Survey Experiments and the Quest for Valid Interpretation

Gustavo Diaz, Christopher Grady, and James H. Kuklinski

When Diana Mutz wrote *Population-Based Survey Experiments* in 2011, she stressed one theme throughout the book. That theme: the use of large random samples with experiments embedded in them is an ideal means by which to generate causal generalizations. The embedded experiment provides the needed leverage to identify true cause and effect, and the random sample of a national population ensures that the results can be generalized to the population from which the sample was drawn.

Mutz's logic remains as compelling today as it was when she wrote the book. However, two significant changes have occurred. First, the increasing influence of the causal inference movement has changed political scientists' priorities with respect to data collection and analysis. Because causal inference emphasizes making the right comparisons, not generalization, researchers increasingly search for unique, often local, research opportunities, thus avoiding the costs and delays associated with the collection of random samples of national populations. Their practice resonates with Campbell's long-ago assertion about social scientific practices: 'There was gross overvaluing of, and financial investment in, external validity, in the sense of representative samples at the nationwide level. In contrast, the physical sciences are so provincial that they have established major discoveries like the hydrolysis of water… by a single water sample' (Campbell, 1988, cited in Rosenbaum, 1999).[1] As a result, individual scholars' own research programs have progressed quickly, and, more significantly, these same scholars have been able to respond to and build on others' work in rapid-fire fashion.[2]

Second, measurement has emerged as a distinct and very active area of experimental survey research, with some of the discipline's best methodologists working in it.[3] Much of the effort has focused on the measurement of sensitive attitudes. The ingenuity of the designs that scholars have used to identify 'true' attitudes has been nothing short of

remarkable. We put quotes around true attitudes, since one of the key developments in measurement has been a continuing change in conceptions of what true attitudes are.

In both areas of experimental survey research, inference and measurement, scholars seek to interpret their results correctly (i.e. validly). In the case of causal inference, the goal is to reach proper conclusions about relationships between treatments and outcomes. In the case of measurement, the goal is to construct methodological approaches that increase confidence in the measurement of concepts such as prejudice, given that respondents often seek to hide their true views or do not consciously understand what they truly feel and think.

In both types of survey experimental study, scholars might misinterpret the empirical results. But why expect scholars to misinterpret their results? After all, scholars have been describing well-designed experiments as the gold standard by which to estimate true causal relationships for centuries. And no one would doubt that most political scientists can capably design strong survey experiments these days.

Considering causal inference experiments first, we identify and discuss three potential sources of misinterpretation of results: factors not included in the experiment moderate the basic treatment-outcome relationship; some people enter the experiment already having been treated in the very external world the researcher seeks to understand; and, respondents enter the experiment with different experiences, which are typically unknown to the experimenter and which shape the way respondents interpret the treatments. As part of our discussion, we evaluate some increasingly complex methods that scholars have proposed to overcome the sources of misinterpretation.

With respect to survey experiments designed to uncover true attitudes when social desirability might be coloring the respondent's true beliefs and feelings, we identify and discuss several problems that might undermine the increasingly complex designs that scholars have brought to bear, thus leading them to misinterpret the results. The biggest obstacle to proper interpretation of results is the lack of full respondent anonymity, as viewed by the respondents themselves. Other problems, such as contamination from earlier questions in the survey, stem from the survey context, not necessarily from features of the experiment.[4] With respect to implicit attitudes, the fundamental problem that can impair interpretations of the results, in addition to some of those discussed in the preceding paragraph, arises when the treatments do not prime the concepts researchers think they have primed.

In both areas of survey experimental research, the designs of the experiments have increased in complexity over time. Two obvious questions, which we keep in mind throughout our discussion: has the increasing complexity of experiments increased scholars' capacities to make right inferences about the outside world? And has this increasing complexity wrought its own set of problems, or at least does the potential exist?

To be clear, we define survey experiments as experiments in which the treatments are delivered through a survey instrument. This excludes field experiments that use surveys to measure outcomes (Broockman et al., 2017). Conversely, if respondents enter a laboratory and are assigned to different treatments via different of a survey item, the study meets the definition of a survey experiment.

We have divided our discussion into three sections. First, we discuss reasons why political scientists can inadvertently misinterpret their results when conducting causal inference studies. Second, we undertake the same task with respect to measurement studies. In both instances, we also discuss designs that scholars have begun to propose to avoid misinterpretation. Finally, we pursue some implications of our earlier discussions. The overall purpose of our discussion is to highlight key areas that require the attention of both experts in the field and junior scholars interested in incorporating survey experiments into their research toolkits.

One of the most difficult aspects of writing this chapter is the knowledge that we will not be able to cite the many meritorious studies that warrant citation. While they originated in American politics, survey experiments are now common across all subfields of political science and international relations, exploring topics as different as immigration (Hainmueller and Hopkins, 2015), vote buying (Gonzalez-Ocantos et al., 2012), corruption (Winters and Weitz-Shapiro, 2013), the democratic peace (Tomz and Weeks, 2013), compliance with international treaties (Findley et al., 2017), and support for extremist groups (Blair et al., 2013). We could dedicate an entire chapter to listing interesting applications. Too keep it simple, we have chosen studies with which we are most familiar and that help us make the points we seek to make. We rely on readers to draw connections with the topics that are salient in their areas of expertise.

SURVEY EXPERIMENTS FOR CAUSAL INFERENCE

Survey experiments for causal inference are experiments that happen to be embedded in a survey instrument. Respondents are randomly assigned to different versions of a treatment, and then they answer one or more outcome questions.[5] Much like in field and laboratory experiments, the researcher can identify an average treatment effect on one or more outcomes of interest (see Bowers and Leavitt, Chapter 41; Morton and Vásquez-Cortés, Chapter 51; Sinclair, Chapter 52; and Wilke and Humphries, Chapter 53, this *Handbook* for details). Because scholars in the discipline use causal inference survey experiments to illuminate real-world social and political phenomena, however, finding a non-zero treatment effect is not enough. Researchers must also convince others that their interpretations of treatment-outcome relationships are valid.

Anyone acquainted with causal inference in the social sciences might respond, 'of course, how could it be any other way?' In answer, scholars have identified at least three distinct challenges to valid interpretation, which, in our view, must be taken seriously and addressed head-on.

