Sociological Methodology
2015, Vol. 45(1) 184–222

DOI: 10.1177/0081175015570095
http://sm.sagepub.com

4

A DESIGN AND A MODEL FOR INVESTIGATING THE HETEROGENEITY OF CONTEXT EFFECTS IN PUBLIC OPINION SURVEYS

Stephen L. Morgan*
Emily S. Taylor Poppe[†]

Abstract

Context effects on survey response, caused by the unobserved interaction between beliefs stored in personal memory and triggers generated by the structure of the survey instrument, are a pervasive challenge to survey research. The authors argue that randomized survey experiments on representative samples, when paired with facilitative primes, can enable researchers to model selection into variable context effects, revealing heterogeneity at the population level. The value of the design, and its associated modeling strategy, is demonstrated by its ability to deepen the interpretation of a treatment effect of international competitiveness framing on long-used items drawn from the Phi Delta Kappa/Gallup Poll and the General Social Survey about the quality of schooling in the United States, confidence in the leaders running public education, and support for spending to improve schools.

*Johns Hopkins University, Baltimore, MD, USA
[†]Cornell University, Ithaca, NY, USA

Corresponding Author:
Stephen L. Morgan, Johns Hopkins University, Department of Sociology, 3400 N. Charles Street, Baltimore, MD 21218, USA
Email: stephen.morgan@jhu.edu

Keywords

context effect, framing effect, heterogeneity, selection, public opinion, survey response

1. INTRODUCTION

The framing literature in public opinion research has grown dramatically in the past two decades (see Chong and Druckman 2007, 2011), following prior psychological research on priming effects in social judgment (see Wyer and Srull 1989) and methodological research on context effects in survey response (see Schuman and Presser 1981; Schwarz and Sudman 1992). Substantive studies of framing consider the extent to which public opinions, and possibly underlying attitudes, reflect the manner and method by which information is delivered to individuals. This literature is dominated by experiments on student populations, which Druckman and Kam (2011) argued have the dual benefits of control over subjects and the timing of frame exposure.

Methodological studies of context effects consider the extent to which the structure and content of survey instruments alter response patterns for particular survey items, such as when early questions trigger information retrieval that determines how respondents interpret and answer later questions (see Tourangeau, Rips, and Rasinski 2000). Like the substantive framing effects literature, the context effects literature has also taken advantage of convenience samples. However, an important goal of this research is to better understand how particular context effects may have structured survey responses to questions in long-running opinion polls and surveys. This goal has led to a preference for representative samples. Schuman (2008) wrote,

> Although much research on context effects can be done with convenience samples such as students, at some points it is important to work with probability samples of a well-defined and heterogeneous population. This is of course expensive and time-consuming, but needed nonetheless. (p. 109)

In this article, we draw together these two traditions of analysis and present a design for survey experiments and an associated model for estimation. Our approach enables modeling of treatment effects that result from frame exposure as well as the variability of response patterns across individuals that are attributable to pretreatment exposure to the same substantive frame. In the sections that follow, we first delineate the essential features of the design. We then offer a conventional treatment effects model, which we estimate for a national survey experiment on the effect of international competitiveness framing on public support for education (previously analyzed, in brief simplified form, in Morgan and Taylor Poppe 2012). We then introduce a model that exploits variation in response to the facilitative prime, which permits investigation of the individual-level heterogeneity that lies beneath the treatment effects estimated by the conventional model. We then show how weighting control group subjects after estimating propensity scores can strengthen interpretations of results by adjusting away the sources of heterogeneity that are not produced by information retrieval relevant to the frame. In conclusion, we discuss limitations of the design.

2. THE RESEARCH DESIGN

2.1. *Precursors*

The survey response literature has long appreciated the value of randomized ballot designs with representative samples for the study of response patterns (see Schuman 2008 for a history). The General Social Survey (GSS), for example, has used randomized ballots to consider the consequences of wording changes and question placement for two of the items we analyze below (see Rasinski 1988, 1989; Smith 1987, 1991, 2006). Randomized ballot designs have also been used to investigate how elicited attitudes vary when they are preceded by alternative priming questions that promote retrieval of subsets of stored information relevant to the attitude. Tourangeau et al. (1989), for example, assessed how respondents from a random sample of adults in Chicago answered a question on legalized abortion after first answering context-setting questions about either gender equity in the workplace or traditional values.

Substantive research on framing effects is now dominated by scholars of political science and communications who seek to determine how political preferences in the electorate are shaped by different issue motivation and persuasion strategies.[1] In these fields, the power of experiments to identify effects is widely acknowledged, a position that has been bolstered by the growing interest in experimental methods in political science in general (see Druckman et al. 2011). And here we also find particularly strong interest in population-level experiments with national samples, which Mutz (2011) argued "may be unmatched in their ability to advance social scientific knowledge" (p. 157).

2.2. *The Design: A National Survey Experiment with a Facilitative Prime*

Following from these precursors, we consider a design with the following essential components:

1. a random sample of subjects drawn from a target population,
2. outcomes from questions that have been used repeatedly in polls and surveys that use random samples from the same target population,
3. randomization of treatment and control conditions across subjects, where
   a. the control condition is a baseline condition that mimics the administration of the questions on extant polls and surveys, and
   b. the treatment condition is structured as a facilitative prime that encourages subjects to reveal if they have been drawn into the real-world frame that is the subject of investigation.

Although the specific components of this design are not novel on their own, their joint adoption allows for the analysis of effects that cannot be considered in most conventional framing experiments. Most important, we show that the pairing of a randomized and facilitative treatment prime with a random sample drawn from a diverse target population allows the analyst to (1) identify and estimate the population-level treatment effects and (2) model response heterogeneity that can reasonably be attributed to pretreatment exposure to the frame of interest in the target population. The latter is possible under the assumption that the facilitative treatment prime triggers retrieval of information among those who have been exposed to the frame prior to the study.

3. ANALYSIS AND DEMONSTRATION

Following and extending the analysis of Morgan and Taylor Poppe (2012), to demonstrate the proposed approach we use a four-part substantive question: Do international competitiveness frames that suggest that the education system in the United States is losing ground to its competitors alter (1) public opinion about the quality of local public schools in the United States, (2) public opinion about the quality of the public schools across the United States, (3) confidence in people running the education system in the United States, and (4) support for spending additional resources to improve the nation's education system?[2]

Data are drawn from the 2011 Cornell National Social Survey (CNSS), which is a random sample of 1,000 individuals ages 18 years or older resident in the continental United States, interviewed by telephone.[3] The Supplementary Appendix, which is available on the authors' personal Web sites as well as this journal's Web site, offers descriptive statistics that demonstrate that the CNSS generated a national sample with typical distributions across demographic characteristics. We analyze the survey data as a self-weighting national sample, but we use estimated weights to adjust for nonresponse for each of our four outcome measures. A small amount of item-specific missing data on covariates for our final set of models is imputed with best-subset linear and logistic regression.

The CNSS interviews began with questions from a split-ballot priming experiment. The two alternative experimental ballots are presented in Figure 1. A randomly selected 47.1 percent of respondents were allocated to the treatment group, and they began the interview with two questions that prime international competitiveness. They were then asked four questions that have been administered repeatedly over the past four decades in high-profile national surveys: the first two in the Phi Delta Kappa/Gallup Poll (PDK/GP) and the second two in the GSS.[4] The remaining respondents were allocated to the control group, and they proceeded immediately to the same four attitude questions that the treatment group answered only following the priming questions.[5]

Consider the structure of the two priming questions on the treatment ballot. Neither priming question gives respondents any information on international differences in economic competitiveness or educational performance. In fact, the second question allows respondents to disagree with the common framing of journalists and political elites that the United States is losing some of its international competitiveness because of a decline in the quality of its public K–12 education system. As such, these two questions constitute what we label in this article a "facilitative prime," rather than a standard manipulative prime. In particular, these two questions prompt respondents to reveal, on the basis of their own information and beliefs, whether they will approach the four subsequent attitude questions after first invoking the frame of reference that is the

Treatment Ballot

1. Which of the following countries is the largest economic threat to the United States?
   - China
   - Germany
   - Japan
   - Russia

   {If another country was volunteered, country name was recorded}

2. In comparison to {insert country from prior question [or China if respondent answered "don't know" or refused]}, how much is our public education system losing ground?
   - None
   - A little bit
   - Some
   - Quite a bit
   - A great deal

3. Students are often given the grades A, B, C, D, and Fail to denote the quality of their work. Suppose the public schools themselves in your community were graded in the same way. What grade would you give the public schools here?
   - A
   - B
   - C
   - D
   - Fail

4. How about the public schools in the nation as a whole? What grade would you give the public schools nationally?
   - A
   - B
   - C
   - D
   - Fail

5. Consider now the people running the public education system in the United States. Would you say that you have:
   - A great deal of confidence in them,
   - Some confidence in them,
   - Hardly any confidence at all in them?

6. We are faced with many problems in this country, none of which can be solved easily or inexpensively. In order to improve the nation's education system, are we:
   - Spending too much money,
   - Too little money,
   - About the right amount?

