Advance Access publication November 16, 2006

Political Analysis (2007) 15:1–20
doi:10.1093/pan/mpl008

The Logic of the Survey Experiment Reexamined

Brian J. Gaines and James H. Kuklinski
Department of Political Science and Institute of Government and Public Affairs,
University of Illinois at Urbana-Champaign,
702 South Wright Street, Urbana, IL 61801
e-mail: bjgaines@uiuc.edu (corresponding author)
e-mail: kuklinsk@ad.uiuc.edu

Paul J. Quirk
Department of Political Science, University of British Columbia,
2329 West Mall, Vancouver, British Columbia, Canada, V6T 1Z4
e-mail: quirk@politics.ubc.ca

Scholars of political behavior increasingly embed experimental designs in opinion surveys by randomly assigning respondents alternative versions of questionnaire items. Such experiments have major advantages: they are simple to implement and they dodge some of the difficulties of making inferences from conventional survey data. But survey experiments are no panacea. We identify problems of inference associated with typical uses of survey experiments in political science and highlight a range of difficulties, some of which have straightforward solutions within the survey-experimental approach and some of which can be dealt with only by exercising greater caution in interpreting findings and bringing to bear alternative strategies of research.

1 Introduction

Most of what we know about public opinion comes from the statistical analysis of cross-sectional survey data and, to a lesser extent, panel survey data. For over half a century, scholars have used these data to explain a wide range of phenomena, including policy preferences, economic assessments, candidate evaluations, and voting decisions, among others. The level of statistical sophistication has increased dramatically since the early days of survey research, but the basic methodological approach has changed little.

Many perils attend efforts to infer causal relationships from cross-sectional survey data, as statisticians and social science methodologists continue to document. Specific

Authors' note: This paper was originally presented at the Annual Meeting of the Midwest Political Science Association, April 2004. The commentators on that panel—Darren Davis, Donald Green, and Diana Mutz—made invaluable comments. We received helpful suggestions during presentations at Columbia University, Purdue University, and Northwestern University. We thank Thomas Rudolph for reading and commenting on an earlier version of the paper and Jamie Druckman for his encouragement from beginning to end. Psychologist Norbert Schwarz, one of the leaders of the survey experiment movement, offered invaluable insights. Robert Erikson and three anonymous reviewers gave useful advice on how to revise the original paper. Our greatest debt is to Paul Sniderman, who, more than any other single individual in political science, brought survey experiments into the mainstream. He will not agree with every argument presented here, but he has supported this project from its infancy.

challenges include selection bias, spurious correlation, correlated measurement errors, censored data, the lack of true counterfactuals, and mutual causation (see, among others, Achen 1986; Fearon 1991; King, Keohane, and Verba 1994; Smith 1999; Brady and Collier 2000). Efforts to solve such problems may merely substitute other problems, such as overcontrolling (Lieberson 1987).

As a result, statistical analyses of cross-sectional survey data are notoriously subject to misleading findings. For instance, economic research comparing cross-sectional survey findings on the effectiveness of government programs with evidence from field experiments has found the survey estimates biased. Various techniques, such as matching, propensity scores, and two-stage least squares, can reduce the biases, but they cannot eliminate them (Rosenbaum and Rubin 1983; LaLonde 1986; Heckman, Ichimura, and Todd 1997, 1998; Heckman, Ichimura, Smith, and Todd 1998). They also require data that rarely exist in survey research.

Panel surveys, which collect data from the same individuals in multiple waves, facilitate across-time analysis to determine whether changes in one variable lead to changes in another. The ability to observe individual-level changes frees researchers from having to treat differences across cases as proxies for overtime differences within cases. However, panel surveys are expensive and difficult to implement, and panel data share many of the limitations of cross-sections. They also pose some unique concerns, including the danger that repeatedly interviewing the same individuals induces changes in their attitudes and behavior.

In light of the travails of conventional survey research, the advent of the survey experiment—introduced to political science largely through the efforts of Paul Sniderman and his colleagues at the Survey Research Center at the University of California at Berkeley—has been good news for students of public opinion and political psychology. Often taking advantage of computer-assisted telephone interviewing (CATI), researchers assign respondents randomly to control and treatment conditions, actively manipulating a treatment. The survey experiment is easy to implement and avoids many problems associated with cross-sectional and panel survey data. It clearly distinguishes cause and effect. When used with representative samples, therefore, survey experiments can provide firmly grounded inferences about real-world political attitudes and behavior.

Having already made substantial contributions to political science, survey experiments will be central to the next generation of public opinion research. To realize their full potential, however, scholars must recognize some limitations of current practices. Crucial to this assessment, political scientists, unlike psychologists, do not study mental processes for their own sake. *Political scientists use survey experiments to identify how citizens make decisions and respond to real-world political objects, in order to enhance understanding of politics*. This focus imposes special demands on findings. For example, an extremely short-lived effect of an experimental treatment, although perhaps of theoretical importance for psychologists, would not often be significant for voting or public opinion.

This paper proceeds in three sections. The first presents an overview of the survey experiment. We revisit how it was developed, briefly identify its defining characteristics, and summarize the varieties of survey experiment and their methodological and substantive contributions to the discipline. The second section argues that certain common practices have limited the scope and importance of survey experiment findings. Political scientists only rarely measure the endurance of "treatment effects"; in those exceptional cases, they have discovered that the effects do not last. Moreover, they almost never repeat treatments over time, even though the real-world political phenomena of interest occur repeatedly. They also do not, generally, consider the possibility of mutual cause and effect.

Surprisingly many studies lack control groups, rendering their results ambiguous. Fortunately, researchers can change these practices.

The third section argues that random assignment alone does not prevent contamination from prior effects. Problematic prior effects can arise from at least two sources: other experiments earlier in the survey and conditions in the real world, in particular, the very conditions the experimental treatment is supposed to represent. Real-life events are most likely to contaminate experimental results when they matter politically, that is, when the causal effects endure. Since researchers have little control over the real world they seek to understand, they must be explicit about how ordinary life intrudes on experimental simulations of it. The conclusion offers recommendations to improve future survey experimentation.

2 The Survey Experiment

Students of public opinion now employ survey experiments as a matter of course, but this was not always true. In part, technological advances turned the sometimes difficult task of manipulating survey items into an easy one. But technology explains only why the survey experiment became more feasible. Two annoying problems in survey research—question-ordering effects and question-wording effects—motivated its adoption.

2.1 *From Vice to Virtue: Taking Words and Orders Seriously*

Prior to the survey experiment reaching its current status as a methodology to study cause and effect, survey researchers had sometimes used split-ballot designs in which they changed either the question ordering or the question wording. In a classic study, American respondents were more likely to say that the United States should admit newspaper reporters from Communist countries when that question was preceded by one about Communist countries admitting American reporters than when it was not (Hyman and Sheatsley 1950). At first, scholars viewed such reversals as nuisances: if changing the ordering or wording of questions changed responses, and if no ordering or wording of questions is more correct than another, then how could a researcher take any finding seriously?

