

# Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability

Daniel J. Hopkins and Gary King

February 7, 2010

## **Abstract**

We report the results of several randomized survey experiments designed to evaluate two intended improvements to anchoring vignettes, an increasingly common technique used to improve interpersonal comparability in survey research. This technique asks for respondent self-assessments followed by assessments of hypothetical people described in vignettes. Variation in assessments of the vignettes across respondents reveals interpersonal incomparability, allowing researchers to improve comparability by rescaling self-assessments relative to vignette responses. Our experiments show, first, that switching the question order so that self-assessments follow the vignettes primes respondents to define the response scale in a common way. In this case, priming is not a bias to avoid but a means of better communicating the question's meaning. Second, we demonstrate that combining vignettes and self-assessments in a single direct comparison induces inconsistent and considerably less informative responses. Since similar combined strategies are widely employed for related purposes, our results suggest that anchoring vignettes could reduce measurement error in many applications where they are not currently used. Data for our experiments come from a national telephone survey and a separate on-line survey.

## Authors' Note

Daniel J. Hopkins is an Assistant Professor of Government at Georgetown University, Washington, DC. Address correspondence to Daniel Hopkins, Department of Government (681 ICC), Georgetown University, Washington, DC, 20057. Email: dh335@georgetown.edu. Webpage: [www.danhopkins.org](http://www.danhopkins.org). Phone: (202) 687-5865. Fax: (202) 687-5858.

Gary King is the Albert J. Weatherhead III University Professor at Harvard University in Cambridge, Massachusetts. Email: [King@harvard.edu](mailto:King@harvard.edu). Webpage: <http://gking.harvard.edu>. Phone: (617) 495-2027. Fax: (617) 821-8581.

## Acknowledgements

Our thanks to Liam Delaney, Chase Harrison, Kristin Javaras, Katherine McCabe, and the *Public Opinion Quarterly* editors and reviewers for many helpful comments. This paper presents data collected by the Center for Survey Research at Indiana University with funding from Time-sharing Experiments in the Social Sciences [NSF Grant 0818839, Arthur Lupia and Diana Mutz, Principal Investigators]. It also presents survey data collected through Polimetrix (now YouGov/Polimetrix). The survey instrument was approved by Harvard University's Committee on the Use of Human Subjects [F11939-101]. Daniel Hopkins gratefully acknowledges institutional support from Yale University's Center for the Study of American Politics and MIT's Department of Political Science. We are grateful to those institutions for the support, assistance, and feedback which made this research possible.

# 1 Introduction

Large, cross-national surveys have become a central tool for social scientists and policymakers, from the European Social Survey and the Afrobarometer to the World Values Survey and the World Health Survey. Yet residents of different countries, and even respondents within a country, often understand the same survey question differently. That may be especially true for the more abstract concepts of interest to social scientists such as political efficacy or economic class. A person who is clearly “middle class” to one respondent could be “upper class” to another and “working class” to a third. This interpersonal incomparability poses a significant threat to the conclusions drawn from some survey questions applied in heterogeneous populations. Seemingly important inter-group differences in survey responses could, in fact, reflect differences in question interpretation. Drawing on educational testing research (Holland and Wainer, 1993), we define a survey question as having interpersonal incomparability if two individuals who are equal on the underlying quantity of interest nonetheless have unequal probabilities of providing the same answer.<sup>1</sup>

Our focus here is on the interpersonal incomparability resulting from different uses of response scales. To attempt to ameliorate this problem, researchers have developed the technique of *anchoring vignettes* (King et al., 2004; King and Wand, 2007). After a standard self-assessment question, the survey respondent learns about hypothetical individuals through brief vignettes and is asked to place those individuals on the same response scale. Variation in vignette responses across individuals reveals interpersonal incomparability and enables researchers to use one of several statistical techniques to rescale the respondent’s own self-assessment. When applied correctly, the technique can greatly increase the comparability in survey responses. Section 2 describes anchoring vignettes, the properties of the resulting incomparability-corrected measure, and the associated statistical methods in more detail. Anchoring vignettes have been implemented in recent research projects in many countries, including state effectiveness in Eastern Europe (Grzymala-Busse, 2007),

job satisfaction in EU countries (Kristensen and Johansson, 2008), community strength in the U.S. (Buckley, 2008), workplace disability in the Netherlands and the U.S. (Kapteyn, Smith and Soest, 2007), problem drinking in Ireland (Soest et al., 2007), and health outcomes throughout the world (Damacena, Vasconcellos and Szwarcwald, 2005; King et al., 2004; Salomon, Tandon and Murray, 2004).<sup>2</sup>

Improving the statistical methods for analyzing anchoring vignette data has attracted considerable scholarly attention (Gupta, Kristensen and Pozzoli, 2008; Javaras and Ripley, 2007; King and Wand, 2007; Soest et al., 2007; Wand, 2007; King et al., 2004), but less work has been devoted to more basic questions of survey administration (e.g. Krosnick, 1999; Schuman and Presser, 1996; Sudman, Bradburn and Schwarz, 1996; Strack, 1991; Tourangeau, 1991), which may also improve the technique (Buckley, 2008). We therefore address two potential improvements of this type for anchoring vignettes, involving question order and question wording. In the first, we ask whether we might generate survey responses with less measurement error if the self-assessment were moved from before to after the vignettes. In the second, we study whether we can extract more information less expensively by asking respondents to compare themselves directly to hypothetical individuals rather than using vignette assessments and a separate self-assessment.

We study these issues via randomized survey experiments embedded in both phone and Internet surveys (to control for mode effects; Tourangeau 2004) with a total of 2,116 respondents and spanning diverse examples that cover external political efficacy, socioeconomic status, and rest. In Section 3, we consider the placement of the self-assessment question within the questionnaire. One might worry that asking self-assessments after the anchoring vignettes will produce priming or other question-order biases (Buckley, 2008). This concern seems to have motivated the 2002 World Health Survey as well as subsequent applications to ask the self-assessment question first. Our idea is to use priming intentionally, in the expectation that exposure to the vignettes prior to the self-assessment will lead respondents to understand the underlying concept in the way intended by the researcher. After hearing or reading the vignettes, respondents might better understand the

specific attribute being assessed as well as the attribute's range. Section 3.4 shows that the relationship between the vignette-corrected responses and related independent variables is stronger when respondents were first primed by the vignettes. We thus conclude, first, that researchers should change current practice and ask self-assessment questions immediately after the vignette battery.