First, survey experimenters can easily overlook or not be able to incorporate factors that moderate the relationship between treatment and outcome (confounding). Second, some respondents might come to the survey experiment having already been pretreated in the very world to which the researcher is trying to infer (pretreatment contamination). Third, respondents might interpret the same treatment in a survey experiment differently due to differences in life experiences and the nature of the environments in which they live, which can be tantamount to receiving different treatments altogether (lack of information equivalence). Ignoring any of these possible complications can lead to wrong interpretations of estimated treatment-outcome relationships.

Confounding

The most obvious challenge to interpretation in causal inference survey experiments is confounding, which arises from the omission of one or more factors that moderate the relationship between treatment and outcome. This is akin to the problem of omitted variable bias in observational studies (King et al., 1995). The analogy might sound counterintuitive, as we are taught that experiments balance the distribution of both observed and unobserved covariates across groups. However, survey experiments randomize constructs that are not necessarily independent from each other outside the survey framework. Consequently, a treatment in a simple two-group design might inadvertently activate elements that correlate with the treatment in the real world, so that the researcher cannot disentangle the effect of a manipulated treatment from the confounder that is activated indirectly.

Consider the study of corruption. A recurrent debate in the literature is whether citizens sanction corrupt politicians with their votes. Since scholars cannot manipulate corruption, they use hypothetical situations in survey experiments to understand voters' reactions to corrupt politicians and hope that the findings translate to actual voting behavior.[6]

The simplest design presents respondents with a vignette describing a current officeholder seeking reelection. The control group receives information about the incumbent's profile only, while the treatment includes additional information about the incumbent's illicit activities. The design logic is straightforward: if voters sanction corrupt politicians in the treatment group, then, by inference, all that prevents voters from sanctioning corrupt politicians in the external world is the absence of credible information. More bluntly, when voters do not punish corrupt politicians, it is probably because they are unaware of their bad deeds.

Simple and elegant as this design is, it ignores the possibility that voters also respond to other activities the politician undertakes – and perhaps to the politician's personal characteristics as well. Any of these factors could moderate the original relationship between corruption and vote. In the extreme case, the inclusion of such other factors eliminates any initial effect between corruption and vote, which raises questions about the validity of the original interpretation (i.e. that information is the key). Thus, subsequent studies on corruption manipulate not only corruption, but also the provision of public goods, shared partisanship, and even gender (Anduiza et al., 2013; Winters and Weitz-Shapiro, 2013; Eggers et al., 2018). The accumulated evidence in these studies suggests that other factors moderate the relationship between corruption and vote.[7]

In almost all experimental studies, it is easy to identify several potential confounders. In the preceding example, other possible confounders include coercion, vote-buying, a politician's experience in office, and the credibility of the source (Botero et al., 2015; Mares and Visconti, 2019; Weitz-Shapiro and Winters, 2017). All of these could simultaneously confound the relationship between corruption and vote – and in different directions. However, traditional survey experimental designs set a low limit on the number of potential confounders that can be included, which invariably raises the annoying and ever-present possibility that the experimenter's conclusion will be wrong, or at least incomplete.

Factorial survey experiments (see Auspurg and Hinz, 2015, and Sniderman, 2018, for overviews) provide one way to address the confounding problem, in that the researcher can manipulate both the explanatory variable of interest and a large number of confounders. However, including many potential confounders comes at the cost of statistical power. The researcher now faces a trade-off between accounting for all potential factors that get in the way of proper interpretation and the capacity to identify a non-zero treatment effect.

Conjoint experiments (Hainmueller et al., 2014) overcome this problem by combining clever design and technological advancements in computer-assisted surveys. In the standard conjoint design, the researcher presents respondents with multiple choice tasks between two or more hypothetical alternatives. The combinations consist of independently randomized attributes. In our earlier example, an experimenter might present respondents with two candidates for office. For each candidate, respondents see information about the candidate's level of public goods provision (low or high), party affiliation (left or right), gender (male or female), and so on. In turn, each one of these attributes is randomly assigned to take one of the values included in parenthesis. Because the exercise is hypothetical, the researcher can repeat this exercise multiple times.

Conjoint experiments, then, can incorporate an unusually large number of factors because respondents answer multiple choice tasks that completely randomize the attributes of each alternative, allowing researchers to explore a wide range of combinations before running

into power limitations. Some tasks, like voting, can be reasonably presented as a choice between two or more alternatives. Others, such as the study of immigration attitudes, where researchers ask respondents to put themselves in the role of an immigration officer to determine which individuals should get priority in the admission process, require more creativity (Hainmueller and Hopkins, 2015).

This increased leverage to address confounding comes with a caution: a seemingly impeccable logic, not empirical reality, drives the methodology. This reality opens the door to possible invalid interpretations of empirical results as they apply to the external world. One potential problem is that the quest for satisfying the logic of the methodology itself can come at the cost of realism, in that some combinations rarely, if ever, exist in the bigger world. For example, if a study randomizes occupation and education levels independently, respondents could potentially encounter a doctor with no post-secondary education (Hainmueller and Hopkins, 2015).[8]

Moreover, scholars can now include virtually as many factors as they can imagine, limited only by statistical power and their own discretion about which factors to include and which to exclude. Which factor has the largest effect size presumably depends heavily on the choices researchers make,[9] and thus misinterpretation of results once again becomes a potential problem. Researchers can be misled into thinking that a given cause is more important than others, although, in fact, any result is the product of choices.

Hainmueller and colleagues offer several valuable examples to emulate. In those examples, the use of well-established theories drives the crucial choices. Unfortunately, we already see a tendency in the work that followed the introduction of conjoint experiments to discard theoretical justification and view every factor in the research design as a treatment. This changes the purpose of the study from proper interpretation centered on one treatment of interest to a horse race to determine which factor has a larger effect size.

Pretreatment Contamination

As we already mentioned, the goal of causal inference survey experiments is to learn about attitudes and behaviors beyond the survey framework. This presents researchers with an interesting dilemma: a research question worthy of pursuit is also likely to be one where respondents encountered the treatment of interest prior to and outside the experiment. This very pretreatment, if not accounted for, can generate wrong interpretations about the effect of the treatment (Druckman and Leeper, 2012; Gaines et al., 2007).

