have often questioned the extent to which the usual subjects in social science experiments resemble broader, more diverse populations. As Carl Hovland famously put it, "College sophomores may not be people."[11] Population-based survey experiments offer a powerful means for researchers to respond to such critiques, one that researchers in fields that traditionally emphasize internal validity and experimental methods will find quite useful.

For example, to study the impact of gender on worker performance evaluations, Rashotte and Webster designed an experiment in which participants were presented with a picture of a male or female individual, David or Diane. The photos were taken from a public website, www.hotornot.com, where, for inexplicable reasons, people post their pictures to be rated by anonymous others. Using these photos allowed the researchers to control for attractiveness effects, by selecting those who were rated as average. Respondents were asked about how intelligent they perceived the person to be, and how capable, along with many other characteristics.

Using this framework with college student samples, Rashotte and Webster had found small but consistent differences in evaluations of David/

[9] Holland (1986, p. 947).

[10] See Conley and Glauber (2005) and Conley (2006) for a summary. Unfortunately, most studies of attractiveness to date have used undergraduate student samples, business school graduates, or those who selected into dating services, all of which makes it difficult to identify "universal" characteristics.

[11] Attributed to Tolman, in Hovland (1959, p. 10). See also, e.g., Sears (1986).

Diane among both men and women.[12] Both groups gave higher ratings to David than Diane, regardless of whether they stated beliefs in gender equality or not. The authors concluded that these gender stereotypes operate below the radar such that both men and women expect greater competence from men.

But whether the effects exist in the population as a whole depends on whether the results for college students can be generalized to an older population, and whether the judgments the students made of other young men and women would hold if the targets were older men and women as well. Some have suggested that gender is losing its status significance over time, especially among younger generations, a claim that is difficult to address without more variance in both the targets and subjects involved in these studies.

When Rashotte and Webster replicated this same study in the general population, their findings were far less consistent by gender of target, thus calling into question the generalizability of previous findings based on student samples. Age of the target had a much stronger impact than gender did, and the findings for gender varied based on whether people were evaluating intelligence, competence, or some other dimension of status. Their results indicated that older men and women were consistently rated as more competent than younger men and women.[13]

Experimentalists can use the larger and more representative samples characteristic of population-based experiments to show that observations generated in a laboratory can be replicated under very different conditions and with far more diverse populations. If they do not replicate well, then researchers learn about the boundaries of their theories, which is also useful to the progress of scientific knowledge.

The Real World Over the Laboratory

Rightly or wrongly, evidence from population-based survey experiments is likely to be viewed by external audiences as more convincing than evidence from laboratory studies. No one disputes that laboratory experiments provide strong tests of causal propositions, but scientific and popular audiences often want more than evidence of causality. In many cases, observers want a demonstration that laboratory observations survive exposure to myriad conditions outside of the lab. For example, the laboratory setting is often assumed to make people act more responsibly than would otherwise be the case without the close supervision of the experimenter. People know when they are being watched, and may act differently as a result. In

[12] Rashotte and Webster (2005a).
[13] Rashotte and Webster (2005b).

addition, the distractions of everyday life can reduce the likelihood that a treatment will have an impact; if an effect can only be obtained when unrealistically high levels of attention are directed toward a stimulus, then the effect probably does not occur as often in the real world.

As I emphasize in Chapter 8, too much importance is routinely attributed to laboratory versus field settings, when other factors probably play a much more important role in external validity. However, this is not to say that settings are completely irrelevant. The importance of situational influences on human behavior is clear and well-documented;[14] despite our tendency to think otherwise, the same people will behave very differently under different circumstances. But this fact aptly illustrates the oversimplification surrounding the idea that conducting research in "the field" means it will easily generalize elsewhere. One field setting may be nothing like another, despite the fact that both are outside the lab.