In Section 4, we examine whether anchoring vignettes might be replaced with “direct comparisons,” where respondents to each question compare themselves directly to hypothetical individuals who represent different levels of the variable of interest (e.g., “Alice is concerned about cars speeding by her house, and would like to see the speed limit street reduced. However, she knows that her elected official is from another part of town, and so is unlikely to help her. Do you have more say in government, the same say, or less say than Alice?”). The rationale for such an approach seems compelling. Direct comparisons eliminate at least one survey question, since there is no need for a self-assessment separate from the vignette questions.<sup>3</sup> They also reduce costs and respondent fatigue, and might also better communicate with the respondent. Yet in practice, we show that this procedure induces its own biases, eliminating much of the information contained in the rescaled response.

Specifically, we find that direct comparisons fail due to an especially pronounced form of response-scale bias. When respondents are asked directly if they are similar to the hypothetical individual described in the vignette, respondents commonly agree that they are, even when saying so is inconsistent with their other responses, or when the vignette describes an outlier. Separating the vignettes from the self-assessment reduces bias in such cases by reducing the tendency for the respondent to report being the same as the hypothetical individual. Doing so also lets the researcher make the comparison, leaving to the respondent questions which are separate and straightforward. In cases similar to those discussed here, answering the research question requires not posing that question directly to the respondent. As the National Election Study, the General Social Survey, and other prominent surveys sometimes employ the “direct comparison” strategy for related

purposes, there appear to be more widespread uses of our results than merely to increase comparability.

## 2 Background on Anchoring Vignettes

We now offer a brief overview and summary of the anchoring vignette survey design strategy and associated statistical methods. For more detailed technical information see Wand (2007), Soest et al. (2007), King and Wand (2007), and King et al. (2004). To fix ideas, we use as a running example the measurement of political efficacy across groups. Unlike variables such as income or height, political efficacy has no single, commonly used metric. If respondents in a certain demographic group report higher levels of efficacy than those in another, it is impossible to know if the true level of efficacy is actually higher or if instead one group interpreted the question differently. For instance, in the 2002 World Health Survey, respondents in China reported significantly *higher* levels of political efficacy than did those in Mexico (King et al., 2004). Given the different objective realities in the two examples, one might suspect that Mexicans' lower efficacy is an artifact of their differing use of the response scale.

### 2.1 Design Logic

Anchoring vignettes address the inter-group incomparability resulting from different uses of the response scale. The technique uses separate assessments of one or more vignettes to recalibrate the responses given to the self-assessment questions. In this example, after being asked about their own efficacy, both Chinese and Mexican respondents were also asked to assess the efficacy of hypothetical individuals. One vignette tells respondents:

[John] lacks clean drinking water. There is a group of local leaders who could do something about the problem, but they have said that industrial development is the most important policy right now instead of clean water.

The survey then asks: “how much say [does John] have in getting the government to address issues that interest [him]?” Provided it has the same response categories as the self-assessment question, the vignette provides a common reference point which allows researchers to rescale the original response. If a Mexican respondent indicates that she herself has “some say in government,” but that John has “little say,” we know that this Mexican respondent has more say than John does. If a Chinese respondent reports having “a lot of say” but also says that John has a lot of say, we know that her efficacy is similar to John’s — and thus lower than that of the Mexican respondent. John’s level of efficacy becomes a fixed anchor on the new scale, allowing researchers to correct for inter-group incomparability by relating each respondent’s efficacy to John’s. And in fact, when researchers applied vignettes in China and Mexico, they concluded that Mexican respondents’ average level of political efficacy is higher than that of Chinese respondents (King et al., 2004).

A central assumption underpinning anchoring vignettes is *vignette equivalence*, which holds that “the level of the variable represented in the vignette is understood by all respondents in the same way” (King and Wand, 2007, p. 49), even if respondents use the response categories in different ways and thus generate incomparability. For example, one respondent might think of an individual who writes a letter to a local official as having significant input in her local government, while another might consider that individual to have little or no input. The key is that the *thresholds* that separate response categories may differ across people, but by assumption, all respondents must understand that writing a letter represents the same point on the underlying continuum, irrespective of the response category they use. Given this assumption, the vignette provides a way to understand how the respondent uses the response scale — or in this case, what the respondent means when he says he has little say in his government.

## 2.2 Analysis Methods

The simplest way to analyze anchoring vignette data formalizes the logic used in the example above, where one compares each individual's vignette responses to the self-assessment (King et al., 2004). When we have only one vignette, we create a non-parametric measure of the underlying variable that is comparable across respondents by coding whether the respondent puts herself at a level higher than, equal to, or lower than the level she put the person in the vignette. More generally, this procedure turns a self-assessment and  $J$  vignettes about a given topic into  $2J + 1$  ordinal categories that are comparable across individuals. For example, suppose we have two vignettes written to convey two different levels of political efficacy, say someone who writes a letter to an official and someone who speaks up at a town meeting. The comparability-corrected variable will have five ordinal categories of political efficacy: below both vignettes, equal to the letter writer, between the letter writer and the town meeting speaker, equal to the town meeting speaker, and above both vignettes.

However, the cost of this additional precision may be respondent confusion in some cases. More specifically, having more than one vignette may cause the responses to be inconsistent or not as informative as we might like. In anchoring vignettes, this means that the rescaled response variable could indicate a range of ordinal values instead of a single value. Consider a respondent who gives the same answer for the self-assessment and both of the vignette questions. We refer to this as a *tied* response. Using the rescaled estimator, we cannot place this individual in a single category, but must acknowledge that, of the five potential categories, his response could be any member of the set  $\{2, 3, 4\}$ . That is, he could be tied with the lower vignette, between the two vignettes, or tied with the higher vignette, but he cannot be above both (in category 5) or below both (category 1). In a similar way, intervals are used to indicate *inconsistent* responses, in which an individual reports that he is above a higher vignette and below a lower vignette.