As one of us noted in earlier work, the effect of pretreatment contamination depends on two factors: when the pretreatment occurs vis-à-vis the survey experiment and the longevity of the pretreatment effect, assuming there is one. We do not repeat the details here, except to say that, depending on the existence and endurance of pretreatments, the same static experiment can generate conclusions ranging from no effect at all to a large effect.

In short, survey experiments and the contexts to which experimenters seek to infer can and often do interact across time in highly complex ways. At the extreme, researchers cannot correctly interpret the experimental results without a thorough understanding of the contextual dynamics. However, if they are intimately familiar with the dynamics that happen in the world to which they are trying to infer, they probably can live without the experiment.

Note that attention to pretreatment moves the focus to dynamics, a shift that should resonate with most scholars. After all, the phenomena in the world that scholars study are dynamic. However, nearly all experimental designs are static, and thus they usually are incapable of addressing the effects of pretreatment. What to do?

One possibility: the researcher could simply include a separate question in the survey asking respondents if they have experienced a version of the treatment recently. This is problematic because the question can trigger different complications depending on

its placement. Including the question before treatment is problematic, in that respondents who did not experience the treatment before might be primed by it and approach the experiment as if they have been pretreated. In other words, the experimenter risks replacing pretreatment contaminations with primacy effects. Conversely, including the question after treatment can trigger a false memory.

Chong and Druckman (2013) propose two possible ways by which to identify and account for pretreatment effects: directly manipulate pretreatment and trace effects over time or find a real-world situation where some respondents have been pretreated and others have not.[10] In a first study, they estimate the effects of two competing frames for and against increased law enforcement. To address pretreatment, they conduct a two-wave panel study in which respondents do or do not receive treatment in the first wave (i.e. they are pretreated or not). In the second wave, they explore whether those not treated in the first wave show greater response to the second-wave treatment than those who had been pretreated. They find a big difference in second wave responses, with those not treated earlier showing greater response to the second-wave treatment. In their second study, they take advantage of a situation where some people have followed a controversy and others have not. Again, the results support the idea that pretreatment affects the experimental results.

Although the Chong-Druckman approach provides leverage on pretreatment contamination, implementing a short panel might be out of reach of many research budgets. Moreover, the field is converging in the opposite way. The norm is to perform increasingly complex one-shot studies, and, when resources permit, the preferred option is to replicate the same experiment with a different sample.

Lack of Information Equivalence[11]

Suppose that a researcher designs a survey experiment that satisfactorily addresses confounding and pretreatment contamination. Can that researcher justifiably claim valid interpretation? The answer is a resounding 'no'. In fact, the final challenge to proper interpretation that we consider, and which only recently has come to the fore, is both the most pervasive and most difficult to resolve. The problem is what Dafoe et al. (2018) call (a lack of) information equivalence. Once expressed, its logic is intuitive, even though solving the problem currently fringes on the impossible.

At the risk of oversimplifying, the idea goes as follows. Respondents routinely interpret and answer survey questions in terms of their life experiences. These experiences will vary greatly, especially when the experiment is embedded in a national survey. To the extent that people's life experiences are sufficiently strong to influence their interpretations of treatments, the result is that, even though the experimental treatment is the same for everyone, different respondents essentially respond to different treatments. How different depends on how much the contextual considerations vary across respondents, and how much those considerations influence their responses.

To return to one of our previous examples, consider an experiment in which information about corruption primes different thoughts in the minds of a respondent from a rich and highly educated district, and a respondent from a poor district with low education levels. To the former, corruption might mean 'committed a heinous and inexcusable white-collar crime'. To a respondent from a poor district with low education levels, where constituents depend on their officeholders for assistance, corruption might mean 'doing a good deed'.

Note that the lack of information equivalence undermines basic tenets of both survey and experimental research. On the survey side, researchers must assume that the same question means the same thing to all respondents (King et al., 2004). On the experimental side, lack of information equivalence violates the Stable Unit Treatment Value Assumption (SUTVA), which states that all units assigned

to receive a treatment experience it in the same way (Cox, 1958).

This lack-of-information-equivalence problem cannot be easily solved without altering the scope of the study. Consider the running example in Dafoe et al. (2018). The democratic peace is a proposition in international relations suggesting that democracies never go to war with other democracies (Russett, 1993). Two alternative explanations underlie this proposition. First, democracies dislike war generally and are less likely than autocracies to go to war with anyone. Alternatively, democracies perceive other democracies as less threatening, so they only go to war less often with fellow democracies. To assess between the two explanations, Tomz and Weeks (2013) used a survey experiment that presented respondents with a hypothetical country in the process of acquiring nuclear weapons. They randomly presented the country as a democracy or dictatorship, and respondents indicate whether they favor or oppose a military intervention from their home countries.

In this study, the treatment is a country's political regime. The study deliberately avoids including explicit country labels to prevent confounding. However, the typical democracy that is suspected of developing nuclear weapons (e.g. Israel) is remarkably different from the typical dictatorship that carries the same suspicion (e.g. North Korea). Note that the limitation here is different from confounding. Even if the researcher manipulates one of the main potential confounders (e.g. economic development), the challenge to interpretation will persist. Rich democracies with nuclear power (e.g. France) still differ from poor democracies with nuclear power (e.g. Pakistan).

Whereas increasingly complex designs help with respect to the first two challenges to valid interpretation (confounding and pretreatment contamination), the verdict on whether they can help with the lack of information equivalence problem remains to be seen. The problem is more encompassing and much more difficult to overcome. The number of possible interpretations of a treatment are countless. 'Good theory' might convince an audience depending on the application, but we currently do not see plausible solutions to the problem itself. The authors themselves could only hint at possibilities.

The lack-of-information-equivalence problem, we might note in closing, is inherent to surveys and survey responses. It is not a derivative of experiments. To overcome the lack of information equivalence problem would not only bring elation to survey experimentalists, it would bring elation to all researchers who use surveys.

Summary

On the surface, conducting survey experiments is straightforward and deceptively easy. Moreover, the cost is relatively low, which makes them especially attractive to graduate students. In fact, the challenges to valid interpretations of survey experiments are many. These challenges, we emphasize, would not be apparent had earlier generations of survey experiments not existed. Increasingly, these challenges have become apparent, and the next generation of survey experimentalists presumably will be more aware of them as they create their own experimental designs. As a result, if the past is any indication, the demand for increasingly complex and sophisticated survey experiments will continue to grow.

