Advance Access publication January 6, 2009

Political Analysis (2009) 17:143–161
doi:10.1093/pan/mpn012

Treatment Spillover Effects across Survey Experiments

John E. Transue
Department of Political Science and Institute for Legal, Legislative, and Policy Studies in the Center for State Policy and Leadership, University of Illinois at Springfield, Mail Stop PAC 350, Springfield, IL 62703
e-mail: jtran8@uis.edu (corresponding author)

Daniel J. Lee
Department of Political Science, Michigan State University, 303 South Kedzie Hall, East Lansing, MI 48824
e-mail: leedan@msu.edu

John H. Aldrich
Department of Political Science, Duke University, 326 Perkins Library, Box 90204, Durham, NC 27708-0204
e-mail: aldrich@duke.edu

Embedding experiments within surveys has reinvigorated survey research. Several survey experiments are generally embedded within a survey, and analysts treat each of these experiments as self-contained. We investigate whether experiments are self-contained or if earlier treatments affect later experiments, which we call ''experimental spillover.'' We consider two types of bias that might be introduced by spillover: mean and inference biases. Using a simple procedure, we test for experimental spillover in two data sets: the 1991 Race and Politics Survey and a survey containing several experiments pertaining to foreign policy attitudes. We find some evidence of spillover and suggest solutions to avoid bias.

1 Introduction

Embedding experiments within surveys has reinvigorated survey research. Researchers use random assignment to conditions to create true experiments within (typically nationally) representative sample surveys; these surveys are often called ''survey experiments.'' They combine the internal validity of experiments with the external validity of national surveys. Thanks to these virtues, scholars using this technique have made impressive substantive contributions to several areas of political science (e.g., Sniderman and Carmines, 1997; Gilens 1999; Herrmann, Tetlock and Visser 1999).

Authors' note: Replication materials and an Appendix are available on the *Political Analysis* Web site. We would like to thank Richard Herrmann for sharing data and helping us understand the data's structure. We also would like to thank Paul Sniderman and Eric Lawrence for their advice, as well as presentation participants at Public Institutions and Public Choice at Duke University, University of Illinois at Urbana-Champaign, and Essex University for helpful comments on earlier drafts of this paper. *Conflict of interest statement*. None declared.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

For obvious reasons, several survey experiments are generally embedded within a survey. Analysts treat each of these experiments as self-contained. That is, although respondents may encounter several experimental treatments within the survey, researchers treat each experiment independently, not taking into account how earlier experimental conditions might influence the current survey question. From the existing literature (e.g., Zaller 1992; Wilcox and Wlezien 1993; Krosnick 1999; Tourangeau, Rips, and Rasinski 2000), we know that question wording and ordering can have large effects. Since experimental treatments are often strongly worded and/or affectively charged, we believe that it is important to investigate whether survey experiments are self-contained in practice or if earlier treatments affect later survey questions and experiments. We refer to this effect as ''experimental treatment spillover.''

The potential for treatment spillover effects is also one of the topics of the theoretical discussion in Gaines, Kuklinski, and Quirk (2007). They show analytically that as long as one has ''good'' random assignment, estimates of an intended experimental treatment are still unbiased in the presence of spillover effects provided that these effects are purely additive. Our paper advances the research on spillover effects in several respects. First, we consider a type of spillover discussed by Gaines, Kuklinski and Quirk, which has general implications for including experiments within any survey. Second, we focus on the more problematic case where the spillover effect is not simply additive. Third, we empirically test for the incidence of spillover, moving from a theoretical to empirical examination. Finally, we consider additional strategies to avoid potential spillover effects.

2 Two Threats to Valid Inference: Mean Bias and Inference Bias

Throughout this paper, we refer to experiments and their position in the survey from the respondent's point of view. When we refer to a ''previous'' experiment, we mean one where the subject has already been exposed to a stimulus and has already encountered the measure of the dependent variable that stimulus is believed to affect. When we refer to an ''intended'' experimental treatment, we mean the experiment at hand, and thus, the treatment that the researcher specifically hypothesizes will affect a particular response. That is, we typically use this term when the subject is in the midst of a stimulus-response chain intended by the researcher. Spillover occurs when a previous treatment influences a subsequent survey question or experiment. In our paper, we consider two types of bias that spillover can introduce, which are now discussed in turn.

2.1 *Mean Bias*

Experimental treatments might shift respondents' starting attitude when they encounter a later experiment's (or even nonexperimental) dependent variable.[1] We call this ''mean bias,'' and its effect is such that the estimate of an experimental treatment's effects on the population's position on an issue could be distorted. For example, responses to a policy issue, which a given experiment is designed to explain, may be shifted in a more conservative direction for those respondents who received a certain condition in an earlier

[1]We do not use the terms ''treatment'' and ''condition'' interchangeably. In our use, treatment refers to the factor being varied and condition refers to a specific stimulus. For example, the welfare mother experiment contains a race-of-mother treatment with two conditions. In one condition, the respondents were told that the welfare mother is white. In the other condition, the respondents were told that the welfare mother is black. Note that in this case there is no control, there are two different levels of treatment (i.e., two conditions). When an experiment only contains one control condition and a single treatment condition (as in Fig. 2 below), we say control and treatment as shorthand to indicate the control and treatment conditions.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

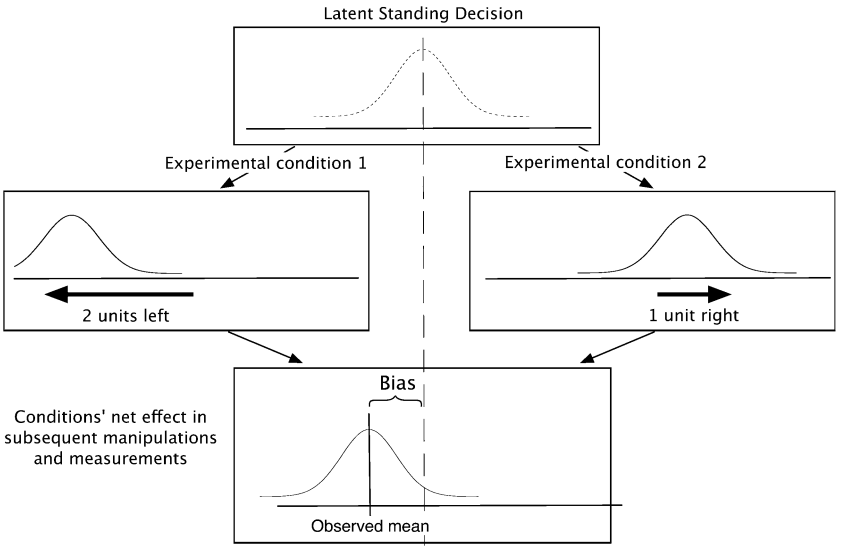


Fig. 1 Illustration of mean bias.

experiment than for those who did not. The problem of mean bias is compounded in the case of experimental treatments that do not include a control group, which we will clarify shortly.