With CATI facilitating the manipulation of question orderings and question wordings in survey instruments, scholars began to identify numerous reversals and, more importantly, discovered an upside. Some political scientists argued that such effects were evidence that people lack true political attitudes on most issues or, less radically, that such attitudes depend heavily on context (Zaller 1992; Lacy 2001). The variability that had been construed as a methodological embarrassment was transformed into an important substantive finding.

Survey experimenters developed another positive interpretation of question-ordering and question-wording effects: as a methodological opportunity to demonstrate real-world cause and effect. By comparing responses to manipulated questions, the argument runs, a researcher can identify causal relationships that exist in the real world. If mentioning affirmative action increases stereotyping in the context of a survey, then real-world discussions of affirmative action programs presumably do the same. With this simple and subtle change in interpretation, scholars transformed a vice into a virtue.

2.2 *Survey Experiments in Political Science*

A survey experiment, then, is nothing more than a deliberate manipulation of the form or placement of items in a survey instrument, for purposes of inferring how public opinion

works in the real world. The word "experiment" also implies random assignment of respondents to control and treatment conditions. Comparing the decisions, judgments, or behaviors of the respondents in the treatment group to those in the control group reveals the causal effects under investigation.[1]

This description of the defining characteristics of survey experiments belies the variety of purposes for which researchers have undertaken them and the kinds of treatments they have employed. Sometimes, scholars conduct survey experiments for methodological purposes. Clarke et al. (1999) randomly assigned respondents to one of two conditions. In the control condition, they used the same values battery that Inglehart (1990) has long used to measure postmaterialism. In the treatment condition, they substituted an unemployment item for Inglehart's inflation item. Interviewing respondents in two countries that were experiencing high unemployment, they found that far fewer people in their treatment (e.g., unemployment) condition than in the control (e.g., inflation) condition met the criteria for postmaterialism. The finding thus raised questions about the validity of Inglehart's postmaterialism measure.

Others have adopted survey experiments to reduce social desirability effects when asking about sensitive social topics like race and sexual preference. In their simplest form, these experiments entail randomly assigning respondents to different target groups—women and African Americans, for example—when asking about policies like affirmative action (Kinder and Sanders 1996; Hurwitz and Peffley 1997; Peffley, Hurwitz, and Sniderman 1997; Gilens 1999; Davis and Silver 2003). This approach allows the researcher to determine whether support of such policies depends on their targets. Other uses of survey experiments to reduce social desirability effects have been more complicated (Kuklinski, Cobb, and Gilens 1997).

Most survey experiments address substantive rather than methodological topics. Some prime a particular thought or idea to determine how (or whether) the priming affects an opinion or attitude. A pioneering example is the experiment of Sniderman and Piazza (1993) on racial attitudes.[2] In the authors' words (102),

> We devised ... the mere mention experiment ... to simulate the kinds of conversations that ordinary people undoubtedly have.... The basic idea is to take advantage of the power of randomization to determine whether references to affirmative action can, in and of themselves, excite negative reactions to blacks.... A sample of a cross section of whites is randomly divided into two halves. One half is asked their view of affirmative action, then their images of blacks. The other half is asked exactly the same questions, except in the opposite order. If a dislike of affirmative action provokes a dislike of blacks, then ... [those] asked first about affirmative action should dislike blacks more than the other[s] ... if the two halves are observed to differ in this way, the reason must necessarily be that the *mere mention* of affirmative action encourages dislikes of blacks—necessarily so, since the two halves of the sample, being randomly composed, are alike in all respects, chance variations aside.

The authors find that the "mere mention" of affirmative action indeed increases negative stereotyping of black Americans among whites. Other survey experiments provide more or less information (Gilens 2001), evoke one motivation or another (Taber and Lodge 2006), or ask respondents either to think seriously before answering a survey question or to react to it viscerally (Kuklinski et al. 1991). Gibson (1998) builds on

[1] Randomized variation is also sometimes used in survey items to create interesting variance. For instance, average price sensitivity can be measured by testing willingness to pay with an item wherein the price presented to each respondent is randomly selected from some known distribution. There does not appear to be consensus on whether or not such items should be regarded as experiments.

[2] It is not the first notable example. That distinction belongs to sociologists Schuman and Bobo (1988). Sniderman (1996) presents an excellent overview of the survey experiment.

Sniderman's early survey experiments to determine whether respondents who gave one or another answer to a survey question can be convinced to change their minds.

Some survey experiments have included more than one randomized element. The "welfare mother experiment" (Sniderman and Carmines 1997, 67–70) manipulated both race (black versus white) and educational achievement (high school graduate versus high school dropout) in a description. The interaction turned out to be crucial: respondents' predictions about the mother's success were most optimistic when she was black and a high school graduate and least optimistic when she was white and a high school dropout.

The emergence of Internet surveys and Time-Sharing Experiments for Social Sciences (TESS) promises to maintain if not increase the use of survey experiments in public opinion research. Indeed, because surveys administered over the Internet can incorporate sounds, video, photographs, and elaborate graphics, the distinction between laboratory experiments and survey experiments grows ever fainter. A decade ago, Ansolabehere and Iyengar (1995) characterized their experiments on the effects of political advertising as "the alternative to the sample survey" (19) and boasted that "our own studies were designed to overcome the limited generalizability of the experimental method ... [because] participants represented a fair cross-section of the electorate, the experimental setting was casual and designed to emulate 'real life', and our studies all took place during ongoing political campaigns" (20). At this stage, the essential features of their research design could be replicated in an online survey. Indeed, respondents could then view political advertisements in their homes, rather than in settings intended to resemble home environments.

3 Practices That Could Produce Misleading Inferences

The survey experiment's resume arguably exceeds that of any other new method in public opinion and political psychology since the development of the sample survey. The method transformed the vexing problems of question-ordering and question-wording effects into a simple approach for determining cause and effect, and it has produced convincing findings on a wide range of substantive topics. In their enthusiasm to use the survey experiment, however, scholars have not always carefully considered whether standard practices might yield misleading inferences. With an eye to improving the next generation of survey experiments, we identify four practices that survey experimenters might want to reexamine. The first two are closely associated with using single-shot, cross-sectional surveys as the primary vehicle for these experiments. It is not altogether clear whether the use of single-shot survey experiments encourages problematic habits or, rather, a lack of concern for the consequences of these practices encourages the use of cross-sectional survey experiments. Ultimately, it does not matter; the limitations exist.

3.1 *Not Measuring the Durations of Effects*

Nearly all survey experiments are embedded in cross-sectional surveys. If those in the treatment group differ, on average, from those in the control group, the researcher normally concludes that the treatment works in a politically significant way in the real world. If the mere mention of affirmative action increases stereotyping in the one-shot survey experiment, the logic goes, then so it does in the real world.

Without knowing the duration of the effects, however, users of survey experiments cannot determine the relevance of their findings for politics. Suppose the effect no longer persists 10 min after treatment. In what real-world scenario would such an effect matter?

Even a voter who hears a campaign commercial on his car radio as he parks at the polling place might not cast his vote within that length of time. The implications of survey-experimental results for politics depend crucially on how long the effects last, with relevant periods measured in weeks, or months, not minutes.