When the vignette-corrected responses are not tied or inconsistent, they can be analyzed like a standard ordinal variable with  $2J + 1$  categories, such as with an ordered probit

model. However, the ties and inconsistencies introduce some added complexity, as we do not have full information for the tied or inconsistent respondents. In the example above, we might know only that a tied respondent is equal to the letter writer, between the letter writer and the town meeting speaker, or equal to the town meeting speaker. But how does one relate political efficacy to other variables if some of the respondents have rescaled efficacy levels that span multiple categories? Establishing the relationship between covariates and the incomparability-corrected measure will prove crucial in evaluating our proposed changes in vignette administration. Yet standard statistical models such as the ordered probit assume that we observe the exact category of each individual’s response.

As King and Wand (2007) show, the nonparametric variables created by anchoring vignettes, which include some interval-valued responses, can be modeled using a censored ordered probit model that they develop. The basic (uncensored) ordered probit model assumes, for each independent observation  $i$ , that there exists an unobserved variable  $Y_i^* \sim N(X_i\beta, 1)$  where  $Y_i^*$  generates the observed ordered dependent variable,  $y_i$ , depending on a vector of threshold (or “cutpoint”) values  $\tau$  (e.g.,  $y_i = 1$  if  $Y_i^* < \tau_1$ ,  $y_i = 2$  if  $\tau_1 \leq Y_i^* \leq \tau_2$ , etc.). In the likelihood,  $y_i$  enters as its ex ante probability: the slice of the normal density bounded by the two cutpoints that define that ordinal value. Formally, we represent this probability (i.e., that the latent variable falls between the two thresholds that correspond to  $y_i$ ) as  $\Pr(Y_i = y|X_i) = \int_{\tau_{y-1}}^{\tau_y} N(y^*|X_i\beta, 1)dy^*$ . The censored ordered probit model is identical to the basic ordered probit model for scalar values of the dependent variable. For censored values—that is, responses that span multiple response categories—the probability changes only by integrating over all adjacent slices of the normal distribution within the set rather than only one slice. In short, the censored ordered probit is a simple generalization to the standard ordered probit to allow for responses that span multiple categories.

All the cutpoints  $\tau$  remain identified in this model, so long as at least some respondents provide information about each cutpoint’s location (i.e., at least one respondent gives each unique value of the ordered dependent variable). One can interpret the resulting

coefficients in exactly the same way as the ordered probit: they are the linear effects of the covariates on the latent dependent variable  $Y^*$ . The experiments we run estimate how changes in survey design impact the results from censored ordered probits as well as the number of responses that are tied or inconsistent. As an alternative to this semi-parametric approach, more fully parametric models are also available to analyze anchoring vignette data (King et al., 2004).

### **3 The Benefit of Priming**

In this section, we build on the survey design literature to develop hypotheses about how best to administer anchoring vignettes, discuss how to validate new approaches, introduce the details of our question order experiments, and present the results.

#### **3.1 Hypotheses**

Research on survey instruments consistently demonstrates the influence of question order on survey responses (e.g. Bradburn, Sudman and Wansink, 2004; Groves et al., 2004; Schwarz, 1999; Krosnick, 1999; Schuman and Presser, 1996; Sudman, Bradburn and Schwarz, 1996; Smith, 1991; Strack, 1991; Tourangeau, 1991). Respondents interpret a given survey question based on its context within the survey, using the previous questions to glean information not provided by the question itself. Perhaps for that reason, the World Health Survey and other surveys using anchoring vignettes have typically asked the self-assessment first and then asked the respondent to assess the hypothetical vignettes. Doing so limits the extent to which the vignettes can prime certain considerations among respondents when assessing themselves. For the same reason, we hypothesize that placing vignettes prior to the self-assessment will clarify the meaning of the self-assessment question and familiarize the respondents with the response scale, further improving measurement (Gerber, Wellens and Keeley, 1996).

Question-order effects come in several variants, and they appear reliably in cases

where respondents are asked a battery of similar questions (Smith, 1991) and where the topics are low-salience (Schuman and Presser, 1996). Certainly, anchoring vignettes typically meet both these criteria: they often ask about concepts that respondents do not commonly think about, and they generally involve asking several vignettes consecutively on the same topic. Of the various question-order effects, part-whole effects are especially relevant (Willits and Ke, 1995; Schuman and Presser, 1996; Schwarz, Strack and Mai, 1991). When respondents are asked specific questions that are part of a larger topic, and then asked to make a more general assessment about that topic, the presence of the specific questions primes them on that topic and shapes their answer to the more general question. For instance, after being asked about their marriage, respondents who are then asked about their happiness will provide an answer that is heavily influenced by their marriage (McClendon and O'Brien, 1988). Having been primed by the specific questions, the respondent interprets the general question in a similar light. The implications for anchoring vignettes are clear. When the more general self-assessment question is asked after the more concrete vignettes, respondents are likely to interpret its meaning with reference to the vignettes. Question order research suggests that, whereas respondents who hear the self-assessment first are likely to answer with reference to their own personal definition of the key concept, *those who first respond to several vignettes will have a more standardized conception of what is being asked*. They will also be more familiar with the range of the attribute being assessed. The benefits of priming are analogous to taring a scale before weighing something.

### **3.2 Experimental Design and Survey Administration**

We choose political efficacy as the main topic for our experiments, as it is an important but abstract concept susceptible to varying meanings. Efficacy is a dependent and independent variable of considerable importance for political scientists (Verba, Schlozman and Brady, 1995; Stewart et al., 1992; Finkel, 1987), but it has also been plagued by inter-personal incomparability and other measurement issues (Niemi, Craig and Mattei, 1991;

Craig, Niemi and Silver, 1990). We focus specifically on external efficacy, which refers to respondents' expectations about governmental responsiveness rather than their feelings about their own capacities (Niemi, Craig and Mattei, 1991, pp. 1407–8). One commonly used survey question measuring external efficacy is: “How much say do you have in getting your local government to consider issues that interest you? Do you have a lot of say, some say, little say, or no say at all?” As the appendix details, the accompanying vignettes provide examples of citizens who want to reduce the speed limit on their street. One citizen does not act because he does not expect a favorable response, one writes a letter to a local official, one raises the issue at a community meeting, and one meets with a local official who promises to address the issue.<sup>4</sup> Each vignette defines a different point along the continuum. Jointly, they allow us to re-scale responses to be comparable across respondents.