Notice that the problem of mean bias spillover can be problematic to survey analysis more broadly through general question wording effects. Researchers might be interested in the mean response to a question, which is itself not an experiment but is asked after a survey experiment. If the earlier experiment shifted opinion, then this shift will be reflected in the mean response of the dependent variable of interest.

Figure 1 depicts the problem of mean bias graphically. Consider a policy scale where respondents choose a position along a left-right or similar dimension. Before encountering an experimental treatment, the policy question would have yielded a distribution of opinion as shown in the top row. A control condition would capture this distribution. In the second row of Fig. 1, half of the respondents have been randomly assigned to one condition of the experimental treatment that moves the distribution of their opinion to the right by one unit. The other condition moves opinion 2 units to the left. To the extent that the treatment's effects persist, the distribution of opinion is no longer the same as before the treatment.

If the conditions' effects are equal in size, but in opposite directions, then no *aggregate* bias is introduced, even though individual responses are biased. If the conditions' effects are unequal, as in Fig. 1, then estimates of the untreated underlying distribution of opinion will be biased. Point estimates of means from survey experiments are rarely discussed in political science research, but since experiments are embedded within a larger survey, experimental treatments may bias responses to nonexperimental measures for which mean responses are generally of more interest.

This example also highlights the pitfalls of not including a true control group in an experiment. In our example, analysts can only observe that the means of the two conditions are 3 units apart. But without knowing the pretreatment distribution, there is no way to know whether that mean difference signifies the situation shown in Fig. 1, a shift of 2 units to the right in one condition and a shift of 1 unit to the left in the other, or some other movement.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Although mean bias is problematic if one is interested in estimating the ''average attitude'' expressed in the survey (such as support for some policy), this form of spillover bias does not hinder the estimation of intended treatment effects in any later survey experiments (but see Gaines et al. 2007). That is, even if the ''starting point'' of respondents is shifted by an earlier treatment, that does not necessarily lead to improper inferences of the intended treatment effect. In contrast, the type of bias discussed in the next section does bias inferences on intended treatment effects.

2.2 *Inference Bias*

Inference bias is a greater cause for concern. The purpose of most survey experiments is to administer stimuli and observe their effects. They test whether treatments cause differences but are not concerned with the particular starting or ending point of the attitude. If previous treatments make intended treatments more or less influential than they would otherwise be, then the apparent effect of the intended treatments suffers from what we call ''inference bias.'' In the Gaines, Kuklinski, and Quirk (2007, 10–12) framework, we are concerned with nonadditive effects of previous experimental conditions. In the presence of spillover inference bias, the estimated effect of the intended treatment is biased. When researchers are unaware of spillover, they will attribute all of the effect on the dependent variable to the intended treatment, ignoring any lingering effects from earlier treatments.

As an abstract example, consider Fig. 2 and suppose that the same experiment (which we will call experiment 2) was given in two different data-gathering contexts. In the study represented by the left column, the experiment of interest is the first experiment the respondents encounter. In a survey containing several experiments, respondents encounter a different experiment before experiment 2. Assume that both experiments have a ''racial

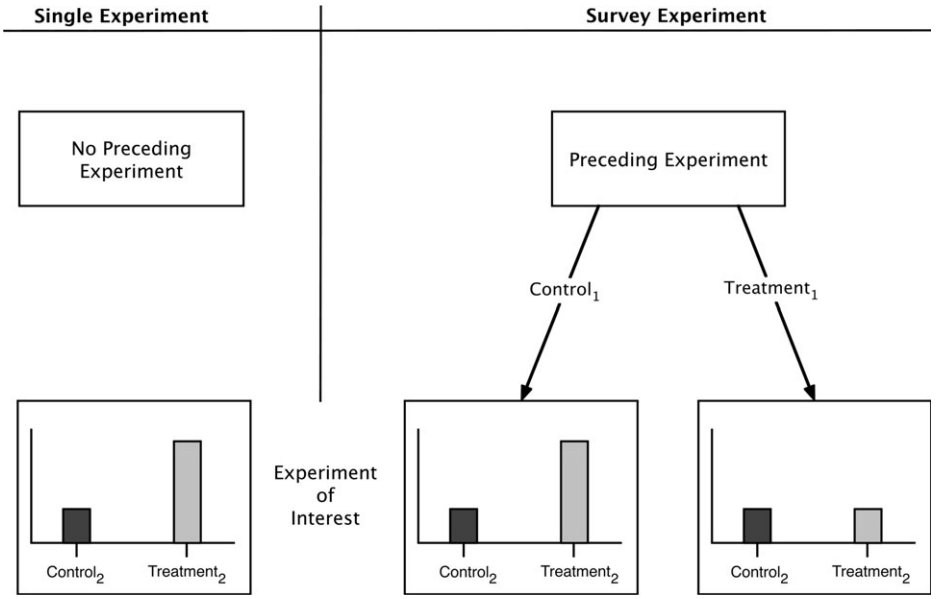


Fig. 2 Illustration of inference bias: later experiment altered by earlier treatment.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 1 Support for school tax (probit)

		Earlier experimental conditions	
	(1) Full Sample	*(2) Increase spending*	*(3) Fight discrimination*
Broad elite support	0.196* (0.099)	0.151 (0.141)	0.243† (0.139)
Minority opportunity	−0.397** (0.098)	−0.534** (0.137)	−0.262† (0.141)
Broad elite support × minority opportunity	−0.176 (0.141)	0.077 (0.201)	−0.412* (0.197)
Constant	0.292** (0.069)	0.327** (0.095)	0.257* (0.100)
Observations	2192	1102	1091

Note. Standard errors in parentheses. Significance levels: †$p < .1$; *$p < .05$; **$p < .01$.

minority'' condition and a control condition. A nonadditive spillover effect is present if the size of the effect of the second minority treatment depends on which condition the respondent received in the first experiment. For example, the second minority treatment might have a larger effect if the respondent was not already primed by the mention of minorities in the first experiment. Alternatively, receiving the first minority condition may take away from the effect of the second minority condition, as in Fig. 2.[2]

2.3 *An Empirical Example of Spillover*

To illustrate the potential substantive implications, we introduce our running example, the School Tax Experiment from the 1991 Race and Politics (RAP) Survey.[3] We discuss this experiment because it showed one of the largest inference bias effects of the entire survey and, therefore, gives a gives a good illustration of the potential inferential problems of spillover.

This experiment was given in the middle of the survey (experiment 6 by our numbering). The question asks whether the respondent supports raising their taxes to improve schools. The two treatments varied the characteristics of those calling for the law (*Some people* versus *Both the President and Congress, including both Democrats and Republicans*) and the beneficiaries of the law (*public schools* versus *minority opportunity*).