7. Do you currently have any children attending the public schools in your community?
   - Yes
   - No

Control Ballot

1. Students are often given the grades A, B, C, D, and Fail to denote the quality of their work. Suppose the public schools themselves in your community were graded in the same way. What grade would you give the public schools here?
   - A
   - B
   - C
   - D
   - Fail

2. How about the public schools in the nation as a whole? What grade would you give the public schools nationally?
   - A
   - B
   - C
   - D
   - Fail

3. Consider now the people running the public education system in the United States. Would you say that you have:
   - A great deal of confidence in them,
   - Some confidence in them,
   - Hardly any confidence at all in them?

4. We are faced with many problems in this country, none of which can be solved easily or inexpensively. In order to improve the nation's education system, are we:
   - Spending too much money,
   - Too little money,
   - About the right amount?

5. Do you currently have any children attending the public schools in your community?
   - Yes
   - No

Figure 1. Alternative randomized ballots for the education module.
Source: Data from Cornell National Social Survey, 2011.

subject of investigation. The label "facilitative" is due to Sniderman (2011), who wrote, "Manipulative designs aim to get people to do what they are not predisposed to do," whereas "facilitative designs involve a directional force in the form of a relevant reason to do what people are already predisposed to do" (p. 108).

This design feature is not common, in part we surmise because it gives less control to the investigator. A manipulative prime delivers precisely the information to respondents that the investigator wishes to deliver, generating a potential response among all respondents. The investigator can then examine the effects generated by the delivery of this information, without considering whether some subjects already had been exposed to the information in the past. In contrast, a facilitative prime triggers the retrieval of stored information and beliefs, which have been shaped before the treatment is delivered. As a result, manipulative and facilitative primes motivate the study of related but distinct effects: the effects of delivered information on responses versus the effects of retrieved information on responses.

To appreciate the difference, consider an alternative manipulative prime that we could have used. As investigators, we would first decide to deliver two pieces of information: (1) recent results from international testing competitions and (2) a statement by a prominent public figure reflecting on the current economic competitiveness of the United States in relation to these test scores. One standard procedure would be to select a newspaper article that contains this information and excerpt from it appropriately (or, in more elaborate form, develop structured vignettes from alternative newspaper articles). In this case, many articles are available, and the following paragraphs from a late 2010 *New York Times* article would have worked well for a 2011 study:

> **Top Test Scores From Shanghai Stun Educators**
> By SAM DILLON
> Published: December 7, 2010
>
> With China's debut in international standardized testing, students in Shanghai have surprised experts by outscoring their counterparts in dozens of other countries, in reading as well as in math and science, according to the results of a respected exam.
>
> . . .
>
> The test, the Program for International Student Assessment, known as PISA, was given to 15-year-old students. . . .
>
> Shanghai students scored 556, ahead of second-place Korea with 539. The United States scored 500 and came in 17th, putting it on par with students in the Netherlands, Belgium, Norway, Germany, France, the United Kingdom and several other countries.

The article continues with authoritative quotations from opinion leaders:

> "Wow, I'm kind of stunned, I'm thinking Sputnik," said Chester E. Finn Jr., who served in President Ronald Reagan's Department of Education, referring to the groundbreaking Soviet satellite launching.
>
> . . .
>
> "We have to see this as a wake-up call," Secretary of Education Arne Duncan said in an interview on Monday.
>
> . . .
>
> President Obama recalled how the Soviet Union's 1957 launching of Sputnik provoked the United States to increase investment in math and science education, helping America win the space race.
>
> "Fifty years later, our generation's Sputnik moment is back," Mr. Obama said. With billions of people in India and China "suddenly plugged into the world economy," he said, nations with the most educated workers will prevail. "As it stands right now," he said, "America is in danger of falling behind."

For a manipulative prime, this article is ideal in some ways: It is detailed, authoritative, and unambiguous. Yet it would be hard to use for a national survey, including those such as the CNSS or the PDK/GP that use telephone interviews. Even for surveys that use face-to-face interviews, such as the GSS, this article would impose substantial cognitive burden on the respondent and absorb too much interview time.[6]

Instead, we use a facilitative prime that is short and easy to administer. With the first question, treatment group respondents are primed to be concerned about economic competitiveness. With the second question, they are asked to reflect on whether the U.S. education system is losing ground relative to the education system of its strongest current economic competitor. To answer this question, some respondents are likely to retrieve information from exposure to authoritative information from the past. Respondents who do not have information to retrieve, because they are not attuned to political discourse of this type, may approach the second question solely on the basis of personal experience and more general attitudes about social services. The question of first order is whether the prime, and the information that it is presumed to cause at least some respondent to retrieve, shifts response patterns for the treatment group relative to the control group. In our first set of analyses, presented in the next section, we offer an affirmative answer to this question.

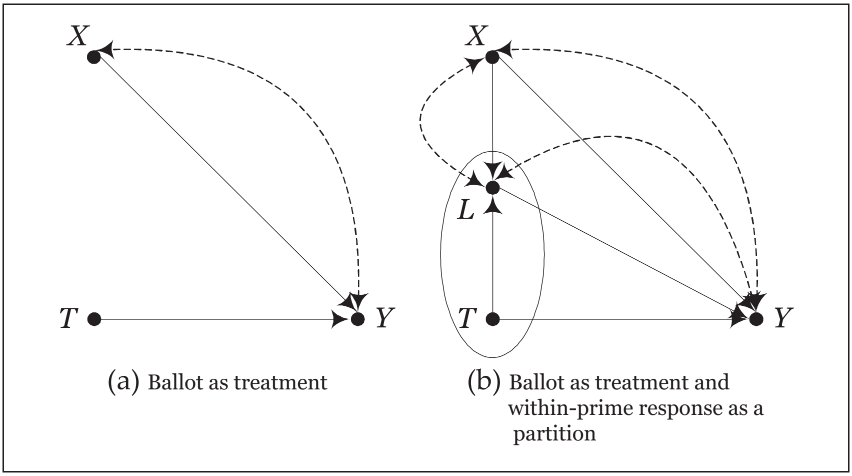

Figure 2. Directed graphs for the causal effect of the facilitative prime on attitudes.

3.1. *A Conventional Group-level Treatment Effect Analysis*

Figure 2a presents a directed graph that motivates the results of this section; see Pearl (2009, 2010) for a full treatment of causal graph methodology and Elwert (2013) for an introduction written for sociologists. The variable T is a dichotomous indicator variable for treatment group membership, equal to 1 for those randomly assigned to the treatment group and 0 otherwise. Note that the variable T encodes *only* exposure to the two questions that constitute the facilitative prime and does not encode any information on the responses that treatment group members provide in response to the questions. The variable Y represents the response distribution of any one of the four subsequent attitude questions (3–6 on the treatment ballot, which correspond to 1–4 on the control ballot). Finally, X represents a set of other observed variables measured for the CNSS, including baseline demographic characteristics of individuals, the response to the final question on each ballot that indicates whether the respondent currently has children attending a local public school, and other variables to be introduced later.

3.1.1. *Identification.* For this model, the treatment effect of T on Y is identified because the values of T are set by the ballot randomization. Because we consider below the more complex directed graph in Figure

2b, it is useful to state this claim more formally. In the causal graph tradition, this identification claim is written

$$P[Y|T] = P[Y|do(T)], \tag{1}$$

where $do(\cdot)$ is an abstract intervention operator (see Pearl 2009, 2010). In this case, the observed probability distribution of Y conditional on T can be given a causal interpretation in the values of T because T is set by an actual intervention. No confounding of the observed relationship between T and Y is present, because the distribution of T is completely random, except as would be produced by chance associations for the finite sample of respondents. Thus, the observed data represented by the left side of equation (1) identifies the causal effects defined by the right side of equation (1).

In the potential outcome tradition (see Morgan and Winship 2015 for an introduction written for sociologists, which also draws the connection to causal graphs), the same identification claim would be written on the basis of an independence assumption,

$$(Y^1, Y^0) \perp\!\!\!\perp T, \tag{2}$$

where Y^1 and Y^0 are potential outcome random variables that correspond to potential treatment exposures for values of T equal to 1 or 0 and where the symbol $\perp\!\!\!\perp$ denotes independence. When equation (2) is valid, which it is for this design because T is completely randomized, comparisons of analogous features of observed distributions of Y across values of T can be given warranted causal interpretations. Again, observed data are subject to sampling error, and the independence assumption in equation (2) applies to the design.

We assume for Figure 2a that the causal effects of the variables in X on Y are not identified, because of our assumption that common unmeasured causes of X and Y exist. In the causal graph literature, these common causes can be represented by a double-headed arrow, $\leftarrow\!\cdots\!\rightarrow$, that connects X to Y. Finally, we do not assume that the effect of T on the distribution of Y is the same for all individuals in the sample. Rather, these effects may vary across individuals, even in interactive fashion with the characteristics measured by X. All such interactions are implicitly embedded in the causal graph. The causal arrows, $\rightarrow$, in Figure 2a signify only that both X and T are causes of Y, and they are mute on whether these assumed effects are interactive or separable.