Some effects generated in survey experiments might last a long time, others not. Many years ago, Hovland and Weiss (1951) found that messages from highly credible sources had a larger immediate effect than those from less credible sources but that the difference dissipated over time (as has, evidently, scholars' memories of Hovland's finding). Hovland explained this phenomenon, which he called the sleeper effect, by distinguishing between accepting a message and learning it: credible sources increased short-term acceptance but did not enhance long-term learning.

Recent evidence from political science echoes this result. Luskin, Fishkin, and Jowell (2002) found that participating in structured discussions changes many people's initial policy preferences. But these changes do not endure; participants soon return to their initial positions, despite permanent changes in their levels of knowledge.[3] Druckman and Nelson (2003) report that the initial elite framing effects they found in their experiments had dissipated within 10 days. Similarly, in a series of creative experiments, Mutz and Reeves (2005) found that exposing subjects to uncivil political debates on television immediately reduced their levels of trust, but once again, the effects did not last: "By the time of the follow-up interview (approximately three weeks), there were no significant differences by original experimental condition" (12).[4]

In all three cases, the authors reported the lack of enduring effects as an aside. Suppose, instead, that they had included the words "transitory effects" in their original titles. Would readers take away different conclusions about the impacts of deliberations, elite framing, and televised uncivil behavior on political attitudes? If so, then survey experimenters would be wise, first, to focus more directly on the durations of effects and, second, to consider the implications for politics of effects that diminish at different rates.

Indeed, determining the rates of decay of various treatment effects and deriving the political implications could be one of the most informative tasks that users of survey experiments undertake in the future. Suppose, for example, that one frame's effects last longer than another's. This might imply that one frame will more widely and substantially affect people's attitudes than the other. But it also might not do so, depending on when politicians offer their statements, such as at the end or at the beginning of a campaign. Currently, political scientists cannot say much about such matters, even thought these are the sorts of dynamics that determine who fails and who succeeds in politics. The lack of permanent effects, in other words, might be less a nuisance than an important datum about the give-and-take of politics.

3.2 *Presenting One-Shot Treatments*

Most real-world political stimuli that survey experimenters attempt to replicate do not occur in a single moment. If the phenomena occurred only once, researchers likely would view them as irrelevant to real-world politics. Yet, cross-sectional survey experiments preclude giving meaningful multiple treatments over time.

[3]Robert Luskin personally communicated this conclusion.

[4]Druckman and Nelson (2003) and Mutz and Reeves (2005) conducted their studies well after we had presented the paper on which this article is based. Whether that paper encouraged them to measure the identified effects' endurances we cannot say. We would like to think it did. Also see Lodge, Steenbergen, and Brau (1995).

The researchers often make up for the lack of multiple treatments by using highly obtrusive treatments. The survey provides a captive audience, attending to more or less everything that the interviewer says; it generally uses exceedingly overt manipulations, for example, frames that the respondent cannot possibly miss; and it often presents the dependent measure immediately after the treatment—typically in the very next question. In effect, researchers assume that a single exposure to a strong treatment in a survey is roughly equivalent to frequent exposure to a weaker stimulus in the real world.

The rationale for using one-shot, obtrusive treatments is well illustrated with an analogy to some research on chemical hazards. Sometimes industries expose people to a chemical. Regulatory officials want to know if this exposure causes cancer, even in small numbers of people, after many years of exposure. Scientists cannot give people the chemical, study huge numbers of animals, or extend a study for many years. Using laboratory experiments, they instead give huge amounts of the chemical—perhaps 1000 times the normal—to a small number of animals for a short time. If they find positive results, they infer that low levels of the chemical cause cancer in small numbers of humans. Such inferences are always vulnerable to the charge of unrealistic treatments.

Scientists studying the effects of chemical exposures face some inherent ethical constraints that survey experimenters do not. The main constraints on survey researchers are those they impose on themselves. Conducting longitudinal studies with multiple treatments is a natural and much-needed next step.

3.3 *Overlooking Mutual Causation*

Political scientists who analyze nonexperimental survey data statistically have become increasingly aware that any two factors of interest might cause each other. Indeed, the development of methods to account for so-called endogeneity has become a cottage industry among social science methodologists. In contrast, experimental logic, especially within the context of one-shot studies, requires the designation of independent and dependent variables. And there is the rub: what is entirely proper from the standpoint of a survey experiment might not capture the real-world causal complexity.

But, some might argue, researchers undertake a survey experiment (or any other type of experiment, for that matter) only when they know that causation goes in one direction and not the other. We doubt that theory is often such a strong guide in this regard. For example, Sniderman and Piazza reasonably assume that the mention of affirmative action increases stereotyping. However, encouraging whites to think about negative stereotypes of black people conceivably could shape their attitudes toward affirmative action as well.

Sometimes survey experiments can shed light on causal direction, even within a one-shot study. Mere mention is a case in point. Sniderman and Piazza (1993) embedded their experiment in a small survey of Kentucky residents (Sniderman and Piazza 1993, 102–4, 182; Sniderman and Carmines 1997, 39). Despite the experiment's nickname, the researchers actually measured support for affirmative action. Respondents were asked two questions, randomly ordered.

> In a nearby state, an effort is being made to increase dramatically the number of blacks working in state government. This means that a large number of jobs will be reserved for blacks, even if their scores on merit exams are lower than those of whites who are turned down for the job. Do you favor or oppose this policy?

> Now I'm going to read a few statements that are sometimes used to describe blacks. Of course, no statement is true about everybody, but speaking generally please say whether you strongly agree, somewhat agree, somewhat disagree, or strongly disagree with each description. How about:

> Blacks have a tendency to be arrogant.
> (Blacks tend to be) family-oriented/lazy/intelligent/irresponsible/pleasure-loving/hard-working/friendly/violent/self-disciplined.

Those who were asked about affirmative action first and stereotypes second endorsed (at least) three of the negative traits—"lazy," "irresponsible," and "arrogant"—at higher rates than those who received the reverse order, with the differences being statistically significant at the 0.05 level in two cases and at the 0.10 level in the third (Sniderman and Piazza 1993, 104).

The authors do not present information about support levels for the affirmative action program under either ordering.[5] Because the cue for respondents to think about affirmative action was a question and not merely a statement, however, the experiment contains this information. Sniderman and Piazza might have used this additional information to tease out more fully the nature of the cause and effect or at least to buttress their conclusion that the mention of affirmative action increases white people's negative stereotyping of black people.

Suppose, for purposes of discussion, that Sniderman and Piazza had asked only about negative (and not positive) stereotypes. Suppose, furthermore, that self-declared opponents of affirmative action (in the control group, treatment group, or both) do not express negative stereotypes at a higher level, overall, than self-declared supporters. One might then question the meaning of the reported difference in negative stereotyping between the two groups.

Or suppose that respondents express less support for affirmative action when the interviewer asks about negative stereotypes first. This finding, when combined with the result that Sniderman and Piazza report, would suggest something other than a simple, one-way causal relationship between stereotyping and affirmative action. At a minimum, the combination of results would imply that the designated independent variable in the mere-mention experiment lacks the stability that researchers take for granted. It would also raise the possibility of a complicated two-way causal process. Or, at the extreme, it would suggest abandoning any distinction between dependent and independent variable and instead thinking in terms of nonseparable beliefs and attitudes (Lacy 2001).