We conducted two randomized experiments, one by telephone survey and the other via a web survey. In each, we randomly assigned respondents to receive the vignettes either before or after the self-assessments. Our interviewer-assisted telephone survey was conducted by the Survey Research Center at the University of Indiana in English from October 19th, 2005 to February 4th, 2006. The American Association for Public Opinion Research RR3 response rate was 30.1%. The 916 respondents who completed the interview were identified by random digit dialing, and the survey targeted the population of U.S. adults with telephones. Ninety-five percent of respondents completed the module in 11 minutes or less. The sample of telephone respondents was more highly educated than the U.S. population as a whole, with 37% claiming a bachelor's degree or higher. It was also 9% black and 5% Latino. (In the 2004 American Community Survey, the respective numbers for the U.S. population were 27% with a bachelor's degree or higher, 12% black, and 14% Latino.)

Our web survey was conducted by Polimetrix (now YouGov/Polimetrix) from March 13th to March 17th, 2006, with 1,200 American adults responding based on a weighted sampling of email addresses collected by Polimetrix from various sources.<sup>5</sup> Potential

respondents were invited to participate so as to match a randomly drawn sample from the 2004 American Community Survey as closely as possible. Differential response rates are the primary reason why 32% of internet respondents had a bachelor's degree or additional education while only 3% identified as black and 2% identified as Latino.

### 3.3 Validation

The challenge of validating a given approach to using vignettes deserves special attention. There is no obvious way to confirm a respondent's report that she has little influence over her local government, as the concept is subjective. This problem is of course a general one: if we had a better measure, especially if we had something approaching a gold standard, we would use that and drop our approach altogether. Since such a gold standard does not exist, we focus on the common approach of relying on construct validity and discriminant validity (Fowler, 1995, p. 139).<sup>6</sup> We draw on theories of political efficacy to generate a list of explanatory variables that should be unambiguously related to the measure in question. Where possible, we identify alternative measurement strategies that have been validated through repeated testing. We then assess whether the relative placement of the vignettes and the self-assessment influences the relationship between these variables and the incomparability-corrected measure. In all experiments, individual vignettes are randomly ordered.

We identified seven variables with known relationships to external political efficacy. In key respects, external efficacy has similar roots to other forms of political participation (Verba, Schlozman and Brady, 1995; Finkel, 1987). For example, we know from past scholarship that more educated and wealthier people should feel more efficacious in politics (Oliver, 2001; Verba, Schlozman and Brady, 1995; Craig, Niemi and Silver, 1990; Finkel, 1985), as should those living in communities with wealthier ZIP codes or with larger shares of homeowners (DiPasquale and Glaeser, 1999; Oliver, 2001). By contrast, heavily populated ZIP codes should dampen efficacy, as residents see their voice as one among many (Oliver, 2001, pg. 228).<sup>7</sup> Other ways of asking about political effi-

cacy should also correlate more highly with measures of efficacy that are relatively free of measurement error.<sup>8</sup>

### 3.4 Results

The section presents results from our question order experiments. Consider the phone respondents to the efficacy module first. For 472 respondents randomly assigned to the control group, we followed prior practice by first asking them to assess themselves. The 431 respondents assigned to the treatment group were first asked the vignette questions and then assessed themselves on the same scale.<sup>9</sup> We found that 52% of those in the treated group (who read the vignette questions first) but only 40% of those in the control group (who read the self-assessment questions first) placed themselves in the most efficacious categories, indicating that exposure to the vignettes did influence responses. In a two-sample t-test, the p-value for the hypothesis of no difference in the two groups is less than 0.001. The treatment produced similar effects in the online sample, with 41% of the treated but just 34% of the control population placing itself in the two most efficacious categories ( $p=0.02$ ). The treatment successfully altered self-reported efficacy, possibly by redefining the response scale.

We then assessed construct and discriminant validity for the vignette-corrected responses for the treatment and control groups using censored ordered probit models. There are four vignettes about efficacy, yielding an incomparability-corrected variable with nine ordinal categories ranging from respondents whose efficacy is lower than the first vignette to those whose efficacy is higher than the fourth vignette. The censored order probit estimates the impact of an explanatory variable on a latent continuous variable, which is then translated into the ordinal response categories through the set of estimated thresholds.

For validation, we modeled the incomparability-corrected responses separately for each independent variable in the treatment group and each in the control group. For both groups, we estimated the change in the latent dependent variable when shifting the explanatory variable from its 10th percentile to its 90th, to make the results more easily

interpretable. An effect of 0.2, for example, indicates that a shift in the independent variable leads to a change of 0.2 standard deviations in the latent, continuous variable. (By construction, the latent variable has a standard deviation of 1.) To facilitate interpretation, positive estimates are hypothesis-confirming while negative estimates cut against the hypothesis.

Figure 1 represents the results graphically. The dark dots on the left show the variable's effect on political efficacy for the control group while the open circles at right show the variable's effect for the treated group. The arrows denote the change attributable to the treatment of first hearing the vignettes compared to first hearing the self-assessment question. For efficacy assessed in the phone survey, every arrow points to the right, indicating that *the treatment always strengthened the expected relationship between the independent variable and the incomparability-corrected measure of efficacy*. In the case of the percent homeowner, the treatment ensured that the coefficient had the expected sign as well (which can be seen by it moving across the vertical line drawn at zero).

As with a standard ordered probit model, the parameters from the censored ordered probit allow us to calculate the percentage of respondents who fall into each response category given the covariates. For instance, in the control group, we estimate that homeowners are 1.6 percentage points *more* likely than renters to fall in the second-lowest efficacy category. In the treatment group, homeowners are 7.6 percentage points *less* likely to be in that category, producing a difference-in-difference estimate of 9.2 percentage points when comparing treatment and control.<sup>10</sup> Put differently, homeownership is far more predictive of efficacy in the treated group as compared to the control group. In Figure 1, the comparable change in the first difference from control to treatment is presented for each independent variable just to the right of the variable labels. All simulated shifts are from the independent variable's 10th percentile to its 90th. When interpreting these predicted probabilities, it is worth remembering that responses are scattered over nine response categories. Thus in the case of homeownership, the 7.6 percentage point change in the treated group reflects a 25% change in the share of respondents with the second-lowest level of

efficacy.