3.1.2. *Results.* To estimate the effect of T on Y, a model for the response distribution of Y must be chosen. Although alternative models (such as ordinary least squares linear regression) would convey the same basic pattern of results that we report below, ordered logit models have become the standard for modeling forced-choice responses to survey items with categories that are ordered but cannot be assumed to be equidistant on a latent response scale. Accordingly, Table 1 presents estimated coefficients for the treatment for four separate ordered logit models with responses to the four attitude questions as the outcome Y.

Grades for schools are the dependent variables for the first two rows, and the estimated coefficients for the treatment group are –.28 and –.21. With the same associated standard error of .12, the first coefficient would be judged significant using a standard two-tailed test with a null hypothesis of zero and the second nearly so. Substantively, the coefficients indicate that respondents who were presented with the two-question facilitative prime gave lower grades to public schools in their communities and in the nation as a whole. Morgan and Taylor Poppe (2012) reported the coefficient for the question on community schools, presenting fitted values that indicated that the coefficient corresponded to 6.8 percent of respondents shifting their grades from A or B to C, D, or F. The second coefficient suggests a slightly smaller substantive shift for schools in the nation as a whole.

The third and fourth rows of Table 1 present analogous results for the two GSS items.[7] For confidence in the leaders of the public education system, the coefficient for the treatment is –.17. With a standard error of .13, the implied negative effect is not statistically significant by the usual standards. However, it is consistent with the direction of the effect for the overall ratings of schools by grades and is larger than its estimated standard error. For the attitudes toward spending to improve education, the estimated coefficient for the treatment is –.30. This is the same coefficient reported by Morgan and Taylor Poppe (2012), on the basis of which they concluded that the international competiveness prime leads respondents to decrease support for additional spending to improve schools. They report a predicted response difference of 7.2 percent for indicating that "too little" money is spent on improving the nation's education system, which they noted is "more than enough to alter the outcome of hypothetical elections for local school board seats and funding levies" (Morgan and Taylor Poppe 2012:265).[8]

Table 1. Treatment Group Coefficients from Ordered Logit Models for Each Outcome Question

Question	Treatment Group Coefficient (*SE*)	*N*	Chi-Square Test Statistic (*df*)
Grades for public schools "in your community"	−.28 (.12)	928	5.1 (1)
Grades for public schools "in the nation as a whole"	−.21 (.12)	926	2.8 (1)
Confidence in "people running the public education system"	−.17 (.13)	971	1.8 (1)
Support for spending "to improve the nation's education system"	−.30 (.13)	968	5.5 (1)

Source: Data from Cornell National Social Survey, 2011.
Note: For grades, the highest response category is A, and the lowest response category is fail. For the confidence question, the highest response category is "A great deal of confidence in them," and the lowest response category is "Hardly any confidence at all in them," with "Some confidence in them" as the middle category. For the spending question, the highest response category is "Too little money," and the lowest response category is "Too much money," with "About the right amount" as the middle category. All models are weighted by the inverse probability of providing a response to the outcome question, as estimated by a supplementary logit model.

3.1.3. *Interpretation.* The conclusions suggested by a conventional treatment effects analysis are straightforward. The facilitative international competitiveness prime causes respondents, on average and in a nationally representative sample, to lower their subjective assessments of the quality of schooling while decreasing support for additional spending to improve the nation's education system. This analysis, although entirely appropriate, does leave one important question on the table: Do the same types of individuals move in response to the treatment prime for all four of the outcome questions, such that those who lower their quality ratings are the same types of individuals who also do not wish to spend any more money on schools? As we show in the next two sections, the response to the second item of the facilitative prime allows us to address this question.

3.2. *A Simple Subgroup-level Response Heterogeneity Analysis*

Figure 2b presents a directed graph that motivates the extended results of this section and the next. The variables T, X, and Y are the same variables defined for Figure 2a. In addition to these variables, a variable L, denoting the attitude "losing substantial ground," is included within an ellipse along with T. The partition of the treatment group represented by L is the key to our analysis in this section and the next. Respondents who conclude their engagement with the facilitative prime by stating the opinion that schools in the United States are losing substantial ground to those of the nation's strongest economic competitor have, in our interpretation, entered into the frame of interest by retrieving stored beliefs based on prior exposure and responsiveness to the frame.

Consider, first, how L is coded for the observed data. L was set equal to 1 for treatment group respondents who answered the second priming question with "quite a bit" (23 percent) or "a great deal" (26 percent), and it was set equal to 0 for all other treatment group respondents—those who answered "none" (9 percent), "a little bit" (11 percent), "some" (23 percent), and "don't know" (3 percent) and those who refused (.6 percent). Accordingly, L is an indicator variable for the 49 percent of the treatment group that has a pronounced belief that the education system in the United States is losing substantial ground.

For subsequent data analysis, we use two distinct treatment subgroups: (1) the "losing ground" treatment subgroup for which $T = 1$ and $L = 1$ and (2) the "not losing ground" treatment subgroup for which $T = 1$ and $L = 0$. Both of these treatment subgroups are compared with the undifferentiated control group for which $T = 0$.

Figure 2b stipulates that L is caused by the observed variables X and unobserved common causes of X, L, and Y (collectively represented by the double-headed arrows in $X \leftrightarrow L$ and in $L \leftrightarrow Y$).[9] In the section that follows this one, we use the variables in X to develop our interpretations for the differences between the control group and the two treatment subgroups. For now, we consider only treatment effects defined in T and L.

3.2.1. *Identification.* The total causal of effect T on Y in Figure 2b remains identified for the same reasons stated in the last section. The effect defined only by the two values of T is still identified by the conditional distribution $P[Y|T]$, regardless of whether the treatment group can be or is partitioned using L. In causal graph terminology, no backdoor paths connect T to Y in Figure 2b, as in Figure 2a.

To begin to understand the complications that arise when we partition the treatment group using L, we need to explain (1) the missing data pattern for L and (2) the special nature of the partitioning variable

L. For the first explanation, note that the observed values for L are completely missing for control group respondents. Yet, because treatment and control group respondents are collectively exchangeable, we can assume that the unobserved distribution of L in the control group would be the same as the observed distribution in the treatment group, subject only to sampling error. In other words, had the control group respondents been exposed to the treatment conditions instead, they too would have responded to the second priming question and chosen values for L that would reproduce the same distribution observed in the treatment group, subject only to variation from finite sampling. Thus, we have a particular form of missing data. Data are missing on L as a deterministic function of T. And because T is set by randomization, whether the data are missing is completely random. However, because the particular values of L have nonrandom causes, which according to the assumptions embedded in Figure 2b include X as well as unobserved common causes of L and X and of L and Y, the individual-specific missing values on L are not missing at random.

For the second explanation, note that the causal effect that we have represented as $T \rightarrow L$ in Figure 2b is different than the other effects in the figure. We signify its special nature by embedding L within an ellipse that includes T.[10] On one hand, respondents who are in the treatment group are exposed to the first priming question, and as such the value that they then provide for L in response to the second question is a function of having been presented with the first priming question in the initial exposure to the treatment conditions. In this sense, being in the treatment group does entail exposure to a question that shapes the particular pattern of responses to the question that then generates L. On the other hand, only treatment group respondents receive the question that generates L. And, in fact, the content of the second priming question is set by the response to the first priming question—the country nominated as the largest economic threat. Thus, no observable data based on this design could ever identify a causal effect of T on L because, even in theory, we cannot intervene separately on T and L without changing the design that we have proposed and that has actually been implemented for this study.[11]

Because the treatment and control groups can be regarded as two independent samples from the same population, the randomization of T allows us to assert that $P[L|T=1]=P[L|T=0]$. Therefore, we can maintain that there is an effect of T on Y within strata of L. In other words, because we can conceive of an abstract scenario in which we could eliminate our missing data problem for L by repeatedly rerandomizing T until all individuals in the population have been exposed to the treatment at least once, we can maintain that causal effects defined within population strata enumerated by L exist in theory and are well defined. With this stochastic conceptualization in the background, the values of L are therefore unobserved latent classes within our observed control group, which we can assume exist because the treatment and control groups are exchangeable.

The causal effects of interest can then be defined with reference to the causal graph or by using potential outcome variables. With the first notation, we are interested in quantities defined by

$$P[Y|do(T), L=1] \tag{3}$$

and

$$P[Y|do(T), L=0]. \tag{4}$$

In Pearl's (2009, 2010) framework, equations (3) and (4) are intervention-induced distributions that result from exposure to the facilitative prime, defined separately for two subgroups that exist in the population: the losing ground group ($L=1$) and the not losing ground group ($L=0$). These effects exist in theory, but the observed data do not identify them because the design does not generate values for L when $do(T=0)$.