The third experiment in the survey, which we call the jobs experiment, asked the respondent if they favored more government involvement in helping blacks get jobs. The treatment varied how the government could get involved: (1) do more to make sure blacks are not discriminated against in getting jobs (*fight discrimination*) or (2) increase spending for programs to help blacks get jobs (*increase spending*).[4]

Table 1 shows how spillover could affect substantive interpretation. The baseline model (model 1), which only includes the intended treatments, shows that support for the law is higher when there is broad elite support and lower among respondents who were given the version stating that minorities are the beneficiaries. Model 1 does not reject the null of no

[2]This situation is an example of a ''ceiling effect'' as discussed in Gaines, Kuklinski, and Quirk (2007, 12).

[3]The 1991 RAP survey contains several survey experiments, and the vast majority of the survey's questions and experimental treatments are devoted to racially relevant policies and attitudes. We discuss later why these features made the RAP a good platform for our first test of the prevalence of spillover. The RAP survey is most thoroughly documented and explained in Sniderman and Carmines (1997). The codebook is available online at http://sda.berkeley.edu/D3/Natlrace/Doc/nrac.htm.

[4]The jobs treatment dummy is coded 1 for respondents given the first version (fight discrimination) and 0 for respondents given the second version (spending).

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

interactive effects between the two treatments. Models 2 and 3, which split the sample by the two conditions from the jobs treatment, illustrate how the estimates are influenced by the type of government involvement the respondents encountered in the third experiment. The baseline model did not find evidence of an interaction effect between the two intended treatments, but we find evidence of an interaction effect among respondents who were asked the version of the jobs question that stressed the government's role in protecting blacks from discrimination.

This difference could have theoretical implications if a hypothesis test hinges on the interaction term. In Table 1, compare model 1 with model 3. Consider if a researcher had a hypothesis that elite support for racially relevant programs generates populist backlash. Model 1 does not support the hypothesis, and that is what a typical analysis of the RAP would yield. However, model 3 shows a significant interaction between mentioning that the President and Congress support a tax and whether that tax will benefit the public schools or enhance educational opportunities for minorities. Which is the truth? Did the earlier ''fight discrimination'' condition rile up the populists in the sample? Did the earlier ''increase spending'' condition calm them down or make them numb to further treatments? Researchers working with these data do not know. If the ''fight discrimination'' condition had no effect on people (i.e., it is a *de facto* control), then it may be the case that model 3 depicts the true treatment effect for the School Tax Experiment. In that case the true effect of the elite support and minority opportunity interaction is masked by the inclusion of the ''increase spending'' condition. In sum, spillover can have a large substantive effect if it affects a parameter that is critical for testing a theory. This example also provides an illustration of some of the problems with experiments that lack a control condition, which we will discuss further in the following section.

Figure 3 shows this pattern graphically. In the full sample, the elite support treatment has little effect on respondents who receive the minority opportunity condition. When the results are separated by conditions from the jobs experiment, the relative levels of support vary and do so in different directions.

2.4 *A More Formal Consideration*

In this section, we show formally how experimental treatment spillover can introduce bias. We assume throughout that treatment randomization is perfect. Thus, we focus on the potentially more problematic forms of spillover biases, which Gaines, Kuklinski, and Quirk (2007) also allude to. We consider, in turn, the cases where the previous experiment includes and does not include a control group. A control group is the subsample of respondents that randomly receives a version of the survey question that does not shift the relevant initial attitude in any direction. We will see that whether a control is included influences estimates of spillover bias.

2.4.1 When the Previous Experiment Has a Control Condition

Building on Gaines, Kuklinski, and Quirk (2007, 10–11), let b denote the initial average attitude of survey respondents on a particular issue. Suppose there are two experiments, each with one treatment. We let $T_1 = 0$ denote the control condition and $T_1 = 1$ denote the treatment condition for experiment 1, and likewise use $T_2 = 0$ and $T_2 = 1$ to denote experiment 2's control and treatment conditions. Let the effect of the treatment in experiment 1, T_1, be p. Let q denote the true effect of the treatment in experiment 2 (T_2) on the response

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

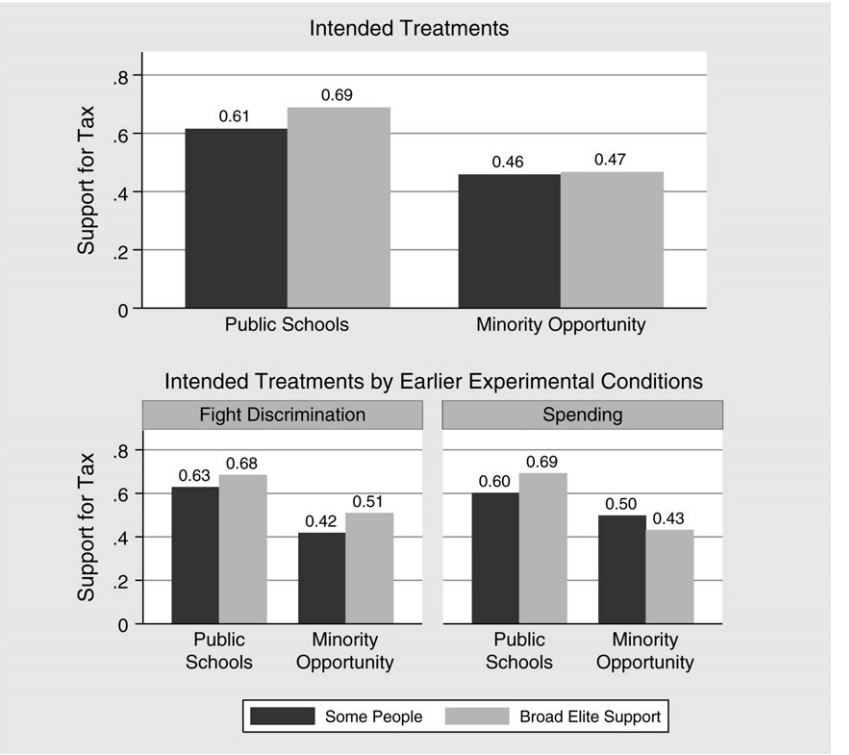


Fig. 3 Spillover into the school tax experiment.

to experiment 2 (y_2). The assumed relationships of treatments and attitude are summarized in Fig. 4.

In this setup, we are assuming that T_1 has a spillover effect, in that its effect lingers into experiment 2. Normally, researchers interested in the effect of T_2 on y_2 would estimate

$$y_2 = \beta_0 + \beta_1 T_2 + \varepsilon, \tag{1}$$

where ε is a normally distributed error term.

Estimating equation (1) would give $\hat{\beta}_0 = b + \frac{1}{2}p$ and $\hat{\beta}_1 = q + \frac{1}{2}r$ (assuming perfectly random and equal cell assignment). The mean bias is simply that the underlying average attitude for respondents who received the treatment in experiment 1 is shifted by p. In

		T_2 = 0	T_2 = 1
T_1	0	b	$b+q$
	1	$b+p$	$b+p+q+r$

Fig. 4 Treatment effects on attitude.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

this example, one would incorrectly infer that the initial attitude is $b + \frac{1}{2}p$, when the true value is b.