SURVEY EXPERIMENTS AS A MEASUREMENT TECHNIQUE

Scholars who use survey experiments to measure attitudes on sensitive issues, or to measure implicit attitudes that respondents themselves fail to see, also seek to design experiments that facilitate proper interpretations of the results. They, too, have encountered not-easily-identified or remedied problems that complicate the task.

Many, although not all, of the potential challenges in the study of measurement are a function of the survey context, as opposed to, as we saw earlier, the features of the larger context that can influence what respondents bring to the survey. With respect to studies that measure attitudes on sensitive issues, scholars routinely assume that eliminating respondents' perceptions of a lack of anonymity is the key to obtaining honest responses and is thus the key to researchers correctly interpreting the experimental results. Scholars who study implicit attitudes face the same challenge, plus the possibility of a lack of information equivalence, which, as we have already seen, is a potentially big challenge in causal inference studies.

List Experiments and Randomized Response Techniques

Scholars take for granted that respondents do not always answer survey questions truthfully when they are asked about sensitive issues for which there are 'right', or socially desirable, answers. To overcome this possible social desirability bias,[12] researchers have developed and refined two types of survey experiment: the survey list experiment and the randomized response technique. Both techniques are designed to convince individual respondents that their responses to questions about sensitive issues cannot be traced to them. There is some, albeit limited, evidence that these techniques induce less bias than direct questions (Blair et al., 2015; Lensvelt-Mulders et al., 2005b; Rosenfeld et al., 2016).

In a list experiment, the researcher randomly assigns respondents to one of two (or more) conditions. Individuals in the control condition are presented with a list of items; individuals in the treatment condition see the same list plus an additional item, which is the item of interest and the one on which the experimenter wants to ensure the respondent of anonymity. The average difference between the treatment and control conditions represents the percentage of respondents who responded to the sensitive item in a 'socially undesirable' way (Blair and Imai, 2012).

The randomized response technique (Boruch, 1971; Warner, 1965) is one of the oldest techniques for asking sensitive survey questions.[13] In the most common version of a randomized response question, the respondent is directly asked a yes or no question about a sensitive topic. The respondent is also given some randomization device, like a coin or die. The respondent is told to answer the direct question when the randomization device takes on a certain value (tails) or to say 'yes' when the randomization device takes a different value (heads).[14] Users of the method assume that respondents will believe their anonymity is protected because the researcher cannot know whether a 'yes' resulted from agreement with the sensitive item or the randomization device. Researchers know the expected distribution of the condition, which allows an estimate of overall agreement with the sensitive item (See Lensvelt-Mulders et al., 2005a and Blair et al., 2015 for summaries).

How likely is it that respondents will perceive their answers to socially-sensitive matters as protected and thus truly anonymous? More specifically, what would it take for them to feel their answers are anonymous, especially if they already harbor suspicions? If they do not perceive the safety of autonomy, they will likely shape their responses to portray themselves in the best light possible, rather than answer honestly (Leary and Kowalski, 1990).

There are conditions under which the basic list experiment will almost surely fail to provide anonymity. Most obviously, if all or none of the items on the list anger respondents, those who seek to hide their true feelings and attitudes must answer dishonestly (Blair, 2015). Respondents might not interpret other response options as fully anonymous, either. If the treatment item is something respondents want to renounce unequivocally, they

might report a very low number to dissociate themselves from that item, on the logic that being associated with 'three of the four [list items] may be interpreted as a 75% chance' that the respondent holds the socially undesirable attitude (Zigerell, 2011: 544).

The most widely used randomized response technique also offers only limited anonymity to respondents. If a respondent answers 'yes', the answer *could* have been dictated by the randomization device, but it could also signal agreement with the sensitive item (Edgell et al., 1982; Yu et al., 2008). Thus, answering 'yes' is not unequivocally protected by the design. This response bias can affect respondents who do not hold the sensitive attitude just as readily as it affects respondents who do hold it. Respondents who hold the sensitive attitude might say 'no' when directed to be truthful, and respondents who do not hold the sensitive attitude might say 'no' when directed to say 'yes' (Edgell et al., 1982).

Knowledge of the various problems has helped researchers sharpen both survey experiments and randomize response techniques as tools for measurement. As researchers have learned more about the ways in which respondents respond to survey experiments designed for measurement, they have developed more complex and penetrating list experiments and randomized response techniques to account for them. In the process, researchers arguably have added some complexity to get closer to the right interpretations.[15]

For list experiments, the added complexity comes from increased attention to preparation and design. In terms of preparation, researchers pay even more attention to piloting to find control items that not only fit with the treatment item, but are negatively correlated with other control items (Glynn, 2013). Negatively correlated control items minimize the number of people who will score very high or very low on the control list, a problem that can compromise anonymity.

Variations on the list experiment have helped to isolate the effect of the treatment item. One variation is the double list experiment (Droitcour et al., 2004; Glynn, 2013), which attempts to solve the problem of respondent interpretation by using two control lists. The treatment item is randomly selected to appear on either the first or the second control list so that some respondents see it on the first list and some respondents see it on the second. If researchers observe the same treatment effect on both lists, there is less risk that the effect depends on the choice of control items or on how respondents interpret the list. Another modification is a placebo-controlled list experiment, which uses a fourth item as a placebo on the control list to ensure that the difference between the two lists is due to the treatment item, not the presence of an extra item (Riambau and Ostwald, 2019).

Users of survey experiments have come up with variations in the randomized response techniques so as, first, to provide what respondents will view as full anonymity and, second, to keep them from viewing one response as riskier. One such variant is the crosswise model (Jann et al., 2011; Yu et al., 2008).[16] In the crosswise model, respondents are presented with two statements, one sensitive statement and one non-sensitive statement, for which the population mean is known. The respondent is asked to say if neither or both statements are true or if one statement is true. Unlike a typical randomized response question, where individuals who agree with the sensitive statement only occupy the 'yes' group, the crosswise model allows people who agree with the sensitive statement to occupy either group.[17]

Beyond Ensuring Anonymity

The jury is out on the effectiveness of these new techniques. They appear to provide something closer to true anonymity, so they come closer to revealing 'the truth' than their predecessors. However, is ensuring anonymity a *sufficient* condition to obtain honest answers

to sensitive questions? It is unlikely to be a *necessary* condition – see coercive measures like the bogus pipeline (Jones and Sigall, 1971) for techniques to obtain honest responses that ignore anonymity altogether – but unspoken in work using list experiments and randomized response techniques is the assumption that respondents will answer honestly if they perceive their answers to be anonymous.