Using potential outcome notation, equivalent causal contrasts are defined by the differences

$$P\left[Y^1|L=1\right] \ v. \ P\left[Y^0|L=1\right] \tag{5}$$

and

$$P\left[Y^1|L=0\right] \ v. \ P\left[Y^0|L=0\right], \tag{6}$$

where the "*v.*" operator is the general "versus" notation used, for example, by Rubin (2005) to allow the analyst to consider any contrasting feature of the two probability distributions for the potential outcome random variables. The potential outcome notation, along with the design, imply that the contrasts in equations (5) and (6) can instead be written

$$P[Y|T=1, L=1] \text{ } v. \text{ } P[Y^0|L=1] \tag{7}$$

and

$$P[Y|T=1, L=0] \text{ } v. \text{ } P[Y^0|L=0]. \tag{8}$$

This way of expressing the causal effects of interest might provide a more transparent explanation for the challenges that confront estimation with our observed data. We cannot form conditional probability distributions for the right-hand sides of these equations because L is not observed for the control group (and even though the sample analog distribution of $P[Y|T=0]$ is consistent for $P[Y^0]$).

3.2.2. *Results*. Although the negative identification results just presented may appear dire, much interpretable analysis is possible, as we demonstrate in this section and the next. Table 2 presents coefficients from four models analogous to those presented in Table 1. Rather than including a single variable T as the sole predictor, we include dummy variables for the two treatment subgroups differentiated by L, in effect representing T and L in the ellipse in Figure 2b as a single cross-classified factor whereby the two latent classes $(T=0, L=1)$ and $(T=0, L=0)$ are collapsed into an omitted reference group. As shown in the first two rows, the "losing ground" treatment subgroup is much more likely to offer lower grades to schools "in your community" and "in the nation as a whole" than is the control group, with ordered logit coefficients of –.81 and –.83 (and with the same standard error of .15). More surprising, perhaps, are the positive coefficients for the "not losing ground" treatment subgroup, which imply that these respondents offer higher grades in comparison with the control group.

Figures 3 and 4 present predicted response probabilities that correspond to the models for the first two questions.[12] These figures indicate that 37 percent of the control group offered grades of C, D, or fail to schools in their communities and that 72 percent of the control group offered grades of C, D, or fail to schools in the nation as a whole.[13] The two treatment subgroups, however, have very different predicted response patterns. Treatment group respondents who did not see schools losing substantial ground offered slightly more positive grades than the control group, with 5 percent and 9 percent fewer respondents offering grades of C, D, or fail, respectively. Treatment group respondents who saw schools losing substantial ground offered much worse grades than

Table 2. Treatment Group by Losing Ground Coefficients from Ordered Logit Models for Each Outcome Question

Question	Treatment Group by Not Losing Ground Coefficient (*SE*)	Treatment Group by Losing Ground Coefficient (*SE*)	*N*	Chi-square Test statistic (*df*)
Grades for public schools "in your community"	.23 (.14)	−.81 (.15)	928	45.8 (2)
Grades for public schools "in the nation as a whole"	.39 (.15)	−.83 (.15)	926	50.5.8 (2)
Confidence in "people running the public education system"	.31 (.17)	−.62 (.16)	971	25.9 (2)
Support for spending "to improve the nation's education system"	−.53 (.15)	−.03 (.16)	968	14.1 (2)

Source: Data from Cornell National Social Survey, 2011.

Note: For grades, the highest response category is A, and the lowest response category is fail. For the confidence question, the highest response category is "A great deal of confidence in them," and the lowest response category is "Hardly any confidence at all in them," with "Some confidence in them" as the middle category. For the spending question, the highest response category is "Too little money," and the lowest response category is "Too much money," with "About the right amount" as the middle category. All models are weighted by the inverse probability of providing a response to the outcome question, as estimated by a supplementary logit model.

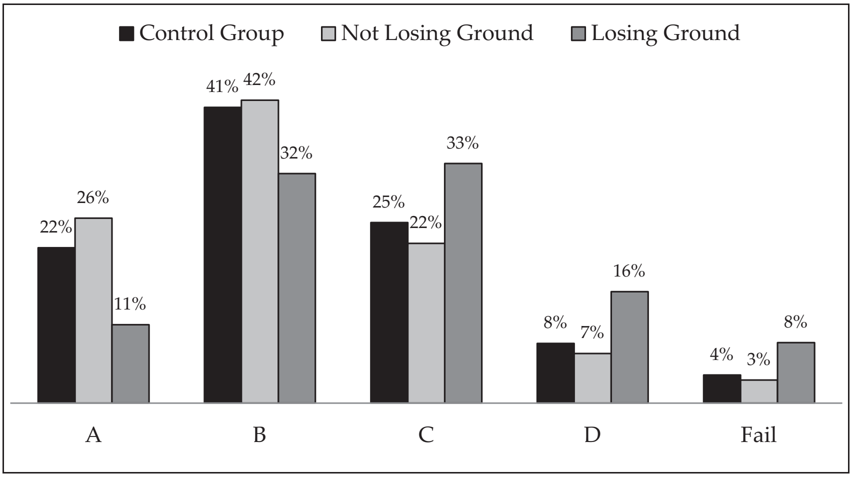

Figure 3. Predicted response probabilities for grades awarded to public schools "in your community," from the model reported in Table 2.

the control group, with 20 percent and 14 percent more respondents offering grades of C, D, or fail, respectively.

It is perhaps not surprising that individuals who express the belief that U.S. schools are losing ground to those of international competitors would then carry on to offer the lowest grades for schools. Yet the differences induced by the treatment are substantial and imbalanced across the two treatment subgroups defined by *L*, such that they combine to generate an overall decline in offered grades, as presented earlier in the first two rows of Table 1.

The third and fourth rows of Table 2 report analogous ordered logit coefficients for the two GSS questions. Treatment group respondents who indicated that schools in the United States were losing substantial ground expressed lower confidence in the people running the public education system, with an ordered logit coefficient of –.62 and a standard error of .16. Figure 5 shows that this difference is 14 percentage points in comparison with the control group, with 44 percent of this treatment subgroup having "hardly any confidence at all" in the people running the education system, in comparison with 30 percent of the control group. The other treatment subgroup again has the opposite pattern of responses, expressing more confidence in leaders than the control group. These results suggest substantial consistency across the PDK/GP

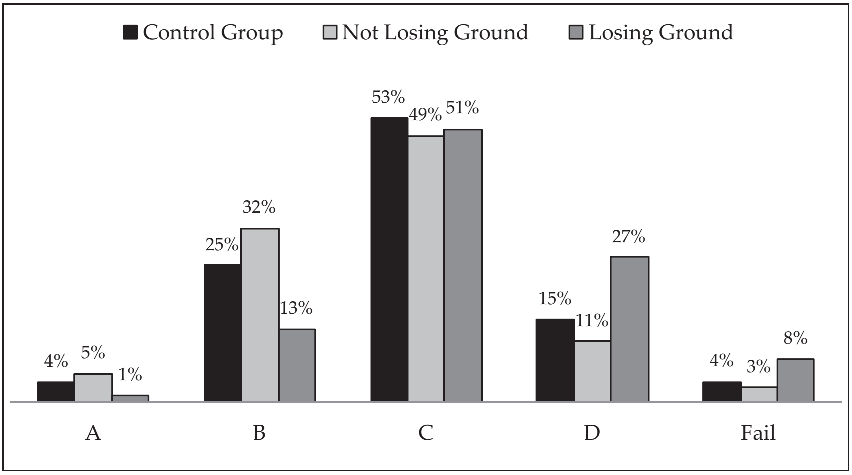

Figure 4. Predicted response probabilities for grades awarded to public schools "in the nation as a whole," from the model reported in Table 2.

and GSS items on rating the quality of schools and how they are run. In combination, the results suggest that leadership is one perceived weakness of current public schooling in the United States but also that overall grades for schools are based on additional perceived weaknesses about which the CNSS does not ask.

The final model in the fourth row of Table 2 assesses support for using additional money to improve the nation's education system. Here, the patterns are different, which reveals the utility of this design and the associated model for this application. For the "losing ground" treatment subgroup, there is no average response difference relative to the control group, given the coefficient of –.03. Instead, and unlike for the prior three questions, a negative coefficient of –.53, with a standard error of .15, applies instead to the "not losing ground" treatment subgroup. As shown in Figure 6, only 49 percent of this "not losing ground" treatment subgroup felt that "too little money" was being spent to improve the nation's education system, in comparison with 62 percent of respondents in the control group. In contrast, the "losing ground" treatment subgroup had a pattern of responses that is almost indistinguishable from that of the control group in Figure 6. In combination, these results imply that the negative treatment effect presented earlier in Table 1 for this question is produced entirely by the "not losing ground" members

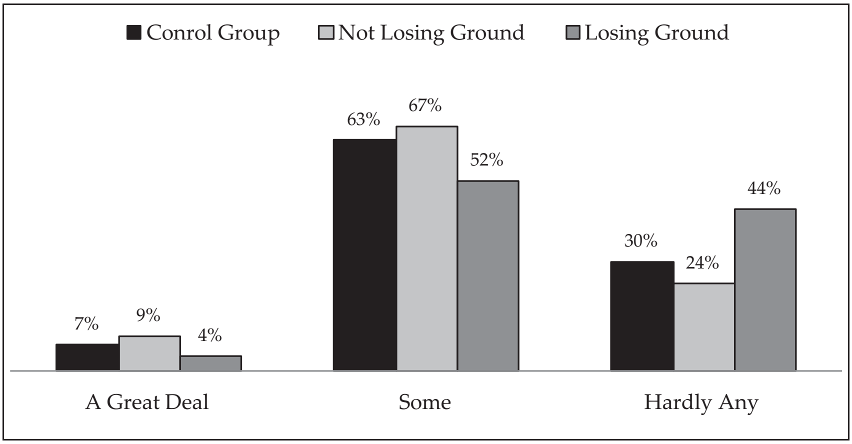

Figure 5. Predicted response probabilities for confidence in "people running the public education system," from the model reported in Table 2.

of the treatment group, which is the opposite of the pattern for the other three outcome variables.