Research Designed Around Specific Subpopulations

Yet another advantage that population-based experiments provide over laboratory experiments is the ability to study specialized subpopulations. Particularly when utilizing respondents from ongoing Internet panels, investigators typically know a great deal about their experimental subjects in advance. These respondents are generally "prescreened" for a variety of characteristics during an initial recruitment interview, so characteristics such as race, region, income, employment status and so forth are already known. This makes it possible to sample specific subpopulations or to block respondents based on characteristics known to affect the dependent variable. Under ordinary circumstances in laboratory studies, sampling subpopulations is either massively inefficient (because large numbers of ineligible participants will need to be screened out and turned away) and/or it makes subjects too aware of exactly why they were selected, which threatens the integrity of the results. The use of ongoing panels eliminates many of these problems.

For example, using equally-sized samples of respondents who had previously identified themselves as white American, black American, or Asian American, Cottrell and Neuberg randomly assigned these three groups to evaluate one of each of these same three ethnic groups.[15] Respondents reported both general favorability toward the specific group, and a list of discrete emotions such as anger, fear, pity, envy, respect, etc. They also reported on a variety of specific kinds of threats that they perceived the randomly assigned group posed to "people like me."

[14] See Ross and Nisbett (1991).
[15] See Cottrell and Neuberg (2005).

The investigators found that outgroup prejudice was not all cut of the same cloth; there were distinct emotional profiles underlying sometimes identical levels of favorability (or the lack thereof) toward a group. Interestingly, in this case their population-based findings closely mirrored results from student samples, with African Americans eliciting mainly fear, pity, and anger, whereas Asian Americans generated pity and resentment, but no fear. This work highlights the fact that prejudice as traditionally conceived often masks very different emotional responses to different groups. Moreover, different groups often have qualitatively different prejudices toward the very same group. This kind of information is quite valuable to those who seek to combat prejudice and would have been far more difficult to obtain within a standard lab setting.

In addition to advantages over laboratory experiments in particular, population-based survey experiments also have some general strengths as a methodological approach. These advantages are not characteristics exclusive to this one method, but they are secondary benefits that help make a case for adding this approach to the broader collection of research tools at our disposal. More specifically, population-based survey experiments are likely to encourage more widespread use of experiments by social scientists, the use of complex experimental designs, more successive studies comprising a larger research agenda, and research that speaks to real world events and policies.

Encouraging Greater Use of Experimental Designs

Experiments play a critical role in advancing social science. Scholars have long recognized the power of experiments, acknowledging them as the best possible approach when attempting to draw causal inferences empirically. For many research questions, experiments are simply the most effective means of evaluating competing causal hypotheses. As much as we would love it if it were true, there simply are no statistical techniques for observational data that provide the power and elegance of an experimental design.

Nonetheless, there remains considerable resistance to the use of experiments in some fields, mainly because of researchers' concerns about the generalizability of research findings produced in a lab and/or with college student subjects. One advantage of population-based survey experiments is that by overcoming these restrictions, they offer the prospect of luring more social scientists into experimental research. By eliminating the most common objections to experimental work, population-based experiments promise to increase the number of scholars and the breadth of fields regularly using experimental designs.

For example, scholars can use population-based experiments to clarify the causal implications of findings from conventional surveys. Experiments often make it possible to resolve the direction of a causal relationship that has been difficult to disentangle. For example, do high levels of social trust lead people to buy products online, or does buying online lead people to have higher levels of social trust? As described further in Chapter 6, there is now experimental evidence that, in fact, both causal processes take place.

In addition to providing a means of resolving direction of causation and ruling out potentially spurious relationships, population-based experiments can help advance theory in research areas where selection bias makes observational studies relatively unproductive. For example, one puzzle in international relations involves the question of whether "audience costs" exist; that is, if one country's leaders threaten another country, are there penalties for the country's leaders if they do not follow through on their threat? Do leaders who make empty threats risk losing the faith of their constituents in addition to losing credibility with the country they have threatened?

Audience costs have long been believed to exist, but it is difficult to document them because if leaders take the *prospect* of audience costs into account when they make foreign policy decisions, then they are unlikely to back down because they believe citizens would react negatively to such a decision. So social scientists doing observational research are denied the opportunity to observe a public backlash against the leader in all but a very few cases; leaders self-select into making only threats that they are willing to back up.