We see several reasons why anonymity is *not* a sufficient condition to obtain honest answers to sensitive questions. First, even with anonymity, respondents have no incentive to answer honestly. If a prejudiced person is presented with a list experiment that uses a treatment item designed to measure prejudice, what incentive does that prejudiced individual have to comply with the instructions of the list experiment? In addition to anonymity, a further assumption must be made: respondents want to express their socially undesirable opinions in a way that eludes social sanctions.

Second, anonymity does not help respondents interpret the question as the researcher intended. When the purpose of a question is unclear, respondents must either increase their own cognitive efforts in order to understand the question or satisfice and provide an answer that seems reasonable, even without understanding the question. All survey questions assume that the respondent interprets the question in the way intended by researchers; techniques to ensure anonymity make that interpretation less likely by obfuscating the question's purpose.

Anonymity does not solve many other pitfalls familiar to survey questions and survey experiments. It does not help researchers to avoid question ordering effects or contamination from earlier questions in the survey; it does not reveal how respondents interpret the sensitive item and thus cannot ensure information equivalence. Who knows what other novel problems it does not address? Future research should further explicate the assumptions necessary to obtain honest answers to sensitive questions. Future research should also reveal further limitations of techniques to measure sensitive attitudes.

One limitation is clear even without further research: these questions do not uncover implicit attitudes. Many sensitive topics appear so sensitive that individual's conscious, explicit attitudes differ from their implicit attitudes (Greenwald and Banaji, 1995). Even many non-sensitive attitudes seem to be beyond an individual's conscious awareness (Nisbett and Wilson, 1977). Techniques like list experiments and randomized response techniques purport to offer anonymity so that respondents feel comfortable revealing their unsavory conscious attitudes, but these techniques do nothing to draw out attitudes that respondents do not know they have. In the next section, we cover survey experimental techniques to reveal respondents' implicit attitudes.

Priming and Implicit Attitudes

Whereas techniques to measure explicit attitudes seek to provide respondents with anonymity, techniques to measure implicit attitudes seek to keep the respondent consciously unaware of the implicit attitude being measured. To do so, researchers use priming experiments.[18]

In a priming experiment, researchers expose respondents to a stimulus representing topic X in order to influence their responses to a survey question about topic Y, without the conscious knowledge of the respondents. A control group is not exposed to the stimuli representing topic X, so the difference between the treatment group and control group is due to exposure to the treatment stimuli. Priming experiments work by directing respondents' consciousness away from topic X and towards topic Y so that respondents do not consciously censor their feelings about topic X (Macrae et al., 1994; Schwarz and Clore, 1983).

The earliest priming experiments simply randomized the order in which questions were asked (McFarland, 1981). For example,

Schwarz and Bless (1991) show that a question about someone's marriage *before* a question about their general life satisfaction increases life satisfaction for people with good marriages and decreases it for people with bad marriages. The priming paradigm would later be used to measure sensitive attitudes about social groups by priming the social group and then asking respondents about a related topic. For example, Hurwitz and Peffley (1997) prime race by randomizing whether the target of law enforcement action was white or black. This allows them to determine the effect of race on judgments of crime and punishment.

While these priming experiments are extremely innovative, they are also so simple and straightforward that suspicious respondents might realize a connection between the sensitive item being primed and the non-sensitive item being solicited. And, because they measure implicit attitudes by estimating the relationship between the prime and outcomes, they also suffer from the flaws of information equivalence and confounding we discussed in the previous section.

To prevent subjects from ascertaining the goal of the study, researchers try to hide the prime amid other, ostensibly more important, information. One way to do this is with an endorsement experiment (Cohen, 2003). In an endorsement experiment, respondents are asked how much they support a policy. In the treatment condition, the policy is 'endorsed' by a group that respondents would not consciously admit to influencing their opinion. In the control condition, the policy is not endorsed by any group. The average difference in support between the endorsed and unendorsed policy represents the change in support for the policy because of the endorsement.[19]

Though endorsement experiments help hide the goal of the study by distracting attention away from the group and towards a substantive policy, they still suffer from lack of information equivalence. In endorsement experiments, the problem manifests itself because a group's endorsement may be used as a heuristic to understand substantive policy details (Lupia, 1994). The basic endorsement experiment cannot differentiate bias towards the endorsing group from use of the endorsing group as an information heuristic. To ensure information equivalence, researchers utilize endorsements as part of factorial experiments that vary substantive details about the policy along with group endorsement. For example, Nicholson (2011) uses this design to show that a group's endorsement of a policy is overwhelmed by information about the social groups who are helped or harmed by the policy.

Priming experiments still suffer from several problems. First, the treatment might not prime the intended concept. Unlike list experiments or randomized response questions, the item of interest is not directly enumerated to the respondent. It is possible that the treatment activates different attitudes than the researcher intends. Second, the mental construct being primed may already be salient in the minds of all respondents (i.e. pretreated), rendering the prime impotent.[20] Unfortunately, methods to validate the estimates of priming experiments do not exist yet.

Summary

Scholars want to measure concepts validly and reliably. Direct survey questions are often used to measure concepts, but direct questions fail when respondents lie or do not have conscious access to the attitude the researcher is interested in. To address the reasons that direct questions fail, scholars began measuring some concepts with survey experiments in lieu of direct questions. As scholars learned more about measuring concepts with survey experiments, they learned the pitfalls of survey experiments for measurement and further adapted their measures to account for those pitfalls. Thus, the history of measuring sensitive concepts has been to increase complexity of design for the purpose of increasing validity.

One problem plagues all measurement: how do we know that our measure is valid? For some outcomes, such as voter turnout, we can compare our measure with population estimates. But for other outcomes, such as racism or the effect that political parties have on citizens' policy preferences, no population estimate with which to validate our measure exists. We are searching for a truth we cannot know.