Inference bias is captured by the parameter r.[5] If $r = 0$, then there is not any inference bias. Notice that when $r = 0$, even if there is mean bias, we still get an unbiased estimate of the effect of T_2 by estimating the baseline model since we assume that random assignment is equal. This fact is formally shown by Gaines, Kuklinski, and Quirk (2007, 11). If $r < 0$, then T_1 suppresses the effect of T_2. Conversely, if $r > 0$, then T_1 increases the effect of T_2 on y_2. Therefore, if $r \neq 0$, then estimating the baseline model will give the biased estimate of T_2's effect of $q + \frac{1}{2}r$.

In order to detect spillover bias and get unbiased estimates of the initial attitude and the effect of T_2, one should estimate

$$y_2 = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 (T_1 \times T_2) + \varepsilon \tag{2}$$

and now $\hat{\beta}_0 = b$, $\hat{\beta}_1 = p$, $\hat{\beta}_2 = q$, and $\hat{\beta}_3 = r$. If the estimate of the interaction term, $\hat{\beta}_3$, is significantly nonzero, then there is inference bias. Likewise, if $\hat{\beta}_1$ is significantly nonzero, then there is mean bias.

Equation (2) diagnoses the presence of spillover. It also provides a correction for both mean and inference bias. By estimating equation (2), $\hat{\beta}_0$ gives an unbiased estimate of the true starting attitude b, which is made possible by the unbiased estimate of p, the effect of T_1 on y_2. The estimated effect of the intended treatment $T_2(\hat{\beta}_2)$ is also unbiased.[6]

A final observation is that, once spillover has been detected by estimating equation (2), one could also get the unbiased estimate of the effect of T_2 by estimating the baseline model only on respondents with $T_1 = 0$. That is, when there is a control group in the earlier experiment, one can recover unbiased estimates of the intended treatment by estimating its effect only on respondents in the control group from the earlier experiment.

2.4.2 When the Previous Experiment Lacks a Control Condition

Now suppose that there is no true control group for experiment 1. If $T_1 = 0$, then respondents are shifted by p_1. If $T_1 = 1$, then respondents are shifted by p_2. Since all respondents are shifted to some degree from T_1, a plausible conjecture is that both groups ($T_1 = 0$ or $T_1 = 1$) might have a nonadditive interaction effect with T_2. That is, both earlier conditions change the true treatment effect of T_2, which would occur if experiment 1 was not included in the survey. These assumed relationships of treatments and attitude are summarized in Fig. 5. Notice that there now are two nonadditive interaction effects between T_1 and T_2, r_1 and r_2. When T_1 included a control condition in the previous section, we assumed that $r_1 = 0$.

If we again estimate equation (2), then $\hat{\beta}_0 = b + p_1$, $\hat{\beta}_1 = p_2 - p_1$, $\hat{\beta}_2 = q + r_1$, and $\hat{\beta}_3 = r_2 - r_1$. Notice that unless $r_1 = 0$, we cannot truly detect or correct for mean or inference biases in this instance. None of our estimated coefficients produce either b or q if $r \neq 0$.

[5] The parameter r captures a similar effect to the parameter d in the discussion by Gaines, Kuklinski, and Quirk (2007) on the "two contexts paradox." In that section of their article, they consider how "real-world treatments" might alter the effects of intended experimental treatments and d (as r does in our discussion) captures a nonadditive interaction effect.

[6] Notice that even if we are only interested in correctly estimating the effect of T_2 (accounting for inference bias), one must also include the constituent terms in addition to the interaction. The reasoning follows the discussion in Brambor, Clark, and Golder (2006).

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

		T_2 = 0	T_2 = 1
T_1	0	$b + p_1$	$b + p_1 + q + r_1$
	1	$b + p_2$	$b + p_2 + q + r_2$

Fig. 5 Treatment effects on attitude: no control group.

2.4.3 Implication of Control versus No Control Condition

This discussion clarifies some limitations to our proposed spillover test. If there is a control condition for the previous treatment, then our diagnostic can detect and correct for mean and inference biases. Conversely, if there is no control group in the previous experiment, then our procedure is less accurate in estimating bias. Using equation (2) in the case of no control, we are estimating the *differences* in the intended effects, depending on assignment to previous conditions. This is still clear evidence that spillover is occurring since earlier treatments are influencing later experiments. However, we are less certain as to whether this spillover introduces any bias in the intercept (mean bias) or the estimated intended treatment effects (inference bias).

Our analysis here suggests that if the two conditions of the earlier treatment work in opposite directions (such as $p_1 < 0$ and $p_2 > 0$), then using equation (2) might overstate the amount of mean bias. If the effects of the two conditions are in the same direction (such as $0 < p_1 < p_2$), then the amount of mean bias might be understated. The same reasoning holds true for estimates of inference bias. Researchers generally do not know how each treatment condition of an experiment will change attitudes relative to the baseline attitude and also do not know how conditions might interact in multiple ways with later experiments, as in the more complex situation in Fig. 5.

Although the lack of a control condition hurts our ability to estimate the amount of mean or inference bias, we believe that our procedure is the best available test for the presence of spillover bias. At the very least, the procedure detects whether an experiment is influenced by an earlier treatment. And in spite of the difficulties discussed, the test still gives the best sense of whether any mean or inference bias might be present. That is, we believe that the absence of a control hurts *correction* more than *detection*. Future emphasis on including a control condition in survey experiments is needed to make better estimates of spillover bias (see also Gaines, Kuklinski, and Quirk [2007]'s discussion of the importance of control groups in survey experiments).

3 A First Test: The 1991 RAP Survey

To the best of our knowledge, we offer the first published empirical examination of spillover effects. Therefore, we chose to explore a survey with embedded experiments that is a ''most likely'' case for detecting spillover. Our reasoning is that if we do not find any evidence of spillover in what would seem to be a most likely case, then other surveys with embedded experiments are probably also safe from spillover effects.

The RAP Survey is well suited to serve as a test for spillover effects because of its content and structure. When a great deal of survey content is about the same topic, it

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

is more likely that respondents will be affected by spillover across experiments. The vast majority of the content on the RAP was on a single topic: race. In addition, because many people have strong attitudes on racial matters, there is a higher chance that respondents will still be affected by (or even elaborating on) earlier experimental treatments later in the survey. Finally, in contrast to our second multiexperiment survey test (which we discuss later), the structure of the RAP makes it more vulnerable to spillover effects because the order of the experiments was not randomized.

4 Hypotheses

We are agnostic about expectations. Given the fragility of survey results to seemingly minor changes in question wording and order, and given the strong affective reactions triggered by some experimental treatments, one might imagine spillover to be fairly likely. Just how spillover would manifest, if at all, is less clear. Our central empirical question is whether there are or are not spillover effects.[7]

4.1 *Procedure*

We began by analyzing the experiments independently as they were (essentially) designed. We refer to these analyses as ''baseline models,'' and we call those stimuli the ''intended treatments.'' For the baseline model, we estimated the model, where I is the current experiment with intended treatments $\{1, \ldots, j\} = J$, and $P(J)$ is the power set of J (but excluding the empty set):

$$y = f\left(\{I_B\}_{B \in P(J)}\right), \tag{3}$$

where y is the dependent variable (policy attitude), and I_{mn} is the interactions ($I_m \times I_n$), if any, between intended conditions m and n. Finally, $f(\cdot)$ is an additive linear function. This simply states that the regressors are all intended treatments and all possible interactions among those treatments.