What is a social scientist to do? We could attempt to persuade leaders to make idle threats and then back down just for the heck of it—or at least for the sake of gathering scientific knowledge on the price that would be paid. Unfortunately (or perhaps fortunately), such an appeal is unlikely to be successful with most national leaders, although some undoubtedly have their price. Alternatively, political scientist Michael Tomz administered a population-based experiment to a representative sample of Americans, asking each person about "a situation our country has faced many times in the past and will probably face again. Different leaders have handled the situation in different ways. We will describe one approach U.S. leaders have taken, and ask whether you approve or disapprove."[16]

Respondents were then told about a crisis in which "a country sent its military to take over a neighboring country." As the author describes:

[16] Tomz (2007, p. 824).

The country was led by a "dictator" in half the interviews, and a "democratically elected government" in the other half. The attacker sometimes had aggressive motives—it invaded "to get more power and resources"—and sometimes invaded "because of a long-standing historical feud." To vary power, I informed half the participants that the attacker had a "strong military," such that "it would have taken a major effort for the United States to help push them out," and told the others that the attacker had a "weak military," which the United States could have repelled without major effort. Finally, a victory by the attacking country would either "hurt" or "not affect" the safety and economy of the United States.[17]

Respondents also learned how the U.S. president had handled the situation, either (a) promising to stay out and then doing so, with the attacking country eventually taking over its neighbor; or (b) promising the help of the U.S. military to push out the invaders, but not doing so, with the attacking country taking over its neighbor. If audience costs exist, the respondents who heard that the president threatened but did not carry out his threat should approve less of his handling of the situation than those who heard that he stuck by his original decision to stay out.

Results of the study suggest that audience costs do indeed exist across the population and under broad conditions; those who heard that the president issued an empty threat were significantly more disapproving than those told he never intended to get involved. Experimental designs such as this one allow researchers to test causal hypotheses that there is no other way to empirically verify outside of historical case studies. As reviews of such experiments reveal, population-based experimental designs have already generated important discoveries in many social sciences.[18]

More Complex Experimental Designs

Population-based experiments can accommodate a large number of experimental conditions. In part, this capacity is due to the larger samples noted above. But it is not only a function of sample size. In a laboratory, an experimental design with nearly 100 different experimental conditions would be considered insane to attempt unless the treatments were administered by computer. But in population-based experiments, such numbers are not uncommon precisely because a computer program determines the

[17] Tomz (2007).
[18] Sniderman and Grob (1996).

treatment a given respondent receives, whether it is administered online or via telephone interview.[19]

For example, Sniderman, Piazza, Tetlock, and Kendrick used a population-based experiment to see whether combinations of individual characteristics combined with race would trigger respondents to be less likely to help black rather than white victims of unemployment.[20] Their laid-off worker experiment varied descriptions of the unemployed person along six different dimensions (age, sex, race, marital and parental status, dependability) that were each manipulated orthogonal to the other, creating a total of 96 conditions. A person was described as having been "laid off because the company where they worked had to reduce its staff" and was either black or white, male or female, in their early twenties, mid-thirties, or early forties, and single, a single parent, married, or married with children, along with being (or not) a dependable worker. Respondents were next asked, "Think for a moment about each person and then tell me how much government help, if any, that person should receive while looking for a new job."[21]

Respondents then had the option of suggesting a lot of government help, some, or none at all. The many possible variations of this highly complex treatment were easily administered in the context of a population-based experiment. Further, the larger sample sizes characteristic of population-based experiments allowed sufficient statistical power to identify even fairly modest effect sizes. The sample sizes alone would have been prohibitive in allowing execution of this study in a lab. Given that the characteristics of the target person were each randomly assigned, each attribute was orthogonal to every other attribute, thus making it possible to assess the statistically independent impact of each characteristic as well as their interactions.

By facilitating more complicated experimental designs, population-based experiments allow researchers to study complex interactions. For example, in the study described above, the "new racism" thesis suggested that white respondents (and white conservatives in particular) would be especially unlikely to recommend government help for black victims of unemployment who violated social norms, either by failing to keep their family intact or by failing to be reliable workers. Such was not the case

[19] Although computer-administered experiments can be done in a lab setting, what would be the point of using a convenience sample in a lab, when the same study could be administered to a random sample via computer as well? Cost would seem to be the only advantage (assuming laboratory subjects are free, which is not always the case).
[20] Sniderman et al. (1991).
[21] Sniderman et al. (1991, p. 427).

according to their results, but without a complex factorial design, it would have been difficult to test this hypothesis.