Despite an inability to validate most measures, we make progress by fully explicating the assumptions of each measurement strategy and then determining if the strategy fulfills them. When measurement strategies do not satisfy their own assumptions, researchers must create measures that do. For example, researchers assume anonymity is the key to obtaining honest answers about sensitive topics. Yu et al. (2008) noticed that randomized response techniques fail to satisfy this assumption and created the crosswise model to provide respondents with full anonymity.

Part of explicating assumptions is explicating the reasons as to why direct questions will not validly measure a concept of interest. Sometimes scholars will find that people are not as squeamish as scholars expect, and thus direct questions work well. Other times the techniques that provide anonymity do not reveal the attitude of interest because the barrier to measurement is respondents lacking conscious access to that attitude. We think it is important that the theoretical assumptions about the concept we are measuring match the assumptions of our measurement strategy.

CONCLUSION

Survey experimental research has matured quickly and dramatically. In both areas of endeavor that we reviewed, researchers appear to be functioning in the best tradition of 'normal science' (Kuhn, 1962), which we applaud. The growing complexity and sophistication in research designs has led to new discoveries, and, subsequently, more challenges, which in turn have generated even more creative and complex designs. The more we learn, it seems, the more complex and sophisticated the next generation of research must be to address the newest, and often unexpected, discoveries. No serious researcher will be surprised to know that a final stopping point is nowhere to be seen.

Increased complexity implies more moving parts, and the more moving parts there are, the less transparent the methods and empirics can become. Recent work on conjoint analysis exemplifies this statement (Hainmueller et al., 2014). The logic of leveraging pair comparisons to isolate independent effects across manipulations fringes on impeccable, and as a result, the potential to remedy some existing challenges is, at least in theory, high. Yet, knowing precisely how respondents respond to and interpret many descriptive pairs remains slightly difficult to ascertain. Likewise, recent measurement advances like the crosswise model (Yu et al., 2008) appear to offer anonymity in a way that is convincing to respondents. Yet, scholars have no way of assessing the validity of this approach on most sensitive attitudes.

We are reminded of the dictum of 'no more than three independent variables', as applied it to the growing complexity of probit and logit models (Achen, 2002). We see no reason to apply it to survey experimental research at this time, although keeping it close at hand as the two areas progress would be wise. Increasing the complexity and sophistication of survey experimental designs has thus far evolved in a logical, progressive, and helpful way. Whether a lack of transparency will begin to obfuscate the power of the designs remains to be seen.

Overall, reviewing the evolution of survey experimental research has been, for us, an eye-opener. Although we have remained conversant with the literature as it has grown, peering into the bowels of the research has increased our understanding of the challenges that await new generations of researchers. It has also increased what was our already-high respect for the work.

Notes

1 We do not address external validity or inferring from an experiment to some (often undefined) world outside it. Suffice it to say that we agree with the contention that the primary purpose of an experiment should be to test theory, not to infer to some outside world (Mook, 1983). However, for the purposes of this chapter, we take the general practice of inferring from experiment to world as given.

2 The speed with which researchers can respond to each other underlines the crucial importance of publishing the results of well-designed survey experiments that generate null findings.

3 Note that this distinction is primarily a working and admittedly artificial one. The study of causal inference requires good measurement and the study of measurement cannot occur without causal inference. Our distinction echoes what is current practice in the field.

4 These problems apply to causal inference studies as well. We bring them up in the discussion of measurement because it is there that the potential consequences of survey contamination are most obvious.

5 Depending on the question of interest, the control group may receive no treatment at all, or an innocuous version of the treatment (placebo). Some survey experiments include both pure control and placebo conditions. The conventional advice is to include some form of control group to identify the direction of the treatment effect (Gaines et al., 2007).

6 Note that survey experiments are not the only way to study corruption. Occasionally, researchers find direct measures in observational data (Fernández-Vázquez et al., 2016; Ferraz and Finan, 2008; Golden and Picci, 2005; Olken, 2007).

7 This example is certainly not the only case of researchers using multiple factors in survey experiments. In the study of American politics, the practice is traceable at least back to the early studies on racial prejudice (Sniderman et al., 1991) and heuristic processing (Mondak, 1993).

8 The standard practice is to prevent these combinations from appearing, but that might introduce bias in the average marginal component effect. An alternative is to allow for illogical combinations with a relatively low probability (Hainmueller et al., 2014)

9 Although previous work provides guidelines for the number of factors that should be included (Auspurg and Hinz, 2015; Hainmueller et al., 2014), we are not aware of any developments regarding which factors to include.

10 Chong and Druckman (2010, 2013) pioneered dynamic studies. Even today, few have followed their footsteps.

11 In their original study, Dafoe et al. (2018) use the term information equivalence. We use lack of information equivalence because a lack of equivalence is the problem they address.

12 Other instances of response bias are due to demand effects (when the respondent learns what the researcher wants to hear and obliges) and acquiescence bias (the tendency of respondents to agree rather than disagree with statements). Survey experiments help overcome these biases by hiding the intent of the researcher. Other reasons for respondent's lying are when the respondent deliberately provides untrue responses for fun, lack of attention, and desire to get through the survey quickly. Most survey experimental techniques will not overcome these issues.

13 While many people do not consider randomized response models as experiments, we view them as experiments in which the researcher does not know the experimental condition of the respondent. Results can still be analyzed in an 'experimental' fashion (and often compared to results from other experiments to reduce bias) because the data-generating process is still known. The same logic applies to non-randomized response models, the close cousins of randomized response models.

14 This is known as the 'forced response' model, introduced by Fox and Tracy (1986). Other models use slightly different procedures.

15 A full discussion of all recent advances in list experiments and randomized response techniques is beyond our scope. Here we focus on design advances and omit work on statistical analysis of survey experiments and the comparison of responses to direct questions (Ahlquist, 2018; Aronow et al., 2015; Blair and Imai, 2012; Blair et al., 2015; Blair et al., 2018; Chou et al., 2017; Corstange, 2009; Rosenfeld et al., 2016).

16 Technically, the crosswise model is a non-randomized response model. Non-randomized response models pair a sensitive question with some nonrandom phenomenon, instead of with random phenomenon like a coin flip. The sensitive question and the non-random phenomenon are paired in such a way that the researcher cannot know if the respondent agrees with the sensitive question or the non-random phenomenon.