3.2.3. *Interpretation.* Extending the warranted causal interpretation on the basis of the results reported in Table 1, there are two ways to interpret the additional insight offered by the elaborated models reported in Table 2. First, from an experimental design perspective, these elaborated models offer a consistency check, while deepening the account of why the effects presented in Table 1 were produced. The "losing ground" treatment subgroup is responsible for the overall negative treatment effect on the quality ratings and leadership of schools, but the "not losing ground" treatment subgroup is responsible for the overall negative treatment effect on support for spending to improve the nation's schools. Second, from a population polling perspective, these elaborated models suggest that the second priming question separates the treatment group into those who retrieve stored information that is in agreement with past exposure to the frame and those who do not.

What pattern of information retrieval could generate the results that differ between the first three questions and the fourth question? Before we offer an answer, we must first learn more about the pattern of responses among treatment group members that generates the distribution of L and then determine whether results similar to those presented in Table 2 persist after adjustment for X.

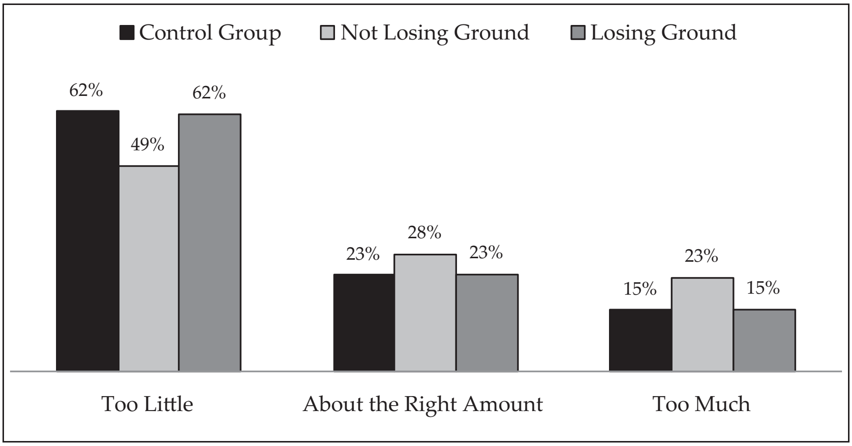

Figure 6. Predicted response probabilities for opinions on current spending "to improve the nation's education system," from the model reported in Table 2.

3.3. *A Response Heterogeneity Analysis with Conditioning on Observed Confounders*

Recall that for our discussion of Figure 2b, we noted that some of the determinants of L have been observed as X. Models that adjust for X may generate stronger interpretations. The key, as we show in this section, is to use the observed variables in X to weight the control group so that it can serve, sequentially, as a comparison group for each of the two treatment subgroups.[14]

3.3.1. *Identification.* The treatment subgroup coefficients reported in Table 2 are ordered logit–based summaries of differences:

$$P[Y|T=1, L=1] \ v. \ P[Y|T=0] \tag{9}$$

and

$$P[Y|T=1, L=0] \ v. \ P[Y|T=0]. \tag{10}$$

It would be incautious to claim that sample analogs to equations (9) and (10) identify the causal effects defined in equations (5) and (6) (or equations 3 and 4). There is no good reason to assume that $P[Y^0] = P[Y^0|L=1] = P[Y^0|L=0]$.[15] As a result, even though $P[Y|T=0]$ is consistent for $P[Y^0]$, this result provides no justification for regarding $P[Y|T=0]$ as consistent for either $P[Y^0|L=1]$ or $P[Y^0|L=0]$.

For the coefficients reported in Table 2, the primary threat to a causal interpretation on the basis of frame-specific information retrieval is the possibility that the "losing ground" partition is a proxy for underlying differences in unrelated characteristics and attitudes across members of the treatment group. Table 3 begins to assess this concern, presenting mean differences for individual characteristics measured in the CNSS for the two treatment subgroups. "Losing ground" treatment respondents were slightly more likely to have children currently attending public schools in their communities (27 percent vs. 24 percent) but were very similar on demographic characteristics, with the largest difference observed for the percentage African American (10 percent vs. 14 percent). Differences in home ownership, having taken the survey via cell phone, level of completed education, party identification, and self-labeled level of ideological conservatism were very small. Differences in family income were slightly larger (.07 on the log scale, which is commonly interpreted as a 7 percent difference in family income).

Table 3 also shows that other attitudes toward specific world affairs differed across the two treatment subgroups. In a later topical module in the CNSS, respondents were asked, "Do you agree or disagree with the statement 'The U.S. needs to play an active role in solving conflicts around the world'?" The "losing ground" treatment subgroup was considerably more likely to disagree (58 percent vs. 41 percent). In the same module, respondents were also asked,

> Some people believe that the war in Afghanistan will make America safer, while others believe that the war will not make America safer. To what extent do you agree with the following statement: "The war in Afghanistan will make America safer"?

Again, the "losing ground" treatment subgroup was more likely to disagree (70 percent vs. 60 percent).[16] Responses to these two items suggest that the "losing ground" treatment subgroup was less likely to favor active involvement of the United States in world affairs, even though the respondents in this subgroup were not more likely to self-identify with a particular political party or ideology.

Overall, then, because the only substantial measured difference between the two treatment subgroups is in attitudes toward world affairs, it is hard to make the case that these two subgroups are very different on the distribution of X. Still, the two subgroups are not identical with respect to the characteristics and attitudes reported in Table 3, and it is

Table 3. Mean Values for Additional Variables across the Two Subgroups within the Treatment Group

Variable	Treatment and Not Losing Much Ground	Treatment and Losing "Quite a Bit" or "a Great Deal" of Ground
Has children currently in public schools in the community	.24	.27
Demographic characteristics		
Female	.49	.52
Hispanic ethnicity	.06	.07
African American	.10	.14
Born in the United States	.93	.92
Age (years)	49.99	49.46
Residential characteristics		
Respondent owns home in which she or he lives	.75	.73
Question: "Do you own or rent the place where you live now?"		
Took interview on cell phone rather than land line	.31	.30
Socioeconomic status		
Family income from all sources (natural logarithm)	8.66	8.73
Education (in years completed)	14.90	14.82
Political affiliations and values		
Republican party identification	3.98	3.90
Seven-point scale with poles "strong Democrat" to "strong Republican" in response to the question "Generally speaking, when it comes to political parties in the United States, how would you best describe yourself?"		
Conservative ideology	4.15	4.12

(continued)

Table 3. (continued)

Variable	Treatment and Not Losing Much Ground	Treatment and Losing "Quite a Bit" or "a Great Deal" of Ground
Seven-point scale as responses to the question "When it comes to social issues, do you usually think of yourself as extremely liberal, liberal, slightly liberal, moderate or middle of the road, slightly conservative, conservative, or extremely conservative?"		
Attitudes toward engagement in world affairs		
Does not agree with interventions to solve conflicts around the world	.41	.58
Respondent expresses disagreement in response to the question "Do you agree or disagree with the statement 'The U.S. needs to play an active role in solving conflicts around the world'?"		
Does not feel the war in Afghanistan makes America safer	.60	.70
Respondent expresses disagreement when asked: "Some people believe that the war in Afghanistan will make America safer, while others believe that the war will not make America safer. To what extent do you agree with the following statement: 'The war in Afghanistan will make America safer'?"		

Source: Data from Cornell National Social Survey, 2011.

Note: The total Cornell National Social Survey sample size is 1,000 respondents, but this table presents results only for the 471 individuals in the treatment group. A small amount of missing data on these covariates was imputed with best-subset linear and logistic regression models.

possible that these small differences are nonetheless important for the pattern of responses to the attitude questions.

Modeling the consequences of these differences is possible because all of the variables in Table 3 are measured in the control group as well, even though the variable L is not observed for the control group. The first step is to model the propensity for members of the treatment group to indicate that public schools in the United States are losing ground, as predicted by X. The second step is to then use the estimated coefficients from the model estimated for the treatment group to generate weights that can be used for the control group, invoking a propensity score weighting rationale. These weights can then be used to weight all members of the control group in two ways: first to align the control group with the distribution of X that is observed for the "losing ground" treatment subgroup and then for the "not losing ground" treatment subgroup.