Facilitating Research Agendas

Relative to traditional surveys, population-based experiments are well suited to take advantage of sequential data collection efforts, mainly because this approach tends to be driven to a greater extent by the desire to test specific hypotheses and because it requires the collection of a relatively small number of variables. As with any experiment, the whole point of a population-based experiment is typically to evaluate a specific hypothesis or a small set of hypotheses. By contrast, much (although certainly not all) non-experimental survey design seeks to collect all the data the researcher can think of (or afford) bearing on a particular issue, and then see what emerges as important later in the analysis. Survey analyses are observational rather than experimental, so a large number of questions must be asked in order to rule out potentially spurious relationships. Because opportunities for survey data collection are limited and may be restricted to just one large effort, survey researchers must anticipate in advance all possible variables needed for many possible hypotheses, and then put them together in one survey.

This characterization is not a critique of survey research or of survey researchers. While it is tempting to think that whatever additions or refinements that could be made on a subsequent survey could have been done on the first survey if the researcher had just thought long enough and hard enough, this is simply not how social science (or natural science, for that matter) progresses. Often the hypotheses that form the basis for the second data collection depend upon what was learned from the first data collection. But given that relatively few social scientists have even one such opportunity for original data collection, still fewer benefit from successive studies of the same topic. Survey research generally *must* be of an omnibus nature due to the need to include potentially spurious "third variables," and the need for a larger sample size due to statistical issues.

Successive population-based survey experiments tend to be more self-contained. The experimental design means that the study requires a much smaller number of questions than an observational study testing the same hypothesis. In addition, the use of simpler statistical models means that one can often get by with a smaller sample size. The treatments are designed to test particular hypotheses, rather than dozens and dozens of potential hypotheses. As a result of the more focused quality of experiments, they tend to be smaller and less expensive, thereby conserving resources. Resources can be put toward subsequent studies that build on what was

learned in the first. For these reasons, population-based experiments are well suited to the gradual, systematic accumulation of knowledge.

Research that Speaks to Real World Events and Policies

Population-based experiments are ideal for taking advantage of quickly unfolding opportunities for experiments embedded in real world events. So-called "firehouse studies" that are executed quickly in response to naturally occurring and often unpredictable events are facilitated by the existence of infrastructure to sample and contact respondents. In the past, opportunities for studies of this kind were scarce because of the substantial lead time needed to contact appropriate samples and get studies into the field. Because Internet-based survey panels have already established contact with potential participants and often have gathered extensive demographics in advance as well, it is possible to contact a large number of respondents quickly as events arise that researchers wish to study.

For example, in the midst of the 2004 controversy over the Georgia state flag, which still featured the Confederate battle emblem at the time, Hutchings and Walton did a population-based experiment focused strictly on residents of the state of Georgia using a blocked design such that half of the respondents were black and half were white.[22] A number of other population-based experiments were designed around real world events such as Hurricane Katrina, which ravaged New Orleans in 2005, as well as the role of partisanship in attributions of blame for the terrorist attacks of 9/11.[23] In short, population-based survey experiments make it easier to jump on real world events that present unique research opportunities.

In addition to event-based studies, population-based survey experiments also offer advantages for studies designed specifically to inform public policy. Take, for example, the study described at greater length in Chapter 3, which examined how much extra money per month people would be willing to pay for health insurance that includes new vaccines for themselves and their children—vaccines that otherwise must be covered out of pocket.[24] The answer to this research question is obviously relevant to the design of healthcare options to maximize public health.

But assuming a study is designed to answer this question, what kind of study would maximize its policy impact? The answer becomes clear after considering the more traditional options for empirical research, surveys,

[22] Hutchings, Walton, and Benjamin (2005).

[23] For example, Huber, Van Boven, Park, and Pizzi (2006, 2007); Shaker and Ben-Porath (2010); Malhotra and Kuo (2008, 2009); Willer and Adams (2008).