17 Lensvelt-Mulders et al. (2005a), Azfar and Murrell (2009), and Gingerich (2015) report other randomized response techniques and advances.

18 Other techniques are also used, such as implicit associations tests (Greenwald et al., 1998) and physiological measures (Rankin and Campbell, 1955). Survey experiments enjoy one major

advantage: they can easily be administered to respondents outside of the laboratory.

19 Endorsement experiments have recently been used to measure explicit attitudes towards groups that may be dangerous to support publicly. Rather than measure implicit attitudes, these 'explicit' endorsement experiments work like list experiments, where individuals can freely express their support for the sensitive group because the researcher cannot differentiate policy support from group support at an individual level. Whereas list experiments hide the respondent's opinion by pairing the sensitive item with non-sensitive control items, endorsement experiments hide the respondent's opinion by pairing the sensitive item with a policy (e.g. Blair, 2015).

20 The outcome being measured could also activate the prime, which would accidentally treat the control group. This happens when the outcome is too close mentally to the prime. For example, the term 'welfare' may make racial minorities salient in the minds of white respondents.

REFERENCES

Achen, Christopher H. 2002. Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5(1): 423–50.

Ahlquist, John S. 2018. List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators. *Political Analysis* 26(1): 34–53.

Anduiza, Eva, Aina Gallego, and Jordi Muñoz. 2013. Turning a Blind Eye: Experimental Evidence of Partisan Bias in Attitudes Towards Corruption. *Comparative Political Studies* 46(12): 1664–92.

Aronow, Peter M., Alexander Coppock, Forrest W. Crawford, and Donald P. Green. 2015. Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence. *Journal of Survey Statistics and Methodology* 3(1): 43–66.

Auspurg, Katrin, and Thomas Hinz. 2015. *Factorial Survey Experiments*. London: Sage.

Azfar, Omar, and Peter Murrell. 2009. Identifying Reticent Respondents: Assessing the Quality of Survey Data on Corruption and Values. *Economic Development and Cultural Change* 57(2): 387–411.

Blair, Graeme. 2015. Survey Methods for Sensitive Topics. *Comparative Politics Newsletter* 12: 44.

Blair, Graeme, C. Christine Fair, Neil Malhotra, and Jacob N. Shapiro. 2013. Poverty and Support for Militant Politics: Evidence from Pakistan. *American Journal of Political Science* 57(1): 30–48.

Blair, Graeme, and Kosuke Imai. 2012. Statistical Analysis of List Experiments. *Political Analysis* 20(1): 47–77.

Blair, Graeme, Kosuke Imai, and Yang-Yang Zhou. 2015. Design and Analysis of the Randomized Response Technique. *Journal of the American Statistical Association* 110(511): 1304–19.

Blair, Graeme, Winston Chou, and Kosuke Imai. 2018. List Experiments with Measurement Error. *Political Analysis*, 1–26.

Boruch, Robert F. 1971. Assuring Confidentiality of Responses in Social Research: A Note on Strategies. *The American Sociologist*, 308–311.

Botero, Sandra, Rodrigo Castro Cornejo, Laura Gamboa, Nara Pavao, and David W. Nickerson. 2015. Says Who? An Experiment on Allegations of Corruption and Credibility of Sources. *Political Research Quarterly* 68(3): 493–504.

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. The Design of Field Experiments With Survey Outcomes: A Framework for Selecting More Efficient, Robust, and Ethical Designs. *Political Analysis* 25(4): 435–64.

Campbell, Donald T. 1988. Can We Be Scientific in Applied Social Science? In Donald T. Campbell and E. Samuel Overman (eds) *Methodology and Epistemology for Social Science: Selected Papers*, 315–33. Chicago: University of Chicago Press.

Chong, Dennis, and James N. Druckman. 2010. Dynamic Public Opinion: Communication Effects over Time. *American Political Science Review* 104(4): 663–80.

Chong, Dennis, and James N. Druckman. 2013. Counterframing Effects. *The Journal of Politics* 75(1): 1–16.

Chou, Winston, Kosuke Imai, and Bryn Rosenfeld. 2017. Sensitive Survey Questions with Auxiliary Information. *Sociological Methods & Research*. Available at https://doi.org/10.1177/0049124117729711 (Accessed 20 January 2020).

Cohen, Geoffrey L. 2003. Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs. *Journal of Personality and Social Psychology* 85(5): 808–22.

Corstange, Daniel. 2009. Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis* 17(1): 45–63.

Cox, D. R. 1958. *Planning of Experiments*. Wiley Classics Library. New York: Wiley.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. Information Equivalence in Survey Experiments. *Political Analysis* 26(4): 399–416.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsley, Wendy Visscher, and Trena M. Ezzati. 2004. The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application. In Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (eds) *Measurement Errors in Surveys*, 185–210. Hoboken: Wiley.

Druckman, James N., and Thomas J. Leeper. 2012. Learning More from Political Communication Experiments: Pretreatment and Its Effects. *American Journal of Political Science* 56(4): 875–896.

Edgell, Stephen E., Samuel Himmelfarb, and Karen L. Duchan. 1982. Validity of Forced Responses in a Randomized Response Model. *Sociological Methods & Research* 11(1): 89–100.

Eggers, Andrew C., Nick Vivyan, and Markus Wagner. 2018. Corruption, Accountability, and Gender: Do Female Politicians Face Higher Standards in Public Life? *The Journal of Politics* 80(1): 321–26.

Fernández-Vázquez, Pablo, Pablo Barberá, and Gonzalo Rivero. 2016. Rooting Out Corruption or Rooting for Corruption? The Heterogeneous Electoral Consequences of Scandals. *Political Science Research and Methods* 4(2): 379–97.

Ferraz, Claudio, and Federico Finan. 2008. Exposing Corrupt Politicians: The Effects of Brazil's Public Released Audits on Electoral Outcomes. *Quarterly Journal of Economics* 123(2): 703–45.

Findley, Michael G., Brock Laney, Daniel L. Nielson, and J. C. Sharman. 2017. External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation. *The Journal of Politics* 79(3): 856–72.