To be clear, we are not asserting that Sniderman's data reveal any such patterns. We have not analyzed the data and have no basis for conjecture that surprises lurk within. The finding that respondents who first hear about affirmative action are 7%–17% more likely to endorse certain negative traits certainly supports the inference that "dislike of particular racial policies can provoke dislike of blacks" (Sniderman and Piazza 1993, 104). Our point is simply that in this context—and others—the deliberate shuffling of questions in a cross-sectional survey can create multiple, intertwined experiments that speak to causal direction or at least to causal complexity. Assuming one-way causal relationships can be an error. Multiple treatments across time can offer even greater insight into the causal processes.

3.4 *Not Including a Control Group*

When scholars embed experiments in cross-sectional surveys, they sometimes compare two (or more) treatment condition results to each other and not to a control condition. The lack of controls is especially prevalent in survey experiments where changed question wording serves as the treatment. A typical framing study, for example, compares the attitudes toward

[5]Neither book, moreover, presents information about endorsement rates of the other negative qualities or any of the positive qualities.

a policy among those who receive one frame with those who receive another. Finding a significant difference, the researcher concludes that framing effects exist.

A study by Kinder and Sanders (1996) of affirmative action framing is a case in point. They show that those who read a description of affirmative action as a remedy for past discrimination supported the policy more than those who read a description of affirmative action as reverse discrimination. They conclude that how politicians frame policy debates shapes citizens' support for that policy.

Substantively, Kinder and Sander's lack of a reported control group makes sense. They wanted to know if the two dominant frames of the affirmative action debate could shape attitudes. But omitting a control condition—in this particular case, asking respondents to express their attitudes toward affirmative action in the absence of any frame—is risky. With no baseline, the researcher confronted with a significant difference between responses across treatments cannot know whether one frame or the other (or both) shapes attitudes. On the basis of the results they report, Kinder and Sanders cannot determine if people respond to only the "reverse discrimination" frame, only the "remedy for past discrimination" frame, or both frames. The two treatment group results could differ significantly from each other even though neither differs from the control group results, if they push people in opposite directions.

A study by Grant and Rudolph (2003) of how people balance the potentially conflicting values of equality and liberty shows the utility and importance of including a control group. The authors randomly assign respondents to one of three experimental conditions: a most-liked-group condition, a least-liked-group condition, and no-specific-group control condition. They then ask respondents a series of questions about equality and freedom of speech as the principles apply to campaign finance reform. They find that people give greater weight to free speech when they consider the speech of their most-liked group, and they give less weight to free speech when they consider the speech of their least-liked group. Significantly, however, they also show that people in the least-liked-group condition responded similarly to those in the control condition. In other words, when people respond to questions about free speech for interest groups generally, disliked groups apparently come to mind. Without a control group, Grant and Rudolph would not have made this discovery.

Both Kinder and Sanders and Grant and Rudolph, then, compare treatment groups. In both studies, these were, substantively, the key comparisons. By including a control group, Grant and Rudolph can tell a more complete story.

4 Contamination of Experimental Settings

Survey experimenters usually intend their studies to reveal the workings of the real world. So Sniderman and Piazza infer from the results of their mere-mention experiment that preferential treatment programs increase white people's negative stereotyping of black people. The now-prominent idea that citizens compensate for their informational deficiencies by using decision shortcuts—taking cues from politicians or interest groups, for example—arose as an inference from survey-experimental research (Sniderman, Brody, and Tetlock 1991; Mondak 1993; Lupia and McCubbins 1998; Mutz 1998). So did the equally important conclusion that people's real-world political opinions vary as a function of how politicians and others frame issues (Zaller 1992; Nelson, Clawson, and Oxley 1997; Grant and Rudolph 2003; Sniderman and Theriault 2004; for a comprehensive review of this literature, see Druckman 2001; for a contrary conclusion, see Druckman and Nelson 2003).

When seeking to infer to the real world, the survey experimenter's primary task is to create experimental contexts (conditions) that closely resemble real-world environments (Druckman and Lupia 2006). When Druckman and Nelson (2003) ask their subjects to discuss issues following exposures to various elite frames, they attempt to create a context—one where people discuss issues with each other—that resembles the real world. However, survey experiments themselves occur within two other contexts: the preceding parts of the survey instrument and the real world about which the researcher seeks to infer. To dramatize, as the respondent is asked a question about affirmative action, he might have just answered several other questions that elicit relevant attitudes, and he might have recently heard a debate about affirmative action on television. Each can contaminate the experimental results and the interpretations of them. We begin with the potential for accidental spillover effects from prior questions in the survey.

4.1 *Accidental Spillover Effects*

If their purpose is to explain real-world politics, survey experimenters presumably hope to find some long-term effects. Paradoxically, such enduring effects can raise havoc when a single survey includes multiple experiments and unintended order effects across (rather than within) experiments arise. Hereafter we describe two scenarios wherein experiments can be contaminated by their predecessors in a single survey. The moral is not that survey experiments must necessarily be isolated—the economics of conducting large-scale opinion surveys and the usefulness of the experimental method jointly rule out such an extreme remedy. But researchers should worry about experimental designs becoming accidentally intertwined, design surveys and conduct analyses accordingly, and sometimes alert readers to lingering possibilities of crossover effects.

Imagine a survey containing multiple items that rely on random assignment of subjects to control and treatment status. If the researcher makes all assignments probabilistically and independently, there will be some chance that later experiments will inherit prior treatment effects. By independent assignment, we *expect* such contamination to be minimal, but the more experiments researchers include, the greater the chance that systematic, cross-experiment effects will pop up, simply by bad luck.

Consider the simplest case, wherein N respondents are twice divided into control and treatment groups by coin toss. In expectation, the two treatment indicators will be uncorrelated, but any given set of realizations can be positively or negatively associated. To simplify exposition, suppose that both experiments are variations on mere mention. Experiment 1 asks respondents if blacks are lazy, with each respondent having a 50% chance of first hearing about affirmative action (treatment number 1). Experiment 2 asks if blacks are irresponsible and features a one-half chance of first hearing about out-of-wedlock birthrates (treatment number 2). In each case, the control group answers the identical question but is not exposed to mention of the treatment topic.