To illustrate concretely how we ran our analyses, we return to the School Tax Experiment that we used earlier (see the online Appendix for the wording of all the experiments in the RAP). The question asks whether the respondent supports raising their taxes to improve schools. Two treatments varied the question wording, varying the characteristics of (1) those suggesting the law (some people versus broad elite support) and (2) the beneficiaries of the tax (public schools versus educational opportunities for minorities). For our baseline model, we simply test the effects of each treatment and the interaction of the two treatments on respondents' support for such a federal law

$$\text{Support school tax} = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_{12}, \tag{4}$$

where ''support school tax'' is the dependent variable (1 support; 0 do not support), I_1 is a dummy that indicates whether the respondent received the ''some people'' (coded as 0) or ''broad elite support'' (coded as 1) condition of the first treatment, I_2 is a dummy that

[7]Gaines, Kuklinski, and Quirk (2007) point to less-than-ideal randomization of survey respondents to experimental conditions as a source of spillover bias. We ignore this concern and focus instead on spillover bias that can occur even with perfect randomization.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 2 Substantive spillover: jobs experiment spillover into school tax experiment (probit)

	Full sample
Broad elite support	0.151 (0.141)
Minority opportunity	−0.534** (0.137)
Broad elite support × minority opportunity	0.077 (0.201)
Jobs	−0.070 (0.138)
Jobs × minority opportunity	0.272 (0.197)
Jobs × broad elite support	0.092 (0.198)
Jobs × broad elite support × minority opportunity	−0.489† (0.281)
Constant	0.327 (0.095)
Observations	2193

Note. Standard errors in parentheses. Significance levels: †$p < .1$; **$p < .01$.

indicates whether the respondent received the ''public schools'' (0) or ''opportunities for minorities'' (1) condition of the second treatment, and I_{12} is an interaction between the two treatment variables, $I_1 \times I_2$.

To test for spillover from the Jobs Experiment to the School Tax Experiment, we simply add the treatment dummy variable from the Jobs Experiment, and interactions between the Jobs Experiment treatment dummy and the dummies for the School Tax Experiment's intended treatments to the baseline model (equation 4). Table 2 presents estimates from this spillover model. The first three variables in the table show the baseline model (i.e., the terms in equation 4) by estimating the simple and interactive effects of the two intended treatments. The other terms test for spillover from the jobs experiment's treatment. First, the dummy for the jobs treatment tests for mean bias and finds no evidence of it. The remaining terms test for inference bias. The interaction between the jobs treatment and the intended treatment interaction is statistically significant at $p < .10$. The substantive implications of this model were discussed in Section 2.3. Our analysis of the tax experiment suggests that the effect of broad elite support is more complicated than the baseline model shows. Its effect is conditional on the specifics of who would benefit from the tax (intended interaction) *and* the government's proposed role in the jobs experiment (spillover interaction).

For each experiment, we proceed by adding the treatment variables to all subsequent experiments, testing both for direct effects (mean bias) and for interactions between the intended experimental treatments and earlier experimental treatments (inference bias). That is, we consider the spillover of *only one experiment at a time*.[8,9]

[8]In many of the RAP experiments, more than one treatment is used. Including all within- and across-experiment interactions quickly multiplies the number of parameters, such that controlling for multiple experimental spillovers within a single model is not feasible. That is, although multiple outside experiments might interact to spillover into the current experiment, for the sake of tractability, we do not statistically model such interactive effects.

[9]We also ran a ''cumulation'' analysis because race is affectively charged for many people, and repeated questions on the topic could increase a running tally of affect (Marcus et al. 1995). To measure cumulation, we coded each experimental condition as either racially charged (mentioned blacks, stereotypes of blacks, or controversial programs associated with blacks) or not. We then kept a running tally of how many racially charged conditions each respondent had encountered. The running tally was then included as a control variable, in addition to the intended treatment dummies and their interaction with the running tally. We did not find evidence of any cumulation effect for mean or inference bias. Results of the cumulation analyses are available from the authors upon request.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Equation (5) presents our spillover model, where P is the previous experiment with treatments $\{1, \ldots, k\} = K$, and $P(K)$ is the power set of K (but excluding the empty set):

$$y = f\left(\{P_A\}_{A\in P(K)} \cup \{I_B\}_{B\in P(J)} \cup \left\{\{P_A\}_{A\in P(K)} \times \{I_B\}_{B\in P(J)}\right\}\right), \tag{5}$$

where P_{mn} is the interaction $P_m \times P_n$, etc., and $f(\cdot)$ is an additive linear function. This expression simply states that the model's regressors are the intended treatments with all possible within-experiment interactions, a previous experiment's treatments with all of its possible within-experiment interactions, and all possible interactions between the two experiments. The simple additive terms $(\{P_A\}_{A\in(PK)})$ control for potential mean bias. The cross-experimental interaction terms $(\{\{P_A\}_{A\in P(K)} \times \{I_B\}_{B\in P(J)}\})$ control for potential inference bias. We do this for every experiment. Thus, with 17 experiments in the RAP survey, we estimate $16 + 15 + 14 + \ldots + 1$ spillover models.[10] Probit and ordered probit models are estimated for dichotomous and ordinal response variables, respectively.[11]

5 Results

5.1 *Statistically Significant Spillover*

When the models described in the previous section reveal statistically significant parameters involving treatments from experiments outside a given dependent variable and its intended experimental treatments, we interpret that as evidence of spillover. Table 3 reports that all but 2 of the 16 experiments in the RAP survey show some sign of the treatments influencing later survey experiments.[12] In Table 3, *tr*i denotes a dummy variable for treatment assignment in experiment i.[13] Taken as a whole, we find evidence of mean bias occurring about 1 time in 9, inference bias occurring 1 in every 10 times, and thus, our overall measure of bias occurs at a statistically significant level 1 in 10 times.[14]

Table 3 shows that, overall, spillover effects do not systematically affect findings from the RAP survey. Treatments from experiments 1, 6, 9, 10, and 11 appeared to have influenced later experiments more than the other experiments, particularly through mean bias. Looking at the content of these experimental treatments, it is not obvious why these experiments appear to have had significant effects on later experiments. The experiments do not share any obvious commonalities in structure or content.

Table 4 shows another perspective on the prevalence of spillover. This table shows the percentages of significant spillover coefficients for all estimated spillover models for each

[10]There is actually an experiment where one set of treatments is used for two separate, but adjacent, dependent variables. Thus, the actual number of estimated spillover models is $17 + 16 + \ldots + 7 + 6 + 4 + 3 + 2 + 1$.