3.3.2. *Results.* As reported in Table A1 in the Appendix, we estimated a logit model in the treatment group with L as the outcome variable and all of the variables reported in the rows of Table 3 as the predictor variables. This model is not strongly predictive. Most coefficients are smaller in magnitude than their standard errors, but the "intervention to solve conflicts" variable has a coefficient that suggests a moderately strong association. And, net of other characteristics in the model, family income has a small positive coefficient that is statistically significant. The model chi-square value is only 24.3, with 14 degrees of freedom, which only narrowly exceeds the relevant .05 critical value of 23.7. These results are not surprising; we showed already in Table 3 that the two treatment subgroups are very similar with respect to the distribution of X.

Table 4 presents treatment subgroup coefficients analogous to those in Table 2 but which have been adjusted for differences in X, as reported in Table 3. For each of the four outcomes, the two treatment subgroup coefficients are estimated as doubly robust inverse probability weighted estimates for treatment subgroup–specific treatment effects. In particular, we take the specification from the logit model reported in Table A1 and use it to calculate estimated weights that can balance the distributions of the variables in Table 3 across the control group and each of the treatment subgroups (see Imbens 2004 and Morgan and Winship 2015, chap. 7, for details of the method). In essence, these two sets of weights, when applied to the control group, reweight control group members such that they represent each treatment subgroup with respect to X. We then reestimated the four models in Table 2, once for each set of weights, while including all of the variables in X as covariates to protect against misspecification of the logit model that estimated the weights. For Table 4, we then present the treatment subgroup coefficients from the models that use the relevant subgroup-specific weights, as well as model-specific chi-square values that correspond to the model from which the coefficient is drawn. Overall, the eight estimated coefficients reported in Table 4 are very similar to the corresponding coefficients reported in Table 2.

3.3.3. *Interpretation.* The results presented in Table 4 imply that differences in responses to the PDK/GP and GSS questions across the two treatment subgroups cannot be explained away by X. There is simply no basis for concluding that the response heterogeneity within the treatment group can be attributed to whether respondents have children in their local public schools, demographic characteristics, socioeconomic status, political party identification, self-rated conservatism, or even attitudes toward the appropriateness of foreign engagement by the government of the United States. These are the obvious sources of observable differences between these groups, and these results give us additional confidence that the response heterogeneity across the two treatment subgroups is specific to the substance of the frame.

3.4. *Discussion of the Demonstration*

To build the case for the value of the design and the modeling that it enables, we conclude the demonstration with a discussion of the deeper interpretation that has been generated. The two-question international competitiveness prime causes respondents, on average in a nationally representative survey, to lower their subjective assessments of the quality of local schooling while decreasing support for additional spending to improve the nation's education system. However, the analysis offered here suggests that it would be unwise to assume that the same individuals move in response to the treatment prime for all four of the outcome questions.

Our first set of results—for which we ignored the available partition in the treatment group—is not incorrect, but because it offers no information to the contrary, it implies that all members of the treatment group responded similarly to the two priming questions when

Table 4. Treatment Group by Losing Ground Coefficients from Ordered Logit Models for Each Outcome Question, with the Control Group Weighted Alternatively by Other Covariates

Question	Treatment Group by Not Losing Ground Coefficient (*SE*)	*N*	Chi-square Test Statistic (*df*)
Grades for public schools "in your community"	−.16 (.15)	928	75.0 (16)
Grades for public schools "in the nation as a whole"	.39 (.16)	926	77.2 (16)
Confidence in "people running the public education system"	.29 (.17)	971	69.2 (16)
Support for spending "to improve the nation's education system"	−.55 (.16)	968	148.6 (16)
	Treatment group by losing ground coefficient (*SE*)	*N*	Chi-square test statistic (*df*)
Grades for public schools "in your community"	−.88 (.16)	928	75.8 (16)
Grades for public schools "in the nation as a whole"	−.85 (.16)	926	100.0 (16)
Confidence in "people running the public education system"	−.60 (.16)	971	61.9 (16)
Support for spending "to improve the nation's education system"	−.09 (.18)	968	128.8 (16)

Source: Data from Cornell National Social Survey, 2011.

Note: For grades, the highest response category is A, and the lowest response category is fail. For the confidence question, the highest response category is "A great deal of confidence in them," and the lowest response category is "Hardly any confidence at all in them," with "Some confidence in them" as the middle category. For the spending question, the highest response category is "Too little money," and the lowest response category is "Too much money," with "About the right amount" as the middle category. All models are weighted by the inverse probability of providing a response to the outcome question, as estimated by a supplementary logit model.

formulating responses to the four outcome questions. In contrast, the results from our subgroup-level treatment effects analysis suggest that there are two subgroups within the treatment group whose average responses differed from each other. This suggests two complementary narratives for two distinct groups of individuals in the population. First, respondents who believe that public schools in the United States are losing substantial ground to those of international competitors offer lower grades for schools but continue to show interest in spending additional resources to improve them at the same level as the population as a whole. Second, respondents who believe that schools in the United States are not losing substantial ground to those of international competitors offer slightly higher grades but then express less support for increasing funding to improve them, again in comparison to response patterns in the population as a whole. This variation clarifies and extends the conclusions in Morgan and Taylor Poppe (2012), demonstrating the utility of a design that pairs a representative sample with a facilitative prime.

Our final set of models demonstrates how the control group can be weighted using estimated propensity scores to generate distinct comparison groups for the two alternative treatment subgroups. These models allow us to rule out response heterogeneity that could have been produced by other observed variables.

What produces these effects? One model for interpreting a context effect, and the one which we favor, is the belief-sampling model of Tourangeau et al. (2000). Here, the context effect is assumed to emerge because the context-setting treatment prime generates the retrieval of information, stored as personal beliefs, that is relevant to responses to the four outcome questions. The CNSS experiment, like nearly all other context effect experiments, does not reveal the specific stored beliefs that are retrieved and thereby made more salient as subsequent questions are interpreted and answered. However, the structure of the treatment prime offers a fairly straightforward set of conclusions nonetheless.

"Losing ground" treatment subgroup respondents are retrieving beliefs shaped by the statements of political elites (candidates for election, authoritative feature journalists, op-ed columnists, etc.) that public schools in the United States are performing below desired levels and falling behind the schools of our international competitors. As a consequence, they then offer lower grades for schools but continue to show interest in spending additional resources to improve them at the same level as the population as a whole.

"Not losing ground" treatment subgroup respondents are explicitly not retrieving this same set of beliefs. They either have not been exposed to these statements of political elites on the flagging performance of schools in the United States, or they have reasons to reject those statements. But why would these respondents offer slightly higher grades to schools than the control group? One answer is that the control group almost certainly includes some members of the population who are aware of the frame and retrieve the belief that schools are losing ground even when only presented with the first PDK/GP question that asks them to grade the schools in their communities, unlike the treatment subgroup that is composed only of individuals who explicitly reject the frame of interest.

The evidence also suggests that these respondents are retrieving beliefs that then prompt them to express less support for increasing funding to improve the nation's education system, again in comparison with response patterns in the population as a whole that are estimated by the control group. The most straightforward interpretation of this pattern is that these respondents are retrieving beliefs on the basis of the statements of political elites that the United States has robust economic competitors, perhaps more so than it did in prior decades. Because these respondents do not believe that shortfalls in the public education system of the United States are leaving the country vulnerable to competitors, some of these respondents may believe that money is better spent elsewhere shoring up whatever institutions they believe leave the country most vulnerable. Some respondents, for example, may believe that the existing federal deficit of the United States is the primary threat to the nation's competitiveness. These respondents, if primed to think about economic threats, may favor spending less money in general on all national priorities, regardless of the need for school reform or for addressing other social problems.

4. CONCLUSION

Context and framing effects in surveys are pervasive and variable across respondents. Because they are assumed to arise from unobserved interactions with beliefs stored in the memories of individuals with different prior experiences, the context-setting triggers embedded in survey instruments cannot be assumed to generate constant effects across respondents. In this article, we have demonstrated that a survey experiment, when paired with a facilitative prime, can enable models of variable context effects and genuine response heterogeneity at the population level.

As we noted above, none of the components of the design we have demonstrated is novel on its own, but their joint adoption allows for an analysis of conventional treatment effects as well as the patterns of response heterogeneity that underlie them. We have also shown how this heterogeneity can be modeled in pursuit of subgroup-level causal effect estimates, even though formal identification of these effects cannot be achieved.

It is a truism that all designs have limitations of some form; this design has three sets of weaknesses. First, it is impossible to conclude that the context effects revealed by the design would occur if similar facilitative primes were inserted into the national surveys from which the outcome questions are drawn. For example, it is impossible to exactly mimic the administration of the PDK/GP and the GSS at the same time, and in our demonstration, there are important differences between the CNSS and these surveys. The GSS is a face-to-face survey and offers an available Spanish-language questionnaire. The PDK/GP is a telephone survey, but it is based on the standing Gallup Panel. Thus, response and cooperation rates, as well as mode of administration, differ across these surveys. In addition, each survey has its own set of context effects, which could not be replicated for the CNSS. In particular, the spending question in the GSS is at the beginning of the survey and asked of everyone, but the confidence in leaders question has most recently been asked on two of the three different GSS ballots, and preceded by slightly different sets of questions based on the ballot. Thus, although it is a strength of our demonstration that it uses questions on which several decades of survey data are available, it is also the case that it is impossible to state definitively that the treatment prime analyzed here would produce analogous context effects if it were inserted into the questionnaires of these two long-running surveys. At most, we can conclude that there is a strong likelihood that such effects would emerge.