Suppose, next, that for all respondents there is a baseline probability of agreeing with any negative stereotype, b, that the first topic increases this probability for all respondents by p and the second increases it by q and that the effects are additive.[6] The usual approach to analyzing such experiments is to compare, for each, the proportions agreeing with the stereotype among the control and treatment respondents, taking their difference to be an

[6]Mathematically, it is immaterial whether we posit that all respondents share a common probability, b, of agreeing with a stereotype or that the proportion of respondents that agrees with the stereotype (with probability 1) is b (and the proportion that agrees with probability 0 is $1 - b$). These two portraits of the public are, of course, very different, but they are indistinguishable from a simple experiment.

estimate of what proportion in the general population can be induced to agree with the stereotype by mere reminder of the topic. Let y_k designate the proportion of respondents agreeing with stereotype k. From the joint distribution of the two treatment indicators,

		T_2		
		0	1	
T_1	0	n_{00}	n_{01}	$n_{0\cdot}$
	1	n_{10}	n_{11}	$n_{1\cdot}$
		$n_{\cdot 0}$	$n_{\cdot 1}$	N

the sequence of the questions, and the assumptions above, we can compute the expected value of the difference for the first question (for a given, fixed set of treatment realizations) to confirm that it is, indeed, p, indicating no bias:

$$E[(y_1 \mid T_1 = 1) - (y_1 \mid T_1 = 0)] = \frac{n_{1\cdot}(b+p)}{n_{1\cdot}} - \frac{n_{0\cdot}(b)}{n_{0\cdot}} = p. \quad (1)$$

Whether the treatment-minus-control difference for the second question will have an expected value of q, however, depends on the particular values taken by those n terms. The difference between agreement rates for the treatment and control groups is

$$E[(y_2 \mid T_2 = 1) - (y_2 \mid T_2 = 0)] = \frac{n_{11}(b+p+q) + n_{01}(b+q)}{n_{\cdot 1}} - \frac{n_{10}(b+p) + n_{00}b}{n_{\cdot 0}} = q + \left(\frac{n_{11}}{n_{\cdot 1}} - \frac{n_{10}}{n_{\cdot 0}}\right)p \quad (2)$$

The implication of the coefficient preceding p is that the effects studied by the first experiment can bias the results of the second experiment. To the extent that randomization failed to produce exactly equal probabilities of assignment to treatment number 2 for the treatment and control groups from experiment 1, bias follows.

For a given n, the probability of deviating enough to create substantial bias in the empirically derived estimate of q will depend on both p and the n terms. More generally, if we include k such experiments in our survey, there will be $\binom{k}{2}$ pairs of treatment indicators to consider and the odds of getting any unintended spillover from one experiment to a later one will be a function of both n and k.

Figure 1 illustrates how the expected size of the maximum bias coefficient falls as n rises for the case of 10 experiments (where maximum means the largest deviation in absolute value across the 45 paired comparisons). With as few as about 600 respondents, one does not expect any of the bias coefficients to exceed 0.10. Of course, absent corrections for such crossover effects, measurements can be biased upward or downward. Also, the figure plots a coefficient, not the actual amount of bias, which is also a function of the size of effect measured by the earlier experiment: when $p > q$, even a coefficient in the 0.05 range can lead to very misleading estimates of q. On the bright side, treatment assignments are observable; thus, as long as one is alert to the possibility of such bias, correction is possible.

A still more worrisome problem occurs if there are multiple avenues to induce changed behavior in (at least some) respondents, but the effects are not additive. Sticking with the

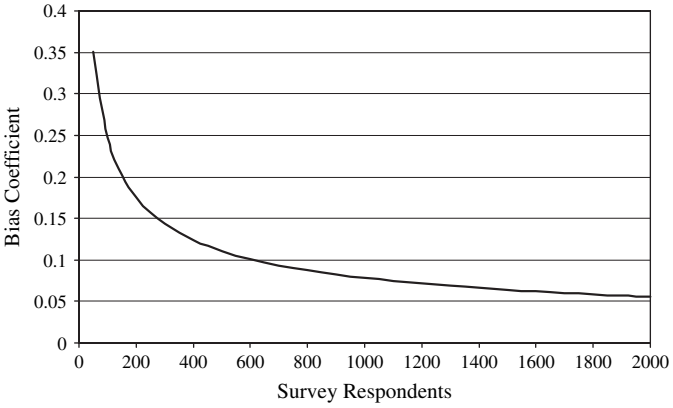

Fig. 1 Bias in downstream survey experiments by accidental association with upstream treatments, 10 (binary) experiments.

context outlined above, it could happen that respondents can be induced to endorse stereotypes by hearing mere mention of either of two topics, but that once the first treatment (mention of affirmative action) has had its effect, the second treatment (mention of out-of-wedlock births) can no longer have an effect. Such a ceiling effect will confound the second experiment, leading to an erroneous conclusion about its potential to sway people's opinions. There is, of course, a simple technical correction: vary the order of the two experiments and the results will show clearly that either topic can increase the average propensity to endorse stereotypes, provided it comes first. For a survey being implemented by pencil-and-paper methods, the problem is that varying the order of k experiments becomes impractical because it requires the preparation of $k!$ different forms. In the CATI or Internet context, a huge proliferation in "forms" is less a problem. To our knowledge, however, standard practice, at present, does not include randomizing survey experiment placement to the full extent. Moreover, even though computing power makes rampant randomization possible, simple mathematics conspire against thorough coverage of all orderings. Six experiments can be ordered in 720 ways and seven in 5040 ways, so the total number of possible forms can easily exceed the number of respondents.

4.2 *The Two Contexts Paradox: Survey Experiments and Real-World Inferences*

Quite apart from the issues of design and practice discussed so far, survey experimenters face complications because, if their research hypotheses have merit, the effects they simulate are likely to have occurred in the real world. In effect, some respondents are likely to have been contaminated by prior exposure to the treatment. If the effects never occurred, there would be no motivation for the research. Of course, if the effects occurred but were only fleeting, respondents would enter the survey essentially uncontaminated. But then the prior effects would be largely irrelevant for political behavior, and there would be little motivation for the research. Put simply, either there is a likelihood of contamination from real-world experience or the survey experiment explores a nonexistent or politically irrelevant phenomenon.

Survey experimenters assume that their active manipulations of the experimental treatments combined with random assignment will ensure valid results; there should be no other systematic differences between the control and treatment subsamples, and hence

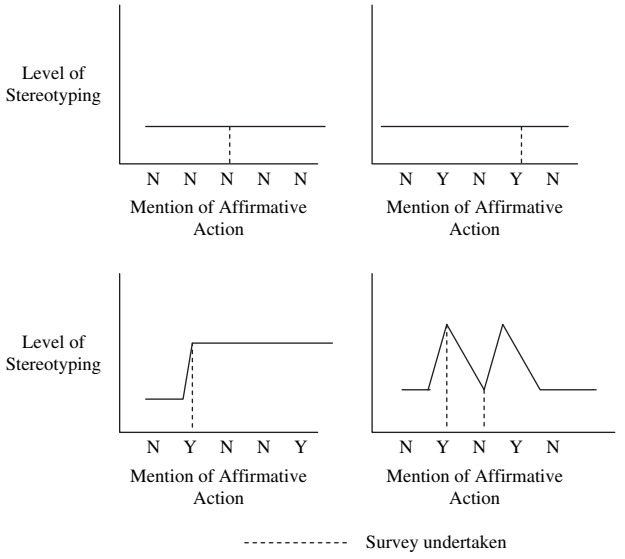

Fig. 2 Alternative real-world relationships between mention of affirmative action and negative stereotyping.

observed differences in their responses *must* be treatment effects. However, there is inevitably some possibility that respondents enter the experiment having already participated in a similar experiment, albeit one occurring in the real world.

To explore this problem of overlapping contexts—that survey respondents interrupt real life to undergo an experimental simulation of real life—we consider a simple framework that makes explicit some assumptions researchers usually leave implicit. Before turning to the framework, we raise a fundamental question that no one else seems to have posed: what forms might real-world cause and effect take? For purposes of discussion, we continue to use the mere-mention experiment as our focal point: the researchers prime some randomly chosen respondents to think about affirmative action and then ask questions about black stereotypes; they ask others about the stereotypes without first priming them. Those in the first condition stereotype more.