[11]These models are run on all the white respondents to the RAP using the postratification weight. We are using white respondents instead of the entire sample because the experimental treatments were primarily designed to influence white attitudes.

[12]Since this is an exploratory study, we report statistical significance at $p < .10$, as well as the conventional $p < .05$ (both two tailed). Results of all estimated models are available online at http://polmeth.wustl.edu/polanalysis/vol/17/PA17_Transue.zip.

[13]See online Appendix for wording of the RAP survey.

[14]There were some treatment randomization problems at the beginning of the RAP's data collection, particularly for the treatments tr8 (rag6), tr9 (rlst), tr15 (rpor), tr16 (rqt), and tr17 (runv). The interpretation of the analyses in Table 3 are not changed by analyses that incorporate a variable denoting whether the respondent received nonrandom assignment. Further, from observing Tables 3 and 4, as well as the table available in the online supplement, there is no clear indication that the experiments that had randomization problems are more or less likely to experience or cause spillover as detected by our procedure.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 3 Percentage of statistically significant spillover parameters at 5% and 10%: effects on later experiments

			Mean bias			*Inference bias*			*Total bias*		
Experiment no.	*Treatments*	*No. of conditions*	*No. of estimates*	*5%*	*10%*	*No. of estimates*	*5%*	*10%*	*No. of estimates*	*5%*	*10%*
1	tr1a, tr1b, tr1c	2 × 2 × 2	119	10.9	17.6	231	5.6	10.4	350	7.4	12.9
2	tr2a, tr2b	3	32	6.3	6.3	62	0.0	4.8	94	2.1	5.3
3	tr3	2	15	0.0	6.7	30	0.0	6.7	45	0.0	6.7
4	tr5a, tr5b	3	28	3.6	3.6	56	1.8	1.8	84	2.4	2.4
5	tr4a, tr4b	2 × 2	39	0.0	0.0	75	2.7	10.7	114	1.8	7.0
6	tr6a, tr6b	2 × 2	36	11.1	11.1	66	9.1	12.1	102	9.8	11.8
7	tr7a, tr7b	3	22	4.5	9.1	40	2.5	12.5	62	3.2	11.3
8	tr8	2	10	0.0	10.0	19	0.0	5.3	29	0.0	6.9
9	tr9a, tr9b	3	18	11.1	16.7	34	2.9	20.6	52	5.8	19.2
10	tr10	2	8	12.5	12.5	16	6.3	6.3	24	8.3	8.3
11	tr11	2	7	14.3	28.6	15	20.0	20.0	22	18.2	22.7
12	tr12a, tr12b	2 × 2	18	0.0	5.6	36	0.0	13.9	54	0.0	11.1
13	tr13a, tr13b	2 × 2	12	0.0	0.0	18	0.0	5.6	30	0.0	3.3
14	tr14	2	3	0.0	0.0	5	0.0	0.0	8	0.0	0.0
15	tr15	2	4	0.0	0.0	6	0.0	0.0	10	0.0	0.0
16	tr16	2	1	0.0	100.0	2	0.0	50.0	3	0.0	66.7
Total			372	6.7	10.8	711	3.9	9.8	1083	4.9	10.2

dependent variable; that is, how often each question was influenced by outside experimental treatments. Similar to the pattern of only a small handful of experimental treatments affecting later experiments, we see only a handful of experiments that were influenced by outside experimental treatments beyond chance. Experiments 2, 6, 7, 8, and 11 appear to have been influenced by other experiments up to 25% of the time.[15]

Although the School Tax Experiment illustrates how spillover effects can alter substantive conclusions in important ways, our fuller analysis over all experiments shows a lack of large systematic biases. For instance, although treatments from the very first survey experiment showed signs of spillover throughout the survey, controlling for spillover does not change the sign of any of the intended treatment coefficients from the baseline models.[16] That is, if one is most interested in hypotheses about the direction of intended effects, then the effects of spillover appear to be minimal.[17]

6 A Second Test: The Herrmann, Tetlock, and Visser Survey

Our first foray into spillover effects suggests that survey experimental findings run a small but detectable risk of mean bias and inference bias. As mentioned earlier, we ran our first analysis on the RAP survey because we considered it the most likely candidate for these

[15]We also performed Wald tests on blocks of our bias parameters. That is, in the models described above, we tested whether all the spillover variables as a group significantly improved model fit over the baseline model. Out of 148 models, only 4 were significant at the $p < .05$ level.

[16]In a few instances, the spillover model causes the intended coefficient to lose statistical significance.

[17]The scarcity of large biases in estimated intended treatment effects is further shown in Table 3, where one sees, as noted earlier, the general pattern of more evidence for mean bias than inference bias.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 4 Percentage of statistically significant spillover parameters at 5% and 10%: how often each experiment was affected by earlier experiments

	Mean bias			*Inference bias*			*Total bias*		
Experiment no.	*No. of estimates*	*5%*	*10%*	*No. of estimates*	*5%*	*10%*	*No. of estimates*	*5%*	*10%*
2	7	0.0	42.9	14	14.3	21.4	21	9.5	28.6
3	9	0.0	0.0	9	0.0	0.0	18	0.0	0.0
4	10	0.0	0.0	20	0.0	0.0	30	0.0	0.0
5	12	0.0	0.0	36	2.8	11.1	48	2.1	8.3
6	15	13.3	13.3	45	15.6	24.4	60	15.0	21.7
7	18	22.2	22.2	36	2.8	16.7	54	9.3	18.5
8	20	10.0	10.0	20	10.0	25.0	40	10.0	17.5
9	21	14.3	28.6	42	4.8	4.8	63	7.9	12.7
10	23	0.0	8.7	23	0.0	4.3	46	0.0	6.5
11	24	12.5	16.7	24	0.0	4.2	48	6.3	10.4
12	25	8.0	8.0	75	4.0	9.3	100	5.0	9.0
13	28	3.6	3.6	84	2.4	8.3	112	2.7	7.1
13a	28	3.6	7.1	84	6.0	9.5	112	5.4	8.9
14	31	3.2	3.2	31	0.0	3.2	62	1.6	3.2
15	32	6.3	12.5	64	0.0	6.3	96	2.1	8.3
16	34	8.8	14.7	34	0.0	14.7	68	4.4	14.7
17	35	2.9	5.7	70	4.3	7.1	105	3.8	6.7
Total	372	6.7	10.8	711	3.9	9.8	1083	4.9	10.2

effects. We believe that its fixed order, focus on a single topic, and the affectively charged nature of its topic all increased the risk of spillover.

Having found some spillover effects, we moved on to an out of sample test by running the same analyses on another data set. We chose the Herrmann, Tetlock, and Visser (1999) data set because it differed from the RAP survey in several relevant ways. First, the Herrmann, Tetlock, and Visser (HTV) data surveys randomized the order of some of their experiments. This randomization allows us to observe whether order randomization can help alleviate spillover bias. Second, the HTV survey covered a range of topics related to foreign policy including value dispositions, references to specific countries, and attitudes about the use of force. Third, most of these topics were less affectively charged than race. Thus, analyzing the HTV data give us a less likely setting to test for spillover.