The treatment effect is noteworthy for other variables as well: the estimated impact of income on political efficacy is four times greater when respondents hear the vignettes first,<sup>11</sup> and we see sizable increases for every independent variable. The two alternate ways of asking about efficacy — labeled “no say” and “don’t care” — also correlate more highly with the dependent variable if respondents were first exposed to the vignettes. By all indications, the vignettes prime respondents to understand the self-assessment differently, producing a more valid measure of efficacy.

We then replicated the experiment using survey data collected online. We randomly assigned 636 respondents to treatment and 551 to control. One difference between the online and phone surveys is that online, respondents were able to return to previous answers if they chose to do so, a fact which has the potential to reduce question-order effects. We were not able to ask respondents’ income or other questions about efficacy in the online survey, so those variables are omitted. Still, we see in the right panel of Figure 1 that the treatment of answering the vignettes again leads to stronger relationships between the independent variables and political efficacy. One variable (the ZIP code’s median household income) achieves the expected sign only when the self-assessment is asked after the vignettes. Here, the change in the first difference is 3.8 percentage points.<sup>12</sup> Other variables see marked increases in their predictive power. We see for instance that shifting from the 10th percentile of ZIP homeownership to the 90th has an effect that is 6 percentage points stronger on the second-lowest efficacy category in the treated group (one-sided  $p$ -value=0.03). Within the treated group, this change is quite significant, representing a 50% decrease in the respondents reporting that level of efficacy. The only exception is for education, where the relationship between the explanatory variable and the incomparability-corrected dependent variable is slightly stronger for those who first answered the self-assessment. Overall, in ten out of eleven trials on efficacy, the treatment produced answers more closely related to explanatory variables of theoretical interest.<sup>13</sup> Priming respondents using vignettes thus does seem to improve their ability to report their

external political efficacy.

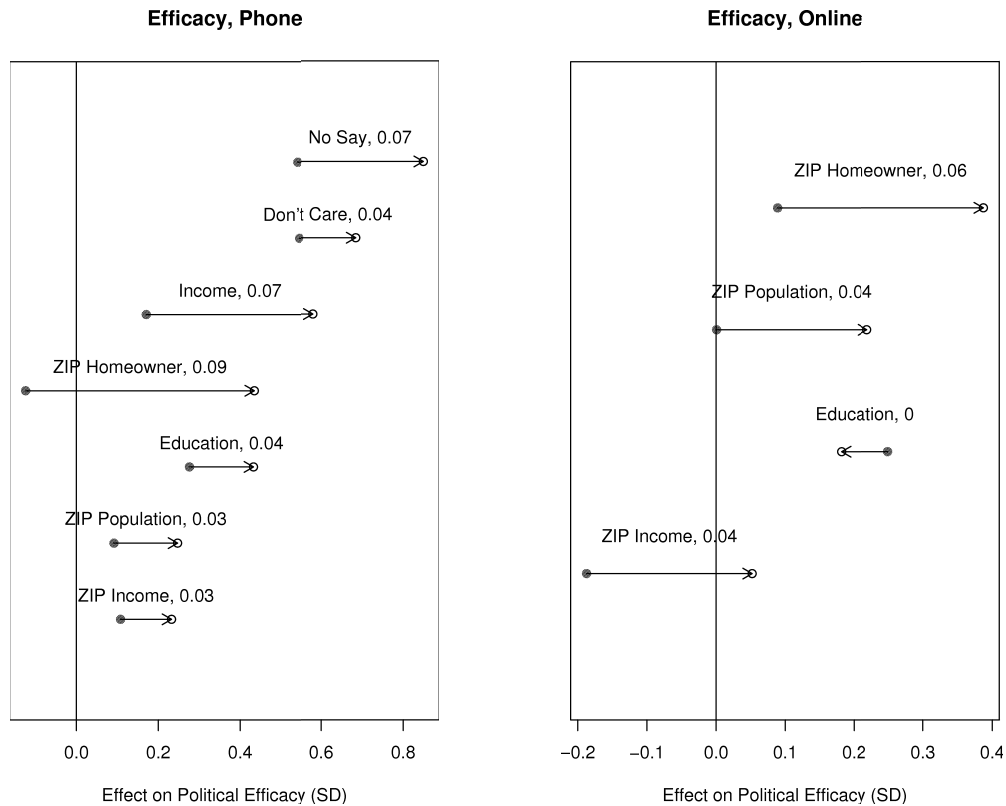


Figure 1: This figure presents results from censored ordered probit models. The x-axis indicates each variable's relationship with political efficacy, which is the latent dependent variable in each model. The black dots show the effect of the explanatory variable on the latent dependent variable in the control group, while the arrows show the change in predictive power for the treated group. Rightward arrows indicate increasing predictive power. The numbers to the right of each label indicate the change in the probability of falling in the second lowest efficacy category under the treatment, with positive numbers indicating expected relationships.

In separate experiments not detailed here, we also evaluated whether the placement of the self-assessment question influenced responses to a battery of questions about economic class. Class and socioeconomic status have been central variables in social science since at least the mid-19th century, and have received increasing attention in public health in recent years as well (e.g. Sorensen, 2000; Grusky and Sorensen, 1998; Krieger, Williams and Moss, 1997; Jackman and Jackman, 1991).<sup>14</sup> Respondents were asked to

identify their economic class, and were randomly assigned to do so before or after assessing five vignettes about hypothetical individuals with differing incomes. The class results are not as striking as those for political efficacy, perhaps because economic class is familiar even to unprimed respondents. Only one variable—race—sees a sharp increase in predictive power. Yet even so, 8 of the 11 trials are in our hypothesized direction, meaning that the expected relationships between demographics and class are stronger in the treated group that encountered the vignettes prior to the self-assessment. Given the clear advantages demonstrated here, and no obvious disadvantages, we conclude that *researchers should move vignettes to before the self-assessments in order to prime and to correct responses.*