What might have been happening prior to the experiment? Figure 2 presents possible causal patterns, all assuming one-way causation. Vertical dashed lines indicate when a survey experiment is conducted. In the first, largely irrelevant case (top left), there has been no real-world mention of affirmative action. We include it to highlight a fundamental assumption that most survey experimenters make and that we noted above: the survey experiment simulates events that have already occurred in the world.

In the remaining panels, there have been real-world mentions of affirmative action, with different consequences for levels of negative stereotyping. The panel in the top right represents the case of no effect; mention of affirmative action never increases the level of stereotyping. The panel in the bottom left, by contrast, shows a once-and-for-all effect such that the first mention of affirmative increases stereotyping to a new level, where it stays whether or not the mention recurs in the period under consideration. In the bottom right, the first mention of affirmative action also increases the level of stereotyping; however, that effect quickly dissipates, and stereotyping returns to its original level. It stays there until the next mention, at which time it rises again. Only when the stimulus is

present does stereotyping reach a high level; as soon as the stimulus recedes (or shortly thereafter), so does stereotyping.

So which of these patterns (or many other possible alternatives) do researchers assume when they find a positive treatment effect? Because researchers have focused little attention on how real-world and experimental treatments are related, it is unclear. A simple model can illuminate how they might interact.

Suppose that 60% of the respondents in the mere-mention treatment group engage in stereotyping, whereas 40% in the control group do. Subtraction yields 20% as the estimated treatment effect, interpreted as the share of the population that can be induced into stereotyping by being prompted by mention of affirmative action.

But what if hearing about affirmative action causes a person to call to mind negative stereotypes for a sufficiently prolonged period that some of the survey respondents would, in fact, have been pretreated by their real-world experiences before the survey? Knowing how to interpret the data requires making additional assumptions about how two treatments interact. Even the terminology of a treatment effect (singular) becomes ambiguous when there might be equivalent treatments inside and outside of the experiment.

The simplest implicit model for the original treatment-control comparison is that the probability of stereotyping is some baseline, b, and that the treatment induces a change (increase) in this probability of p. Now let r and e designate real life and experiment, T and C designate treatment and control, and S designate agreeing with a negative stereotype. Each respondent will have a probability of stereotyping reflecting one of four states, according to the presence or absence of the two forms of treatment.

$$\begin{aligned} &P(S \mid C_{\mathrm{r}}C_{\mathrm{e}}) = b, \qquad P(S \mid C_{\mathrm{r}}T_{\mathrm{e}}) = b + p, \\ &P(S \mid T_{\mathrm{r}}C_{\mathrm{e}}) = b + p, \quad P(S \mid T_{\mathrm{r}}T_{\mathrm{e}}) = b + p + pd, \qquad 0 \leq d \leq 1. \end{aligned} \tag{3}$$

Here, we assume that real-world and experimental treatments have precisely the same effect in isolation (p) and that the combined effect of the two treatments is something between fully additive (when the discount parameter d is 1) and identical to the effect of only one treatment (when $d = 0$). The experimentally observed estimates of $P(S \mid T_{\mathrm{e}}) = 0.6$ and $P(S \mid C_{\mathrm{e}}) = 0.4$ can be understood as weighted averages of the terms above, with the weights reflecting what proportion of the respondents are assumed to have been pretreated. By treatment effect we mean p, so it could perhaps be more accurately described as the first-treatment effect.

Figure 3 shows, for this particular example, how estimates of b and p change as a function of assuming different levels of contamination of our sample by unmeasured real-world pretreatment, for three possible values of d. When $d = 1$, each treatment boosts the likelihood of stereotyping equally.[7] In that case, the usual treatment-minus-control difference in proportions produces an unbiased estimate of the original real-world effect no matter how prevalent the pretreatment. What changes as we change our assumptions about the proportion exposed to a real-world treatment, given fixed results, is the estimate of b, the baseline probability of stereotyping. This value is rarely discussed but is potentially interesting in itself, quite apart from the estimated treatment effect.

By contrast, when $d = 0$, that is, when treatments boost probabilities once and only once, the usual method of estimating the treatment effect produces an underestimate,

[7]This assumption would be untenable in an example with larger numbers of treatments or a sufficiently large baseline probability and large treatment effects, given the ceiling of 1 on all probabilities. For the sake of simplicity, we ignore that bound here.

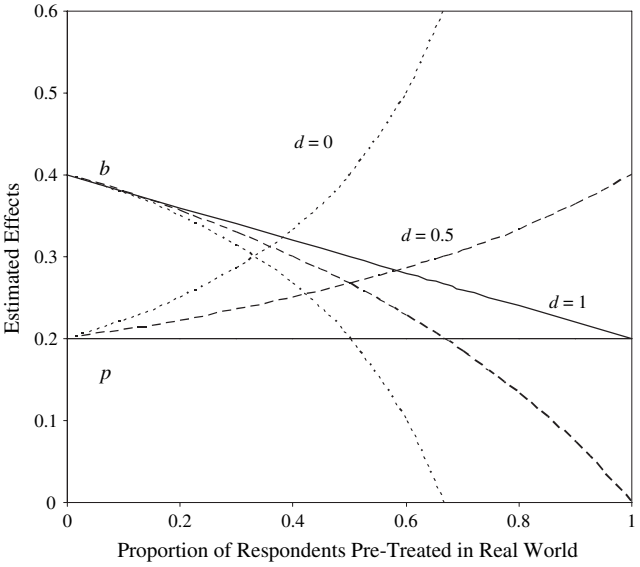

Fig. 3 An example of the impact of real-world pretreatment on estimated baseline (b) and treatment (p) effects.

unless none of the respondents arrived at the survey already having been exposed to a discussion of affirmative action. Indeed, for any $d < 1$, the treatment-control difference underestimates the actual p given any preexperiment, real-life treatment. An estimate of what proportion of the sample had been exposed to the treatment in real life could potentially be obtained from another survey item, although in practice it would probably prove extremely difficult to measure this trait.

One might conclude that the moral of the story is that estimates generated by survey experiments are inherently conservative, since any amount of pretreatment causes a downward bias in the estimate of p under any assumption except the extreme case of purely additive effects.[8] However, a different set of assumptions can produce a simple, experimentally generated estimate of the treatment effect that is too large rather than too small. For instance, the usual estimate of the treatment effect can be too large if the artificially clean environment of the survey question makes treatments easier to receive than in real life, where cues, frames, and communications can be misunderstood or missed entirely despite unambiguous exposure. Then, we might posit that the real effect in which we are interested is some parameter p that is inflated by the experiment by some factor (say f) so that the experimental treatment is pf, with $f > 1$. Carrying through the same algebraic exercise as above, there are then competing forces: pretreatment and possibly diminishing effects of repeat treatments causing the usual treatment-minus-control estimate to be too low, on the one hand, and the exaggerating effect of the distraction-free survey environment pushing the estimate up, on the other. A judgment on whether the observed treatment-control gap is likely to be an over- or underestimate of the actual real-life treatment effect, p, would then require assumptions or guesses about (1) how many respondents are

[8]Another bound shows up here: given $P(S|T_c) - P(S|C_c) = 0.2$, the assumption $d = 0$ is inconsistent with assuming that two-thirds or more of the respondents were pretreated, since b would become negative. The observed gap of 0.2 in the sample data would then have to be regarded as an unusually large, and thus highly improbable, measurement or sampling error.

likely to have experienced real-world pretreatment; (2) the magnitudes of d, the deflation parameter that captures how treatment can be less powerful given pretreatment; and (3) f, the inflation parameter that reflects the posited exaggeration of the treatment effects by the sterile survey context. Hence, it is wise to be as explicit as possible about the nature of the real-world effects being simulated in the experiment, to be mindful of the possibility of preexperimental exposures.