The HTV survey includes three blocks of experiments. The order of the blocks was randomized, whereas the order of experiments within each block was not randomized. In addition to these three blocks, for all survey respondents, two experiments were given before the randomized blocks, and another group of experiments were given after the three randomized blocks. That is, the very first and very last experiments were not randomized.

We focused on the experiments reported in Herrmann, Tetlock, and Visser (1999). The first two experiments were the very first two experiments given to all respondents—before any of the three randomized blocks. The other three published experiments were in the same block, which as a block was randomized relative to the two other blocks (see Fig. 6).[18]

[18]The order of the individual items was not randomized within a given block. The randomization changed the ordering of each block relative to the other two blocks.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Block	Experiment
First, Non-randomized	Repel Defend
I	Unpublished Unpublished Nukes Persian Gulf North Asia
II	Other Experiments
III	Other Experiments
Last, Non-randomized	Other Experiments

Note: Order of blocks I, II, and III were randomized.

Fig. 6 Structure of the HTV survey.

6.1 *Results*

We were only able to test for spillover on one of the experiments in the HTV survey. The nuclear weapons experiment is the first experiment in the randomized block of published experiments (which makes it the third of the published experiments). Both of the later experiments in the block contain so many conditions that the interactions with the earlier treatments caused an explosion in the number of necessary parameters.

Table 5 reports the baseline model estimate of the nuclear weapons experiment. Tables 6 and 7 test for spillover from the first and second experiments, respectively. As before, *tr*i denotes a dummy variable for treatment assignment in experiment *i*. Model 1 in both tables includes the full sample. The set of terms before *tr3* are the coefficients that test for mean bias from the treatments in the first (second) experiment. That is, they test whether being assigned to a condition in a previous experiment significantly altered respondents' position on the nuclear weapons dependent variable. *tr3* is the intended coefficient of the nuclear weapons experiment. The remaining terms test for distortions of *tr3*'s effect caused by previous experimental treatments.

In Table 6, we see that the first experimental treatments appear to have some spillover effect on the nuclear weapons experiment. In model 1, 3 of the 14 spillover coefficients are statistically significant at $p < .05$.[19] Models 2, 3, and 4 split up respondents depending on where the nuclear weapons experiment was placed in their survey experience. Looking at the effect of the first experimental treatments on the nuclear experiment, it appears that spillover effects fade as the target experiment moves farther temporally from the earlier experiment. When the nuclear weapons experiment is closest to the first experiment (i.e., when it is in block 1), three inference bias parameters suggest significant ($p < .05$) spillover; and more (four) inference bias parameters are significant at $p < .10$. In block 2, only

[19]Wald tests for both spillover models are insignificant.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 5 Nuclear experiment: baseline model (ordered probit)

	Coefficient (standard error)
tr3	0.522** (0.058)
Observations	1493
Pseudo R^2	.0187

Note. Significance levels: **$p < .01$.

one parameter suggests spillover at $p < .05$. When the nuclear weapons experiment is in the last block (most distant from the treatments we are testing), none of the parameters that test for bias are significant at $p < .05$. This pattern is consistent with an explanation that as memory of the initial question fades or is affected by the ''distraction'' of responding to other survey questions. However, this pattern of decaying spillover effects is not evident in Table 7, which tests for spillovers from the second experimental treatments on the nuclear weapons experiment.[20]

Table 6 Spillover effects on nuclear weapons experiment (ordered probit): spillover from first experiment

	Location of nuclear weapons experiment in survey							
	(1) Full sample		*(2) Block I*		*(3) Block II*		*(4) Block III*	
	Coefficient	*pValue*	*Coefficient*	*pValue*	*Coefficient*	*pValue*	*Coefficient*	*pValue*
tr1a	0.256†	.10	0.522†	.07	−0.189	.49	0.319	.21
tr1b	0.220	.18	0.495†	.08	0.035	.91	−0.145	.62
tr1c	0.389*	.02	0.253	.39	0.544	.05	0.344	.20
tr1a × b	−0.223	.32	−0.699†	.08	0.196	.64	0.215	.57
tr1a × c	−0.439*	.05	−0.696†	.09	0.013	.97	−0.522	.15
tr1b × c	−0.504*	.03	−0.532	.19	−0.318	.46	−0.357	.36
tr1a × b × c	0.411	.19	0.462	.41	−0.041	.94	0.466	.38
tr3	0.754**	.00	0.835**	.01	1.180**	.00	0.239	.40
tr1a × tr3	−0.358	.11	−0.690†	.08	−0.288	.48	0.019	.96
tr1b × tr3	−0.228	.33	−0.961*	.01	−0.123	.78	0.755†	.08
tr1c × tr3	−0.430†	.08	−0.460	.29	−1.041*	.02	0.243	.56
tr1a × b × tr3	0.399	.22	1.220*	.03	0.466	.45	−0.887	.12
tr1a × c × tr3	0.509	.12	0.858	.15	0.555	.34	−0.020	.97
tr1b × c × tr3	0.532	.12	1.225*	.04	0.393	.52	−0.404	.49
tr1a × b × c × tr3	−0.642	.17	−1.256	.12	−0.382	.65	0.175	.83
Observations	1493		501		481		511	
Pseudo R^2	.0208		.0245		.0493		.0215	

Note. Significance levels: †$p < .1$; *$p < .05$; **$p < .01$.

[20]We note that the intended treatment coefficient in Table 7 is only statistically significant for respondents in block II, which is the group that showed signs of spillover effects. However, pooling blocks I and III and estimating the spillover model for these two blocks gives a p value of .061.

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

Table 7 Spillover effects on nuclear weapons experiment (ordered probit): spillover from second experiment

			Location of nuclear weapons experiment in survey					
	(1) Full sample		*(2) Block I*		*(3) Block II*		*(4) Block III*	
	Coefficient	*pValue*	*Coefficient*	*pValue*	*Coefficient*	*pValue*	*Coefficient*	*pValue*
tr2a	0.143	.35	−0.046	.86	0.417	.14	0.129	.63
tr2b	−0.02	.86	−0.203	.47	0.111	.69	0.089	.73
tr2c	−0.012	.94	−0.144	.58	0.189	.52	−0.022	.94
tr2a × b	−0.108	.63	0.026	.95	−0.345	.36	−0.046	.90
tr2a × c	−0.162	.47	0.196	.60	−1.037*	.02	0.021	.96
tr2b × c	0.146	.51	0.095	.80	0.144	.72	0.165	.66
tr2a × b × c	−0.291	.36	−0.265	.64	0.358	.53	−0.701	.19
tr3	0.489**	.00	0.365	.15	0.898**	.00	0.336	.20
tr2a × tr3	−0.167	.46	0.001	1.00	−0.825*	.05	0.248	.53
tr2b × tr3	0.123	.58	0.212	.59	−0.143	.73	0.159	.67
tr2c × tr3	0.019	.94	−0.007	.99	−0.284	.51	0.253	.52
tr2a × b× tr3	0.062	.85	−0.107	.86	1.042†	.07	−0.688	.22
tr2a × c × tr3	0.306	.35	−0.145	.79	1.852**	.00	−0.453	.42
tr2b × c × tr3	−0.292	.36	0.162	.77	−0.359	.54	−0.602	.26
tr2a × b × c × tr3	0.214	.64	−0.019	.98	−1.096	.19	1.462	.06
Observations	1493		501		481		511	
Pseudo R^2	.0216		.0167		.0528		.0184	

Note. Significance levels: $^{\dagger}p < .1$; $^{*}p < .05$; $^{**}p < .01$.