## 4 The Trouble with Direct Comparisons

To use anchoring vignettes, one must supplement a self-assessment question with one or more vignettes, increasing survey costs. Thus, it is worth investigating techniques that could reduce the total number of questions required (Bowling, 2005). In this section, we consider one potential way to correct for incomparability with fewer questions, by allowing respondents to directly compare themselves to the hypothetical individual described in each vignette. Doing so removes the need for a separate self-assessment. With  $J$  vignettes, using direct comparisons requires  $J$  questions, while the usual anchoring vignette approach requires a self-assessment for a total of  $J + 1$  questions. As with the approach to vignettes discussed above, this approach also allows one to create a corrected measure with  $2J + 1$  categories, from a respondent who is lower than the lowest vignette to a respondent who is higher than the highest vignette. This section demonstrates that, in practice, direct comparisons are subject to significant center-seeking biases, rendering them unhelpful. This seemingly innocuous choice dramatically increases inconsistent responses. Asking a separate self-assessment produces marked improvements in measurement.

To assess these trade-offs, we used a separate randomization to identify 236 respondents to the phone survey who were to report their rest and energy levels using both response formats. Rest is especially well suited to our purposes as it may be somewhat less susceptible to question-order or question-wording effects. Of course, in any real-world application of these methods, researchers would need only  $J$  questions (if using the direct comparison approach) or  $J + 1$  questions (if using the separate vignettes approach). But to know which approach is preferable, it is essential that the same respondents assess themselves using both question formats.

In this module, we first administered a set of vignettes about rest and energy along with a self-assessment question. We then asked respondents to compare themselves directly to the same vignettes about rest and energy levels. Using the vignettes given in the appendix, we asked after each: “Thinking about the last week, would you say that you typically felt more energetic than (Name), about equally as energetic as (Name), or less energetic than (Name)?” If anything, the relative placement of these direct comparison questions after the vignettes should give them an advantage, as respondents are already familiar with the questions and the response scale.

One initial criterion in evaluating these two modes of administration is the share of responses that are inconsistent. And as Figure 2 demonstrates, switching from separate vignettes to direct comparison questions leads to marked increases in responses on the comparability-corrected scale that are logically inconsistent. For every possible combination of the four vignettes (ordered vertically), it plots the percentage of respondents who gave inconsistent or tied responses using both formats. The triangle on the left-hand side of the figure indicates the share of the 236 responses that are inconsistent using the separate vignettes approach. The head of the arrow indicates the share of responses from the same respondents that are inconsistent using direct comparisons. Clearly, responses cannot be inconsistent with only one vignette, explaining the absence of an arrow for the single vignettes at bottom. But as we scan up the left-hand side of the figure, we see that with two or more vignettes, there is strong evidence that the direct comparison approach

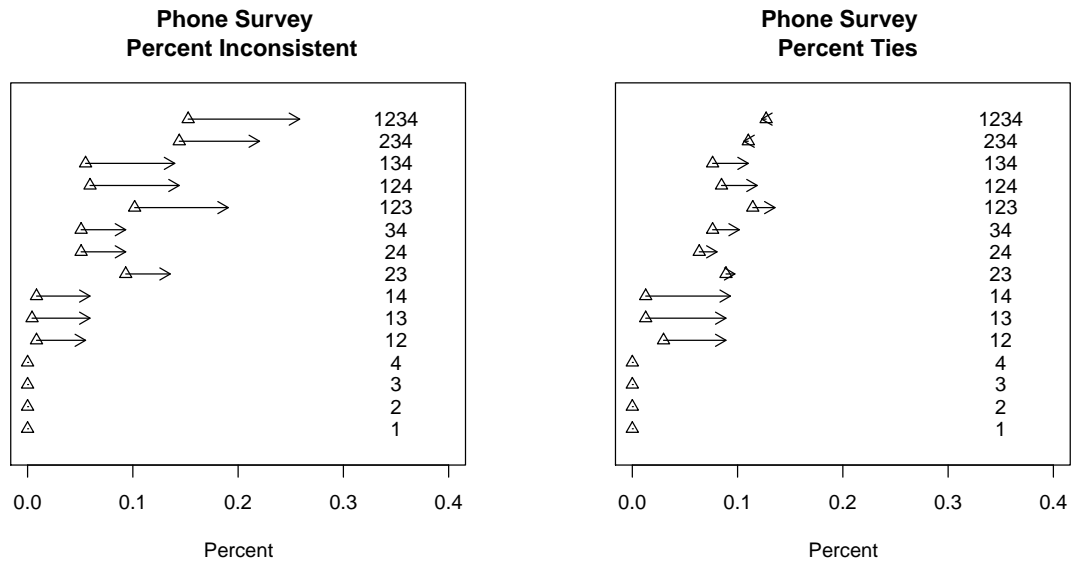


Figure 2: For all combinations of vignettes, these figures show the increases in inconsistent and tied responses when shifting from the basic vignettes approach (triangles) to direct comparisons (arrow heads). The horizontal axis indicates the percent of responses that are tied or inconsistent while the vertical axis indicates which of the four vignettes were used. These results are for the phone survey.

induces more inconsistency than the traditional approach with a separate self-assessment. As shown by the triangle at the top left of Figure 2, when using all four vignettes, only 15% of respondents gave inconsistent answers when asked vignettes and self-assessments separately. However, 26% of respondents gave inconsistent answers when making direct comparisons between themselves and hypothetical vignettes, as shown by the arrowhead at right. (A t-test confirms that such differences would appear by chance alone with a probability less than 0.005.) The right-hand side of the figure shows that direct comparisons also typically increase the number of tied responses, although this result varies to some extent by the subset of vignettes under discussion.