Fox, James, and Paul Tracy. 1986. *Randomized Response*. London: Sage.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. The Logic of the Survey Experiment Reexamined. *Political Analysis* 15(1): 1–20.

Gingerich, Daniel. 2015. Randomized Response: Foundations and New Developments. *Newsletter of the Comparative Politics Organized Section of the American Political Science Association* (The Organized Section in Comparative Politics of the American Political Science Association) 25(1): 16–27.

Glynn, Adam N. 2013. What Can We Learn with Statistical Truth Serum? *Public Opinion Quarterly* 77(S1): 159–72.

Golden, Miriam A., and Lucio Picci. 2005. Proposal for a New Measure of Corruption, Illustrated with Italian Data. *Economics and Politics* 17(1): 37–75.

Gonzalez-Ocantos, Ezequiel, Chad Kiewiet De Jonge, Carlos Meléndez, Javier Osorio, and David W Nickerson. 2012. Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American Journal of Political Science* 56(1): 202–217.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74(6): 1464–80.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review* 102(1): 4.

Hainmueller, Jens, and Daniel J. Hopkins. 2015. The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants. *American Journal of Political Science* 59(3): 529–48.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis* 22(1): 1–30.

Hurwitz, Jon, and Mark Peffley. 1997. Public Perceptions of Race and Crime: The Role of

Racial Stereotypes. *American Journal of Political Science* 41(2): 375–401.
Jann, Ben, Julia Jerke, and Ivar Krumpal. 2011. Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism. *Public Opinion Quarterly* 76(1): 32–49.
Jones, Edward E., and Harold Sigall. 1971. The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin* 76(5): 349–64.
King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review* 98(1): 191–207.
King, Gary, Robert O. Keohane, and Sidney Verba. 1995. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
Leary, Mark R., and Robin M. Kowalski. 1990. Impression Management: A Literature Review and Two-Component Model. *Psychological Bulletin* 107(1): 34.
Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, and Peter G. M. Van Der Heijden. 2005a. How to Improve the Efficiency of Randomised Response Designs. *Quality and Quantity* 39(3): 253–265.
Lensvelt-Mulders, Gerty J. L. M., Joop J. Hox, Peter G. M. Van der Heijden, and Cora JM Maas. 2005b. Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation. *Sociological Methods & Research* 33(3): 319–348.
Lupia, Arthur. 1994. Shortcuts versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections. *American Political Science Review* 88(1): 63–76.
Macrae, C. Neil, Galen V. Bodenhausen, Alan B. Milne, and Jolanda Jetten. 1994. Out of Mind but Back in Sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology* 67(5): 808.
Mares, Isabela, and Giancarlo Visconti. 2019. Voting for the Lesser Evil: Evidence from a Conjoint Experiment in Romania. *Political Science Research and Methods*, 1–14. Available at https://doi.org/10.1017/psrm.2019.12 (Accessed 20 January 2020)
McFarland, Sam G. 1981. Effects of Question Order on Survey Responses'. *Public Opinion Quarterly* 45(2): 208.
Mondak, Jeffery J. 1993. Source Cues and Policy Approval: The Cognitive Dynamics of Public Support for the Reagan Agenda. *American Journal of Political Science* 37(1): 186.
Mook, Douglas G. 1983. In Defense of External Invalidity. *American Psychologist* 38(4): 379.
Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.
Nicholson, Stephen P. 2011. Dominating Cues and the Limits of Elite Influence. *The Journal of Politics* 73(4): 1165–77.
Nisbett, Richard E., and Timothy D. Wilson. 1977. Telling More than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84(3): 231.
Olken, Benjamin A. 2007. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy* 115(2): 200–249.
Rankin, Robert E., and Donald T. Campbell. 1955. Galvanic Skin Response to Negro and White Experimenters. *The Journal of Abnormal and Social Psychology* 51(1): 30.
Riambau, Guillem, and Kai Ostwald. 2019. Placebo Statements in List Experiments: Evidence from a Face-to-Face Survey in Singapore. Working Paper. http://guillemriambau.com/Placebo%20Statements%20in%20List%20Experiments.pdf (Accessed 20 January 2020).
Rosenbaum, Paul R. 1999. Choice as an Alternative to Control in Observational Studies. *Statistical Science* 14(3): 259–304.
Rosenfeld, Bryn, Kosuke Imai, and Jacob N. Shapiro. 2016. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science* 60(3): 783–802.
Russett, Bruce M. 1993. *Grasping the Democratic Peace: Principles for a Post-Cold War World*. Princeton: Princeton University Press.
Schwarz, Norbert, and Gerald L. Clore. 1983. Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States. *Journal of Personality and Social Psychology* 45(3): 513.

Schwarz, Norbert, and Herbert Bless. 1991. Constructing Reality and Its Alternatives: An Inclusion/Exclusion Model of Assimilation and Contrast Effects in Social Judgment.. In Leonard L. Martin and Abrahan Tesser (eds) *The Construction of Social Judgments*, 217–245, New Jersey: Lawrence Erlbaum Associates..

Sniderman, Paul M. 2018. Some Advances in the Design of Survey Experiments. *Annual Review of Political Science* 21(1): 259–75.

Sniderman, Paul M., Thomas Piazza, Philip E. Tetlock, and Ann Kendrick. 1991. The New Racism. *American Journal of Political Science* 35(2): 423.

Tomz, Michael R., and Jessica L. P. Weeks. 2013. Public Opinion and the Democratic Peace. *American Political Science Review* 107(4): 849–65.

Warner, Stanley L. 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association* 60(309): 63–69.

Weitz-Shapiro, Rebecca, and Matthew S. Winters. 2017. Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil. *Journal of Politics* 79(1): 60–74.

Winters, Matthew S., and Rebecca Weitz-Shapiro. 2013. Lacking Information or Condoning Corruption: When Do Voters Support Corrupt Politicians? *Comparative Politics* 45(4): 418–36.

Yu, Jun-Wu, Guo-Liang Tian, and Man-Lai Tang. 2008. Two New Models for Survey Sampling with Sensitive Characteristic: Design and Analysis. *Metrika* 67(3): 251.

Zigerell, Lawrence J. 2011. You Wouldn't like Me When I'm Angry: List Experiment Misreporting. *Social Science Quarterly* 92(2): 552–562.