According to our analysis of the HTV survey, question order randomization appears to reduce spillover. One pattern that does hold across Tables 6 and 7 is that fewer spillover coefficients are significant in the full sample compared to the subsample that showed the most significant signs of spillover. This pattern is particularly strong for the inference bias parameters. However, question order randomization is not a perfect solution. Notice the variation in the size of the intended treatment coefficient across the baseline (Table 5) and two spillover models (Tables 6 and 7, model 1). We particularly see a large difference between the baseline estimate and the spillover estimate in Table 6.

7 Implications and Concluding Remarks

We made three claims in this paper. First, we extended the study of Gaines, Kuklinski, and Quirk (2007) of how experimental treatments from one survey may induce a bias we call mean bias in one or more subsequent experiments to include a second and more serious bias that we call inference bias. Second, we proposed a straightforward and general diagnostic test for both forms of spillover bias and, at least when the experiments have a control condition, an unbiased estimator of the intended effects and of the bias parameters. Third, we tested for the incidence of spillover effects in the RAP multi-experiment survey and then, as an out-of-sample test, did so in the available portions of the HTV multi-experiment survey.

Although we found some evidence of spillover, there was no consistent pattern that would suggest a particular mechanism for spillover. We find a tapering-off effect in one spillover model using the HTV survey, for example, but we do not find a similar pattern

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

in another part of the HTV survey or in the RAP survey. From these analyses, we conclude that spillover can affect substantive findings in several ways, and the extent of spillover varies significantly across survey experiments.

An important step for future investigations of spillover effects is to run a survey with embedded experiments that each contains a control condition. Having clear baseline readings would allow researchers to use the procedure suggested in this article to estimate spillover biases more precisely. That is, whereas our procedure detects whether spillover is present, regardless of whether a control group is present, having a control group would increase our confidence that spillover effects are, indeed, introducing mean or inference bias by allowing more accurate estimates of the bias terms.

We believe that there are several ways to deal with spillover effects. When working with existing survey experimental data, the first step is to diagnose whether spillover effects are present and biasing intended treatment effects. We have outlined a simple spillover diagnostic: Include the potentially problematic outside treatments and their interactions with intended treatments as regressors. This test is straightforward since it relies on conventional statistical significance tests. We also note that this test provides more information when the potential spillover treatment includes a control condition. Experiments that show signs of spillover will then necessitate closer scrutiny by the researcher to ensure that results are robust.

What is less straightforward with our diagnostic is deciding which previous experiments are most likely to induce spillover. Future research should turn toward gaining a better understanding of what characteristics make experiments more or less likely to introduce such bias—that is, to develop a testable theory about the source of spillover effects. The mechanism of spillover might also depend on the characteristics of the experimental treatments—different types of stimuli might lead to different types of spillover. As always in survey and in experimental work, careful pretesting—in this case, to pretest for the possible spillover effects—is an important practice.

Our study, along with Gaines, Kuklinski, and Quirk (2007), points out important considerations for the development of future survey experiments. The best way to deal with spillover effects is to avoid, or at least minimize, the problem in the first place. Part of the answer, as just alluded to, is to decipher what experimental characteristics contribute to the likelihood of spillover. But until researchers develop a more nuanced understanding of spillover, there are immediate solutions that can lessen the chance of spillover bias.

One way to avoid spillover appears to be randomizing the order of experiments as done for some of the HTV experiments. This prevents experiments from systematically affecting each other by distributing whatever influences that might exist, as seen in our analysis of the HTV survey. Another technique that is used to reduce or compartmentalize the effect of experimental treatments is to give the respondents a distracter task such as a memory test to draw their attention to other topics. Since time is precious on surveys, this is often not a practical solution. However, separating experiments as far from each other in the survey as possible may achieve some of this effect. Lastly, we believe that including several different substantive topics on a survey could reduce the risk of experimental spillover. Questions and experiments on other topics could serve as *de facto* distracter tasks. However, this carries more risk than a distracter task that does not gather data because there is a chance that respondents will draw connections across survey/experiment topics that researchers have not anticipated. When using any of these techniques, we still suggest estimating spillover models as a robustness check against potential spillover.

Replicating an experimental treatment in a fresh data collection remains the best way to ensure the validity of experimental findings. The replication can occur in another survey

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press

experiment as long as the preceding questions are not very similar to the initial survey's. When there are concerns that experimental findings (or nonfindings) are artifacts of the other stimuli in a particular survey, replicating the experiment is the gold standard for ensuring that the experiments' effects have internal and external validity.

Survey experimental research has been a major advance in political science, as well as the rest of the social sciences, providing new insights into many social and political phenomena. As we know the fragility of survey response to question wording, so too should researchers consider how survey experiment design might affect the data generating process, as they develop new surveys. Concern with data structure and model misspecification is at least as important in survey research as in any other area of empirical analysis. Attention to survey and experimental design is therefore important to getting the most precise, valid, and robust results possible.

Funding

John Transue would like to thank the Center for State Policy and Leadership for summer research support.

References

Brambor, Thomas, William R. Clark, and Matt Golder. 2006. Understanding interaction models: Improving empirical analyses. *Political Analysis* 14:63–82.

Gaines, Brian J., James H. Kuklinski, and Paul J. Quirk. 2007. ''The logic of the survey experiment reexamined''. *Political Analysis* 15:1–20.

Gilens, Martin. 1999. *Why Americans hate welfare: Race, media, and the politics of antipoverty policy, studies in communication, media, and public opinion*. Chicago, IL: University of Chicago Press.

Herrmann, Richard K., Philip E. Tetlock, and Penny S. Visser. 1999. ''Mass public decisions on going to war: A cognitive-interactionist framework''. *American Political Science Review* 93:553–74.

Krosnick, Jon. 1999. ''Survey research''. *Annual Review of Psychology* 50:537–67.

Marcus, George E., John L. Sullivan, Elizabeth Theiss-Morse, and Sandra Wood. 1995. *With malice toward some: How people make civil liberties judgments*. New York: Cambridge University Press.

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

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge: Cambridge University Press.

Wilcox, Nathaniel, and Christopher Wlezien. 1993. ''The contamination of responses to survey items: economic perceptions and political judgments''. *Political Analysis* 5:181–213.

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

https://doi.org/10.1093/pan/mpn012 Published online by Cambridge University Press