For the Internet sample, 1,052 respondents provided answers to both the vignettes about sleep and energy and to a separate set of logically equivalent questions in which they were asked to make direct comparisons between themselves and hypothetical individuals. Here as well, the vignettes were administered first, giving respondents increased familiarity with the questions when asked to make direct comparisons. Again, the re-

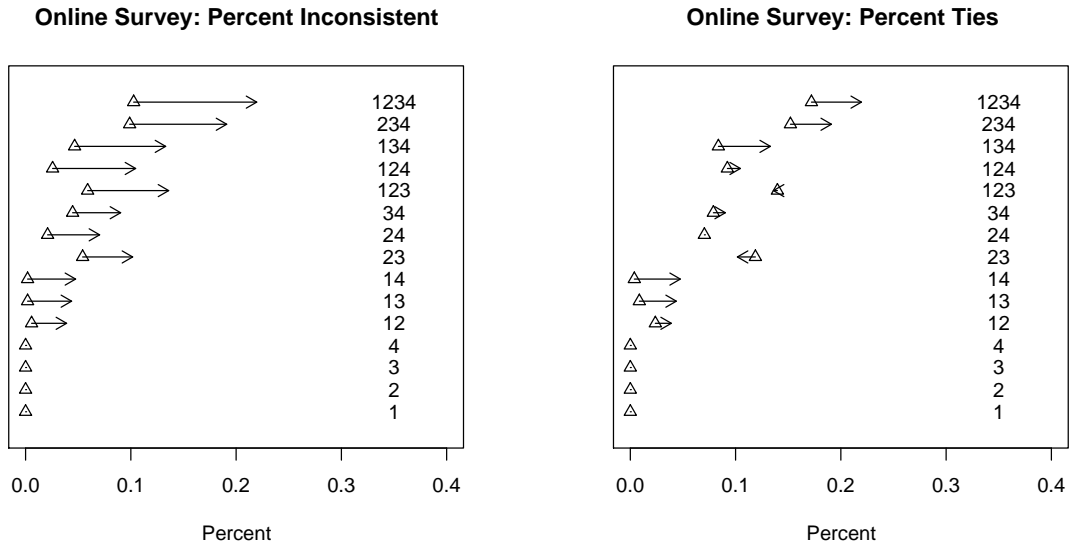


Figure 3: For all combinations of vignettes, these figures show the increases in inconsistent and tied responses when shifting from the basic vignettes approach (triangles) to direct comparisons (arrow heads). The horizontal axis denotes the percent of responses that are tied or inconsistent while the vertical axis indicates which of the four vignettes were used. These results are for the online survey.

sults are unambiguous, as Figure 3 demonstrates. For every combination of vignettes, the number of inconsistencies rises markedly when respondents make direct comparisons. Using all four vignettes, 22% of the direct comparison responses but only 10% of the vignette responses were inconsistent. Here, the p-value from a t-test is less than 0.0001. In most cases, asking respondents to make direct comparisons also increases the share of responses that are tied, again reducing the information inherent in the responses. This is yet more evidence in favor of asking vignettes separately.

#### 4.1 Why Direct Comparisons Fail

The question now becomes *why* the separate vignettes approach yields incomparability-corrected responses with so much more information. Examining the distribution of responses using each approach suggests an answer. When comparing themselves directly, respondents in both the phone and Internet surveys are consistently more likely to say

they are the same as the individual in the vignette. For instance, in the online survey, when asked directly, 22% of respondents said their level of energy was comparable to someone who “wakes up almost once every hour during the night. In the morning, he does not feel well-rested and is tired all day.” When asked via vignettes, only 15% of people likened themselves to someone with the same low level of rest. Put differently, the direct comparison induces respondents to say that they are like the hypothetical individual, wherever he happens to fall on the spectrum.

As another example of this tendency, 33% of phone respondents indicated through the direct comparison that they were equal to the most rested vignette and yet also indicated through the separate approach that they were more rested than the same vignette. This could be a result of the three-category response scale, which encourages respondents to opt for the middle ground (Weisberg, Krosnick and Bowen, 1996; Schuman and Presser, 1996; Schwarz, Hippler and Noelle-Neumann, 1991; Bradburn, 1983; Payne, 1951). Respondents infer typical behavior from the response scale, leading to a center-seeking tendency (Krosnick, 1999; Schwarz et al., 1985). In this case, that tendency is exacerbated by the vignette itself, which seems to provide additional information about what is normal, since it indicates one individual’s story. Encouraged by both the response scale and the vignette, many respondents to direct comparisons say that they are like the individual in the vignette, even when doing so is inconsistent with other responses. The loss of information grows worse still when employing more than one vignette, since individuals who indicate being the same as multiple vignettes are either tied or inconsistent on the incomparability-corrected scale. Although they appeal as a way to reduce survey costs, *direct comparison questions produce substantial losses of information and biases, and should be avoided.*

This finding obviously applies to the growing number of surveys that employ anchoring vignettes, but it also applies to the larger set of surveys that currently employ direct comparisons. For example, the 2004 National Election Study asked respondents: “Compared to the average person do you have fewer opinions about whether things are good or

bad, about the same number of opinions, or more opinions?” Here, too, a respondent is asked to compare himself to a hypothetical individual, and is encouraged by the response scale and by social pressure to say that he has the same number of opinions. Given the results above, it is not surprising that 55% of all respondents indicated that they had the same number of opinions as the average person. Nor is it surprising that in 2000, the share in the center category was identical.

A General Social Survey (GSS) time-series question asking respondents to compare their family income to American families in general risks provoking the same biases. In 1991, for example, 48% of all respondents responded to that question by saying that their family income was average, again demonstrating a strong center-seeking tendency. That figure was 48% in 1996, 47% in 1998, 49% in 2001, and 48% in 2006. In year after year, a strong plurality of respondents claimed to be “average,” reducing the information available to analysts. We have identified many other surveys that ask respondents to make direct comparisons between themselves and others as well, from *New York Times* surveys to *National Geographic* surveys. Our results seem to indicate that anchoring vignettes would improve measurement in these situations.

## 5 Concluding Remarks

In this study of how to use anchoring vignette technology to correct for interpersonal and cross-cultural incomparability, we reconfirm the benefits of the venerable advice: paying attention to the small things in survey research can have substantial pay-offs. We showed that question-order effects can be used to prime respondents into understanding survey questions as intended, and question wording can be tuned to improve certain question types.