This simple framework can accommodate far more complicated assumptions about the two contexts, including differences in the magnitudes of the two kinds of treatment effects, measured or unmeasured heterogeneity in b and p, selection effects in the real-world but not the experiment (e.g., if more sophistication is required to receive treatment cues in the noisy real world), and so on. In some cases, the nature of the phenomenon under study might dictate which assumptions are most plausible and allow for some qualitative claims about likely direction, if not magnitude, of bias.

It could be that pretreatment of this sort is implausible, either because occasions for exposure to the treatment in normal life are rare or because the effects of treatment are likely to dissipate quickly. But either of these accounts is fatal to the political significance of the research, implying that the relevant real-world phenomena are nonexistent or inconsequential for political behavior most of the time. Considering that real-life pretreatment can suppress experimental treatment effects, there is a dark possible interpretation of large observed effects: namely, that most real-world effects are short lived. In a world where treatment is frequent and the effects last, demonstrating (additional) effects in the experiment should often be difficult.

Many survey experiments incorporate designs sufficiently complicated that the manner in which real-life experience could interfere with the survey is less obvious than in the example above. In the "list experiment," for instance, respondents are asked how many items from a short list anger them, with random insertion of an item about racial affirmative action (Kuklinski, Cobb, and Gilens 1997). Since the listed items are unrelated, it is hard to imagine how real-world pretreatment could occur.

Consider the "identity priming" experiment that Sniderman, Hagendoorn, and Prior (2004) undertook in the Netherlands. In asking respondents whether it was advisable to make immigration into the country more difficult, the researchers assigned them either to a lead-in about the importance of national identity or to one about personal uniqueness.[9] The experiment thus lacks a control group, with the attendant ambiguities already discussed. But what would constitute pretreatment in this instance? People differ in their senses of national and personal identity. Insofar as someone treated to the "personal" frame happens, through the accumulation of ordinary life experiences, to have a strong sense of being Dutch, that person has effectively been pretreated with the rival frame. But, so the argument goes, randomization will ensure that the two treatment groups contain equal numbers of both types. True, but that is not the point. Suppose that the two mindframes—think of them as real-world, enduring treatments—are not equally likely; that real life, much like the survey item, nudges people in the direction of seeing themselves as Dutch rather than as unique individuals; and that how much one is affected by exposure to a particular frame depends, in part, on how often such exposure takes place. These suppositions imply the possibility of confounding pretreatment as delineated above.

[9]The precise wordings were as follows: "People belong to different types of groups. One of the most important and essential of these groups is the nation which you belong to. In your case, you belong to the Dutch nationality. Each nation is different." And "People differ in many ways and each human being is unique. One person likes music, another likes to go for a walk, still another likes to go out. Everyone is different." (Sniderman, Hagendoorn, and Prior 2004, 44).

If respondents are asked a question about "poor people," "black people," or "poor black people," by random assignment, the fact that each individual gets only one version means that "not black" and "not poor" are (at best) implied in the former conditions. Hence, a rough analog of starting the survey with recent experience of having discussed affirmative action is having strong (unmeasured) associations of "black" and "poor." Similarly, if respondents are randomly asked about job set-asides either for African Americans or for Mexican Americans, some respondents may defy the contrast by inferring that such programs generally cover both blacks and Mexican Americans, despite no explicit mention of the other group.

Survey experimenters implicitly assume, first, that respondents enter the survey as clean slates and, second, that they can easily manipulate the latent tendencies or considerations floating beneath the surfaces of people's consciousness. But in a rival view, precisely because treatments are interesting when they are true to life, researchers frequently brush up against respondents recently manipulated in essentially the same ways as the experimental questions cue, poke, frame, and nudge them. This fact both underscores the relevance of the experiment (the good news) and, potentially, complicates the interpretation of the results (the bad news).

5 Conclusion: Moving On

To be clear, here are several conclusions that do *not* follow from the points we have raised above. We have not advanced the case that survey experiments serve little purpose. To the contrary, they represent a methodological breakthrough of great importance to public opinion and political psychology research. Their very importance motivated this article. We have also not claimed that survey experiments, as practiced, necessarily produce wrong substantive conclusions about real-world politics. We think there are grounds for concern about some standard practices, but we did not raise questions about practice to impugn past (groundbreaking) work. Instead, we hope to contribute to the goal of maximizing the survey experiment's future intellectual returns. Finally, we have not prescribed a new way of doing survey experiments. If there is to be a new way, it will evolve from the academic community, not from this article.

What of the concerns we raised about current practices? These challenges vary both in how likely they are to arise in any given survey experiment and in how seriously they will compromise the conclusions drawn from the experiment. We regard omitting a control condition, to preserve cases, as an understandable but inherently risky temptation. When researchers choose deliberately to omit a control, they must qualify their findings appropriately. The twin problems of not knowing how long a survey experiment's treatment effect persists and how multiple treatments affect the outcome of interest are more fundamental. A minimal and ultimately unsatisfying corrective, in both cases, is to qualify conclusions. In our view, researchers should measure the endurances of experimental effects and undertake overtime studies with multiple treatments as a matter of routine. These must be the signatures of the next generation of survey experimentation.

Panel studies can serve both ends. Given the existing limitations of widely used cross-sectional studies, in fact, they are a natural next step in the evolution of the survey experiment. TESS and similar vehicles increase the feasibility of such studies. Moreover, as long as survey experimenters do not tie themselves to expensive national surveys, they can conduct smaller scale panel studies that address politically important research questions.

Although we found it useful to make the classic mere-mention experiment a running example, experimental manipulation that consists of randomizing the order of survey

items is comparatively rare. We see strong potential in such designs and urge that researchers put the rich data thus created to full use in exploring the nature of potentially complicated causal relationships.

The danger of experiments contaminating one another is unavoidable as long as surveys contain multiple experiments. Potential question-ordering and -wording effects, after all, still remain. Of course, any survey item might, by chance, induce unequal effects in the control and treatment groups of a later item, but the deliberate creation of heterogeneity within the sample makes earlier experiments especially worrisome with respect to biasing results of later experiments. Here, technical fixes are available. On the one hand, randomization schemes over the entire survey that utilize quotas can ensure an absence of accidental imbalance. Otherwise, since treatment is observable, one can perform simple calculations to determine if other items have contaminated the straightforwardly estimated treatment effect.