Second, the design itself has some inherent weaknesses, as judged relative to alternative designs in the substantive framing literature. In defending the value of student samples (and other convenience samples), Druckman and Kam (2011) privileged a broad criterion of external validity, stating "External validity refers to generalization not only of individuals, but also across settings/contexts, times, and operationalizations" (pp. 42–43). Although one might argue that such a broad definition is an attempt to paper over the very limited capacities of convenience samples to sustain narrower definitions of external validity, it is still the case that demonstrations such as the one offered here are based on a single operationalization, undertaken at a single point in time, and in an artificial interview context over which we, as investigators, have limited control. Carefully designed experimental studies, such as Druckman, Fein, and Leeper (2012), can generate results from repeated administrations that uncover the evolution of opinions in response to rich information sources. Such studies may still poorly mimic the true processes by which individuals form opinions in their nonexperimental lives, but it is without question that such artful and carefully controlled studies have some advantages that cannot be easily accommodated in this design.

Third, our favored interpretation is just that: an interpretation. One alternative interpretation, which is a plausible alternative, is that the facilitative prime only triggers an anchoring point for subsequent judgments. For this interpretation, respondents do not make any connection between economic competitiveness and school quality, regardless of the content of the statements that elites have offered in public and regardless of the content of the questions that explicitly prime economic competitiveness. Instead, respondents change only the reference point for their judgments and invoke an international standard as an anchoring point that, for whatever reason, causes them to lower the grades they offer to schools as well as their confidence in leaders. We believe that this alternative interpretation is less persuasive than our favored one, because we do not see how it generates the "losing ground" partition that then generates a particular pattern of variation across all four subsequent questions. In particular, it does not suggest as natural an interpretation for why the "not losing ground" treatment subgroup lowers its interest in spending money to improve the nation's schools. Even so, the larger point we wish to make here is that an interpretation must still be offered after results are generated using this design, and more than one interpretation will almost certainly be plausible. The design does not reveal which specific beliefs, if any, have been retrieved in response to the facilitative prime, even though it induces individual-level variation in response to the content of the prime that can help motivate alternative interpretations.

These limitations notwithstanding, the design demonstrated here does offer potential substantive insight into framing effects at the population level, demonstrating their relevance for responses to long-used items in important surveys. Moreover, the design's facilitative prime allows subjects to sort themselves in ways that the analyst can plausibly assume reflect real-world information storage and belief formation in response to frame exposure prior to the study. As such, even though the study does not come close to mimicking real-world frame exposure and issue motivation processes, it does offer the analyst leverage to determine which respondents are most likely to have observed and stored context-setting information in response to frame presentation by political elites and others. The analyst can then offer estimates of the proportion of the target population that has likely stored information in response to the frame prior to the study as well as the effect that this information, when retrieved, has on responses to relevant attitude items. When the items under study are long-used questions from important public opinion surveys and polls, such results offer more than just methodological insight into survey response artifacts. They offer substantive insight into how the preexisting beliefs of respondents shape their attitudes, in variable patterns within the target population. Results such as these can move models of context effects from studies that demonstrate their existence toward those that model their prevalence and magnitude. Such results have the potential to inform substantive research and also prompt the augmentation of survey instruments to directly measure any inferred heterogeneity.

APPENDIX

Table A1. Coefficients from a Logit Model That Predicts Whether Members of the Treatment Group Believe That the U.S. Education System Is Losing Substantial Ground

Variable	
Constant	−2.70
Has children currently in public schools in the community	.07 (.23)
Demographic characteristics	
Female	.09 (.20)
Hispanic ethnicity	.30 (.40)
African American	.51 (.33)
Born in the United States	.04 (.36)
Age	−.001 (.01)
Residential characteristics	
Respondent owns home in which she or he lives	−.34 (.26)
Took interview on cell phone rather than land line	−.15 (.22)
Socioeconomic status	
Family income from all sources (natural logarithm)	.28 (.14)
Education (in years completed)	−.02 (.05)
Political affiliations and values	
Republican party identification	.02 (.06)
Conservative ideology	−.001 (.07)
Attitudes toward engagement in world affairs	
Does not agree with interventions to solve conflicts around the world	.69 (.21)
Does not feel the war in Afghanistan makes America safer	.32 (.22)
N	471
Chi-square (*df*)	24.3 (14)

Source: Data from Cornell National Social Survey, 2011.

Acknowledgments

We thank the team at the Survey Research Institute at Cornell University for fielding the CNSS. We also thank the editor and anonymous reviewers for their helpful suggestions.

Notes

1. In this literature, the conceptual distinctions between the terms *framing* and *priming* have been a matter of discussion and debate; see Entman (1993) for an early review and Druckman, Kuklinski, and Sigelman (2009) for a later review. In this article, we do not make fine distinctions between frames and primes.
2. Morgan and Taylor Poppe (2012) considered parts 1 and 4 of this question, and they analyzed the same Cornell National Social Survey experiment but used only models for conventional treatment effects. In this article, we consider all four parts

of the question and extend the modeling strategy to fully reveal the pattern of results that generated their conclusions.

3. The sample was provided by Marketing Systems Group as a random-digit dialing list of telephone numbers drawn from telephone exchanges in the continental United States (including cell phones but excluding known nonhousehold numbers). Within contacted households, one respondent from each household was selected using a "most recent birthday" selection rule. Telephone data collection by the Survey Research Institute at Cornell University began on September 10, 2011, and was completed by December 10, 2011. All interviews were conducted in English, and the cooperation and response rates were 62.4 percent and 24.1 percent, respectively (calculated using definition 2 of the American Association for Public Opinion Research).
4. Although the PDK/GP questions are verbatim copies of their originals (see Bushaw and Lopez 2011, 2012), time constraints on the CNSS required changes to the two GSS questions. The first question asks only about confidence in leaders "running the public education system in the United States," rather than offering a menu of types of leaders associated with particular institutions in a battery of questions. The second question asks only about spending on education, not spending across the full menu of items on the GSS. These changes may generate context effects of their own, as we discuss in the final section of this article when detailing limitations of the design.
5. Both ballots conclude with the same question that asks respondents to indicate whether they currently have children attending public schools in their own communities. The PDK/GP asks a similar question as this last one, and reports based on these data suggest that respondents with children currently enrolled in school award higher grades to schools in their own communities, presumably based either on current information to which they have access or a more diffuse loyalty to the institutions that care for their children.
6. Other primes could also be used that are visual but not textual. The best example in this context would be a brief video of Mitt Romney's opening statement in the October 3, 2012, presidential debate, in which he offered a plan for economic growth: "My plan has five basic parts. One, get us energy independent, North American energy independent. That creates about 4 million jobs. Number two, open up more trade, particularly in Latin America, crack down on China, if and when they cheat. Number three, make sure our people have the skills they need to succeed and the best schools in the world. We're a far way from that now. Number four, get us to a balanced budget. Number five, champion small business. It's small business that creates the jobs in America." Note that Romney's remark on schools is sandwiched between statements that prime economic competitiveness with China and the debt-funded spending of the federal government. It was delivered to a television audience widely estimated to include at least 70 million viewers, during a performance that is regarded as the best two hours of his campaign. It is authoritative (although partisan), and it has the benefit of being real, short, and closer to a facilitative prime. Yet, like the *New York Times* article, it would be difficult to administer this prime to subjects selected as part of a national sample.

Respondent burden would be lower than for the *New York Times* article, but survey costs would remain high.