[24] See Davis and Fant (2005) or Chapter 3 of this volume.

and experiments. Would a survey simply asking a representative sample of people how much they would pay for this benefit produce a convincing course of action? Possibly, but given that most people have no idea up front how much vaccines should cost, their answers are likely to be all over the map. Perhaps more importantly, using a traditional experiment, would policymakers really be swayed to alter the structure of health care options by evidence based on a sample of college students, most of whom do not have children and have never purchased health insurance or paid premiums through their place of employment? This seems highly unlikely. But by conducting an experimental study with a representative sample of parents, researchers are able to provide policymakers with greater confidence in the findings while also enhancing the relevance of their research to real world policy questions.

Overview

In this introductory chapter, I have outlined the development of population-based experiments and highlighted some of their advantages over traditional experiments and surveys. In the remainder of the book, I make the case for this method as a unique contribution to social science methodology, and provide extensive examples of how it has been used thus far. I also discuss some of the problems that are common to population-based survey experiments, and suggest best practices for future users.

There is a tendency to think about population-based survey experiments as simply a hybrid methodology that melds certain characteristics of surveys and experiments. But to say this tells us nothing about *which* advantages and disadvantages of each methodology are inherited. As I further argue in the concluding chapter, population-based experiments are not simply a mix of two methods in the sense that quasi-experimental designs are a hybrid of observational and experimental techniques.[25] Instead they are more akin to an agricultural hybrid that produces something that was not present in either of the two original plants. To the extent that population-based survey experiments can be implemented with effective treatments and with the same degree of control over random assignment as in the lab, it is the only kind of research design capable of simply and straightforwardly estimating population average treatment effects without complex statistical machinations. This characteristic makes population-based experiments unique. They are not without their limitations, to be

[25] For example, some suggest that quasi-experiments inherit the weaknesses of both experiments and field studies (Marsh, 1982).

sure, but those limitations do not lie in our ability to draw important conclusions from them so much as in our ability to execute this method well.

Toward that end, Part I of the book focuses on the greatest challenge facing population-based experiments—how to implement effective treatments outside the laboratory via remote communication technologies. I provide an overview of many different kinds of population-based experiments, drawing on the wealth of creativity provided by TESS participants as well as other researchers. These designs are by no means the only or necessarily the best possible ways to execute population-based experiments. But I provide this sample template of designs as a means of stimulating readers' own creativity as to the possibilities that population-based experiments offer their field of research. In Chapter 2, I describe population-based experiments designed to improve measurement. These are descendants of the early split-ballot approach, also geared toward improving measurement of attitudes and behaviors, but the approaches are now far more sophisticated and complex. In Chapter 3, I describe direct and indirect treatments, that is, treatments that either directly and overtly try to move the respondent in a particular direction, or that indirectly do so in a more unobtrusive (and often clever) manner. Chapters 4 and 5 cover two approaches to implementing highly complex population-based experimental treatments: vignette treatments and game-based treatments, respectively. Vignettes offer the possibility of easily executing complex, multi-dimensional factorial designs. Game-based treatments are an outgrowth of conducting experiments online, where gaming seems only natural, and where highly complex, multi-stage experimental treatments can be experienced by participants. Ideally, Part I of the book should serve as an organizational framework and as an idea generator, helping those in one discipline see the promise of what has been pilot-tested already in another.

In Part II, I address a series of practical matters in the design and implementation of population-based survey experiments. The TESS staff benefited from its involvement in hundreds of different population-based experiments, and we learned a lot from these experiences. In Chapter 6, I try to eliminate the need for others to learn by trial and error as we did. Hopefully some of our errors will allow others to avoid wasting time and money by falling prey to the same problems. The practical issues addressed range from "How do I explain a population-based experiment to my Institutional Review Board (IRB)?" to "What can I do to maximize the effectiveness of a treatment in a population-based experiment?" to "How should I think about measurement differently from when I am designing a survey?" Not surprisingly, different disciplines had different problems adapting to the idea of a population-based experiment, and I use various war stories from TESS to illustrate the kinds of problems most likely to plague users from different disciplines.