Interference from real-life experience poses potentially bigger problems. Very often, such pretreatment cannot be tested or measured. Not all experiments are prone to real-world pretreatment effects, but we suspect that, more often than not, some respondents will begin a survey still under the influence of a true-life variation on some of the experiments contained within. Hence, experimenters ought to think of treatments not as binary but as continuous and, alas, unobserved in their real-world form. In some cases, theory and reasonable assumptions might permit qualitative assessments about likely direction of error, based on guesses about frequency of real-world treatments and the durations of their effects. Overall, however, real-world contamination represents a formidable challenge to survey experimenters.

One approach to the problem is so simple that it fringes on the laughable. When researchers conduct field experiments, they naturally know something about what has already occurred in their research contexts, for the research and real-world contexts are one and the same. Survey experimenters can gain considerable leverage solely by reviewing relevant events in the real world to which they seek to infer.

Survey experiments will continue to play a major role in public opinion research. They will also continue to evolve. Users of this methodology can shape this evolution to advantage by critically evaluating current practices and changing those practices where warranted. We fully expect the next generation of survey experimentation to take the study of public opinion and political psychology to new and as yet unanticipated heights.

References

Achen, Christopher H. 1986. *The statistical analysis of quasi-experiments*. Berkeley, CA: University of California Press.

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going negative: How attack ads shrink and polarize the electorate*. New York: Free Press.

Brady, Henry E., and David Collier, eds. 2000. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman and Littlefield.

Clarke, Harold D., Allan Kornberg, Chris McIntyre, Petra Bauer-Kaase, and Max Kaase. 1999. The effect of economic priorities on the measurement of value change: New experimental evidence. *American Political Science Review* 93:637–47.

Davis, Darren W., and Brian D. Silver. 2003. Stereotype threat and race of interviewer effects in a survey on political knowledge. *American Journal of Political Science* 47:33–45.

Druckman, James N. 2001. The implications of framing effects for citizen competence. *Political Behavior* 23:225–56.

Druckman, James N., and Arthur Lupia. 2006. Mind, will, and choice. In *The Oxford handbook on contextual political analysis*, ed. Charles Tilly and Robert E. Goodwin, 97–113. Oxford: Oxford University Press.

Druckman, James N., and Kjersten R. Nelson. 2003. Framing and deliberation: How citizens' conversations limit elite influence. *American Journal of Political Science* 47:729–45.

Fearon, James D. 1991. Counterfactuals and hypothesis testing in political science. *World Politics* 43:169–95.

Gibson, James. 1998. A sober second thought: An experiment in persuading Russians to tolerate. *American Journal of Political Science* 42:819–50.

Gilens, Martin. 1999. *Why Americans hate welfare: Race, media, and the politics of anti-poverty reform*. Chicago, IL: University of Chicago Press.

———. 2001. Political ignorance and collective policy preferences. *American Political Science Review* 95: 379–98.

Grant, J. Tobin, and Thomas J. Rudolph. 2003. Value conflict, group affect, and the issue of campaign finance. *American Journal of Political Science* 47:453–69.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* 64:605–54.

———. 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65:261–94.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66:1017–98.

Hovland, Carl I., and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly* 15:635–50.

Hurwitz, Jon, and Mark Peffley. 1997. Public perception of race and crime: The role of racial stereotypes. *American Journal of Political Science* 41:375–401.

Hyman, Herbert H., and Paul B. Sheatsley. 1950. The current status of American public opinion. In *The teaching of contemporary affairs*, ed. J. C. Payne, 11–34. Washington, DC: National Council for the Social Studies.

Inglehart, Ronald. 1990. *Culture shift in advanced industrial society*. Princeton, NJ: Princeton University Press.

Kinder, Donald R., and Lynn M. Sanders. 1996. *Divided by color: Racial politics and democratic ideals*. Chicago, IL: University of Chicago Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Kuklinski, James H., Michael D. Cobb, and Martin Gilens. 1997. Racial attitudes and the "New South." *Journal of Politics* 59:323–49.

Kuklinski, James H., Ellen Riggle, Victor Ottati, Norbert Schwarz, and Robert S. Wyer, Jr. 1991. The cognitive and affective bases of political tolerance judgments. *American Journal of Political Science* 35:1–27.

Lacy, Dean. 2001. A theory of nonseparable preferences in survey responses. *American Journal of Political Science* 45:239–58.

LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604–20.

Lieberson, Stanley. 1987. *Making it count: The improvement of social research and theory*. Berkeley, CA: University of California Press.

Lodge, Milton, Marco R. Steenbergen, and Shawn Brau. 1995. The responsive voter: Campaign information and the dynamics of candidate evaluation. *American Political Science Review* 89:309–26.

Lupia, Arthur, and Mathew D. McCubbins. 1998. *The democratic dilemma: Can citizens learn what they need to know?* Cambridge: Cambridge University Press.

Luskin, Robert C., James S. Fishkin, and Roger Jowell. 2002. Considered opinions: Deliberative polling in Britain. *British Journal of Political Science* 32:455–87.

Mondak, Jeffery J. 1993. Source cues and policy approval: The cognitive dynamics of public support for the Reagan agenda. *American Journal of Political Science* 37:186–212.

Mutz, Diana C. 1998. *Impersonal influence: How perceptions of mass collectives affect political attitudes*. New York: Cambridge University Press.

Mutz, Diana C., and Byron Reeves. 2005. The new videomalaise: Effects of televised civility on political trust. *American Political Science Review* 99:1–15.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review* 91:567–83.

Peffley, Mark, Jon Hurwitz, and Paul M. Sniderman. 1997. Racial stereotypes and whites' political views of blacks in the context of welfare and crime. *American Journal of Political Science* 41:30–60.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.

Schuman, Howard, and Lawrence Bobo. 1988. Survey-based experiments on white racial attitudes toward residential segregation. *American Journal of Sociology* 94:273–99.

Smith, Alastair. 1999. Testing theories of strategic choice: The example of crisis escalation. *American Journal of Political Science* 43:1254–83.

Sniderman, Paul M. 1996. Innovations in experimental design in attitude surveys. *Annual Review of Sociology* 22:377–99.

Sniderman, Paul M., Richard A. Brody, and Philip E. Tetlock. 1991. *Reasoning and choice: Explorations in political psychology.* Cambridge: Cambridge University Press.

Sniderman, Paul M., and Edward G. Carmines. 1997. *Reaching beyond race*. Cambridge, MA: Harvard University Press.

Sniderman, Paul M., and Thomas Piazza. 1993. *The scar of race*. Cambridge, MA: Belknap Press.

Sniderman, Paul M., and Sean Theriault. 2004. The dynamics of political argument and the logic of issue framing. In *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*, ed. Willem E. Saris and Paul M. Sniderman, 133–65. Princeton, NJ: Princeton University Press.

Sniderman, Paul M., Louk Hagendoorn, and Markus Prior. 2004. Predispositional factors and situational triggers: Exclusionary reactions to immigrant minorities. *American Political Science Review* 98:35–50.

Taber, Charles S., and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50:755–69.

Zaller, John. 1992. *The nature and origins of mass opinion*. New York: Cambridge University Press.