7. As noted earlier, the GSS has used split-ballot designs in the past to examine context effects, including for these questions. For the question that asks respondents to rate a series of national spending priorities ("I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount"), respondents were randomly assigned to one of two question wordings for the spending areas, generally in the pattern of "improving the nation's education system" (the original GSS wording used from 1973 through 2012) and "education" (a terse alternative introduced first in 1984 and used on a split ballot through 2012). Question wording effects for the education item (which is one of the questions analyzed in our experiment) were small. However, as the experiment has continued, additional power has accumulated to identify smaller effects, which are generally less than four percentage points (see Smith 1991, 2006).
8. Similar to this result, Simon and Davey (2010) offered a vignette-based framing study with an online national sample that evaluated a variety of framing strategies for generating support for higher education. Their results suggest that "a commonly advanced value in public discourse, Global Competitiveness, on the higher education level actually depresses support for progressive policy reform" (p. 3). Their vignette, however, stresses preparing the next generation of children for competition in a global economy, not the economic threat that other nations pose to the United States.
9. We could also use latent variables with hollow nodes (see, e.g., Morgan and Winship 2015) to allow the common causes represented by the double-headed arrows in $L \dashleftarrow\dashrightarrow X$ and in $X \dashleftarrow\dashrightarrow Y$ to be determined by distal common causes U. We do not offer such an elaborated graph, because the need for an explicit representation of U is vitiated by our inclusion of $L \dashleftarrow\dashrightarrow Y$, which renders the effect of L on Y unidentified, regardless of whether conditioning on X would induce additional collider-stratification bias; see Pearl (2009) and Elwert (2013).
10. An alternative representation would be to join T to L by defining a new three-valued treatment variable as $Z=0$ if $T=0$ and $Z=1+L$ if $T=1$. We chose the representation in Figure 2b because it remains consistent with Figure 2a by showing that the total effect of T on Y is still identified, while also revealing that the particular value of L is endogenous with respect to common causes that also determine both X and Y. An alternative graph that relied only on Z to represent both T and L would not reveal this consistency between Figures 2a and 2b. Relatedly, our position is that L is not a collider variable of the usual kind because it is missing entirely for the control group and is best regarded as inherent to the treatment conditions. By using L to partition the treatment group only, we are not generating induced nuisance associations between T and Y across the full population, as would be the case if we conditioned on L in the control group as well. Instead, we attach substantive interpretations to partition-defined treatment effects, even though some of the same reasoning for understanding collider-induced associations does obtain. As explained below, we give interpretations to the contrasts that we feel are substantively justified.

11. Alternative designs would allow for a genuine separation of T from L, but with associated costs that would not aid in identification. For example, the two priming questions could be asked of the control group after the four attitude questions, as in a question-order experiment. This design would generate values for L for the control group. However, we would then have a context effect whereby beliefs about the quality of schools may then exert their own effects on L. Any attempt to draw a full causal graph would then include cycles, such as $L \rightarrow Y \rightarrow L$. Without introducing further assumptions—such as assuming that the effect of Y on L is null—this alternative design would simply have twice as many unidentified effects.
12. See also Table S3 in the online Supplementary Appendix for the specific values depicted graphically in Figures 2 through 5.
13. This 35 percentage point difference between grades for schools in respondents' communities and in the nation as a whole is comparable with the differences reported in the PDK/GP (33 points in 2011 and 29 points in 2012; see Bushaw and Lopez 2011, 2012). The scale of these lower grades, however, is higher by about 10 percentage points in the PDK/GP, which may reflect a specific negative context effect in the PDK/GP. The questions on grades are the third through fifth questions on the PDK/GP poll (W. J. Bushaw, personal communication). The first question filters respondents on the basis of whether they have children in the local public schools. The second question, which may generate a negative context effect, asks respondents, "What do you think are the biggest problems that the public schools of your community must deal with?" It is unclear what the response categories to this question have been over the years, but Bushaw and Lopez (2012:10) reported that 43 percent of respondents indicated "lack of financial support," while another 16 percent listed "lack of discipline,""overcrowded schools,""fighting/gang violence," or "drugs."
14. In effect, we use the information in X to impute the nonrandom missing data in L for the control group, on the basis of the relationship between X and L in the treatment group. We do not actually impute any data. Instead, we reweight the control group so that its distribution of X matches, as closely as possible, the distribution of X for each treatment subgroup for each relevant model. When estimating subgroup-level casual effects, this is equivalent to imputing L in a way that would partition the control group according to X but is more efficient because it uses all control cases for each subsequent weighted model.
15. In other words, we can decompose $P[Y^1]$ using only observed quantities, such that $P[Y^1] = \lambda P[Y|T=1, L=1] + (1-\lambda)P[Y|T=1, L=0]$, where λ is the proportion of the treatment group that sees schools as losing ground. However, we have no partition of the control group by L and therefore we cannot similarly decompose $P[Y^0]$ across L using observable quantities only. In particular, we have no way to estimate either $P[Y^0|L=1]$ or $P[Y^0|L=0]$ with the observed data, even though we can borrow the estimate of λ from the treatment group because both the treatment and control groups are representative samples from the same target population. We observe no data for sample analogs to either $P[Y|T=0, L=1]$ or $P[Y|T=0, L=0]$.
16. Although not impossible, it is unlikely that the treatment prime altered these responses. The education module was the first module in the CNSS interview

while the world affairs module was separated from it by 25 other questions on a variety of topics.

References

Bushaw, William J., and Shane J. Lopez. 2011. "Betting on Teachers: The 43rd Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes toward the Public Schools." *Phi Delta Kappan* 93:8–26.

Bushaw, William J., and Shane J. Lopez. 2012. "Public Education in the United States: A Nation Divided. The 44th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes toward the Public Schools." *Phi Delta Kappan* 94:9–25.

Chong, Dennis, and James N. Druckman. 2007. "Framing Theory." *Annual Review of Political Science* 10:103–26.

Chong, Dennis, and James N. Druckman. 2011. "Public-elite Interactions: Puzzles in Search of Researchers." Pp. 170–88 in *The Oxford Handbook of American Public Opinion and the Media*, edited by R. Y. Shapiro and L. R. Jacobs. New York: Oxford University Press.

Druckman, James N., Jordan Fein, and Thomas J. Leeper. 2012. "A Source of Bias in Public Opinion Stability." *American Political Science Review* 106:430–54.

Druckman, James N., Green, Donald P., Kuklinski, James H., and Arthur Lupia, eds. 2011. *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.

Druckman, James N., and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base.'" Pp. 41–57 in *Cambridge Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. New York: Cambridge University Press.

Druckman, James N., James H. Kuklinski, and Lee Sigelman. 2009. "The Unmet Potential of Interdisciplinary Research: Political Psychological Approaches to Voting and Public Opinion." *Political Behavior* 31:485–510.

Elwert, Felix. 2013. "Graphical Causal Models." Pp. 245–73 in *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan. Dordrecht, the Netherlands: Springer.

Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43:51–58.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86:4–29.

Morgan, Stephen L., and Emily S. Taylor Poppe. 2012. "The Consequences of International Comparisons for Public Support of K–12 Education: Evidence from a National Survey Experiment." *Educational Researcher* 42:262–68.

Morgan, Stephen L., and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge, UK: Cambridge University Press.

Mutz, Diana C. 2011. *Population-based Survey Experiments*. Princeton, NJ: Princeton University Press.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge, UK: Cambridge University Press.

Pearl, Judea. 2010. "The Foundations of Causal Inference." Pp. 75–149 in *Sociological Methodology*, Vol. 40, edited by Tim Futing Liao. Boston, MA: Wiley-Blackwell.

Rasinski, Kenneth A. 1988. "The Effect of Question Wording on Public Support for Government Spending." GSS Methodological Report No. 54. Chicago, IL: National Opinion Research Center.

Rasinski, Kenneth A. 1989. "The Effect of Question Wording on Public Support for Government Spending." *Public Opinion Quarterly* 53:388–94.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:322–31.

Schuman, Howard. 2008. *Method and Meaning in Polls and Surveys*. Cambridge, MA: Harvard University Press.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. New York: Academic Press.

Schwarz, Norbert, and Seymour Sudman, eds. 1992. *Context Effects in Social and Psychological Research*. New York: Springer-Verlag.

Simon, Adam F., and Lynn F. Davey. 2010. "College Bound: The Effects of Value Frames on Attitudes toward Higher Education Reform." A Frameworks Research Report. Washington, DC: The Frameworks Institute.

Smith, Tom W. 1987. "That Which We Call Welfare by Any Other Name Would Smell Sweeter: An Analysis of Question Wording on Response Patterns." *Public Opinion Quarterly* 51:75–83.

Smith, Tom W. 1991. "Context Effects in the General Social Survey." Pp. 57–72 in *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. New York: John Wiley.

Smith, Tom W. 2006. "Wording Effects on the National Spending Priority Items across Time, 1984–2004." GSS Methodological Report No. 107. Chicago, IL: National Opinion Research Center.

Sniderman, Paul M. 2011. "The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation." Pp. 102–14 in *Cambridge Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. New York: Cambridge University Press.

Tourangeau, Roger, Kenneth A. Rasinski, Norman Bradburn, and Roy D'Andrade. 1989. "Belief Accessibility and Context Effects in Attitude Measurement." *Journal of Experimental Social Psychology* 25:401–21.

Tourangeau, Roger, Lance J. Rips, and Kenneth A. Rasinski. 2000. *The Psychology of Survey Response*. New York: Cambridge University Press.

Wyer, Robert S., and Thomas K. Srull. 1989. *Memory and Cognition in Its Social Context*. Hillsdale, NJ: Lawrence Erlbaum.

Author Biographies

Stephen L. Morgan is the Bloomberg Distinguished Professor of Sociology and Education at Johns Hopkins University. His current areas of research include education, inequality, demography, and methodology. Along with Christopher Winship, he is the author of *Counterfactuals and Causal Inference: Methods and Principles for Social*

Research (Cambridge University Press, 2007; revised and enlarged second edition, 2015).

Emily S. Taylor Poppe is a doctoral candidate in sociology at Cornell University. Her research interests include inequality, sociology of law, and methodology. Her dissertation focuses on legal representation in residential foreclosure cases during the Great Recession.