In particular, when implementing anchoring vignettes, our randomized experiments suggest that the self-assessment question be moved to after the set of vignettes. They also suggest that the comparison between vignettes and self-assessments not be combined

into a single question, as that tends to discard a considerable amount of information. As such direct comparison questions have been used in many different types of surveys, this second result suggests numerous other applications for anchoring vignette technology well beyond its intended purpose of correcting for interpersonal incomparability.

We encourage other scholars to go beyond our results to study exactly how vignettes are posed, self-assessments are asked, and response categories are constructed. For example, it would be worth evaluating whether online surveys can eliminate inconsistencies by presenting all the vignettes to a respondent simultaneously (something infeasible for a phone survey). We could also experiment with comparison questions without a middle category option, which would eliminate the possibility of tied responses. Overall, the years of advice built up by social psychologists and other survey researchers provides a host of valuable ideas to try and methods to evaluate.

## Notes

<sup>1</sup>In educational research, interpersonal incomparability is widely known under the name “Differential Item Functioning” or DIF. For details, see Holland and Wainer (1993) and King et al. (2004, fn 1).

<sup>2</sup>In addition, other studies have adopted similar approaches for dealing with inter-group differences in response category use, including research on national identity in the U.S. and the U.K. (Javaras and Ripley, 2007), happiness in China (Hsee and Tang, 2007), and binge eating in the Boston area (Javaras et al., 2008).

<sup>3</sup>In most cases, the removal of the separate self-assessment question reduces the total number of questions by one. However, direct comparisons might also facilitate an adaptive questionnaire, where individuals are only asked the minimum questions needed to locate their position relative to the hypothetical individuals. Direct comparisons thus have the potential to reduce questionnaire length by more than one question.

<sup>4</sup>We randomized the order of the vignettes within each set, matched the names of the hypothetical citizens to the respondent’s gender, and randomly varied among a set of common names. This encourages respondents to think of the vignettes as someone like themselves but reduces the likelihood that respondents would draw unintended information from the use of one name over another.

<sup>5</sup>The AAPOR RR1 response rate was 32.6%.

<sup>6</sup>Fowler (1995) defines construct validity by noting that “if several questions are measuring the same or closely related things, then they should be highly correlated.” He defines discriminant validity as “the extent to which groups of respondents who are thought to differ in what is being measured, in fact, do differ in their answers.”

<sup>7</sup>Although there are of course more ZIP codes in more populated areas, there is still a strong positive correlation between ZIP code population and a city or county’s population.

<sup>8</sup>We adapted two National Election Study questions asking respondents to agree or disagree with certain statements. The first statement was “people like me don’t have

any say about what our local government does,” while the second statement was “Local officials don’t care much what people like me think.” See Craig, Niemi and Silver (1990) and Niemi, Craig and Mattei (1991) for more on the measurement of efficacy.

<sup>9</sup>In this experiment and elsewhere, two-sample t-tests confirmed that randomization achieved covariate balance as designed.

<sup>10</sup>The 95% confidence interval for the difference-in-difference estimate runs from 3 to 16 percentage points. The corresponding one-sided p-value is 0.001.

<sup>11</sup>Here, when treated, those reporting incomes between \$15,000 and \$25,000 are 7 percentage points more likely to be in the second-lowest efficacy category than those reporting incomes over \$75,000. The corresponding one-sided p-value is 0.001.

<sup>12</sup>The one-sided p-value on the change in first differences is 0.11.

<sup>13</sup>When considered collectively, these results, across different questions and modes of survey administration, are obviously unlikely to have occurred by chance, although we do not offer a formal hypothesis test because of complications caused by the dependence among the separate analyses.

<sup>14</sup>Using vignettes to measure economic class provides another potential advantage to survey researchers. While surveys asking about income routinely encounter refusal rates of over 10%, only 2% of our sample refused to report their class status.

## Appendix: Survey Questions

### Political Efficacy

Self-assessment: How much say do you have in getting your local government to consider issues that interest you? Do you have a lot of say, some say, little say, or no say at all?

Below are the vignettes for political efficacy. All end with the following question: How much say do you think (name) has in getting (his/her) local government to consider issues that interest him/her?

- (Name) is concerned about cars speeding by (his/her) house, and (he/she) would like to see the speed limit on (his/her) street reduced. However, (he/she) knows that (his/her) town, city, or county elected official is from another part of town, and so is very unlikely to help him/her.
- (Name) is concerned about cars speeding by (his/her) house, and (he/she) would like to see the speed limit on (his/her) street reduced. (He/she) writes a letter to (his/her) town, city, or county elected official and receives a form letter in reply.
- (Name) is concerned about cars speeding by (his/her) house, and (he/she) would like to see the speed limit on (his/her) street reduced. (He/she) brings the issue up at a public town meeting. The issue is thoroughly debated by (his/her) town, city, or county elected officials.
- (Name) is concerned about cars speeding by (his/her) house, and (he/she) would like to see the speed limit on (his/her) street reduced. (He/she) meets with (his/her) town, city, or county elected official, who promises to work on the matter.

### Economic Class

Self-assessment: Which describes your economic class best? Would you say you belong to the upper class, upper middle class, middle class, working class, or lower class?

The vignettes for economic class are all of the following form, with the amount of money being 18 thousand dollars, 34 thousand dollars, 55 thousand dollars, 87 thousand dollars, and 154 thousand dollars. These represent the 20th, 40th, 60th, 80th, and 95th percentiles in U.S. household income in 2003 according to the U.S. Census Bureau.

(Name) makes 55 thousand dollars a year. Which economic class describes (Name) best?

## **Rest/Energy**

Self-Assessment: In the last week, during the day, would you say you typically felt very energetic, somewhat energetic, somewhat tired, or very tired?

Below are the vignettes for rest/energy. All end with the following question: In the last week, during the day, would you say that (Name) typically felt very energetic, somewhat energetic, somewhat tired, or very tired?

- (Name) wakes up feeling well-rested, and remains alert throughout the day.
- (Name) has no trouble falling asleep, but every morning (he/she) finds it difficult to wake up. (He/she) is sometimes late to work and feels tired in the mornings.
- Two nights a week, (name) wakes up in the middle of the night, and cannot fall back to sleep. On these days, (he/she) is exhausted at work.
- (Name) wakes up almost once every hour during the night. In the morning, he does not feel well-rested and is tired all day.

## References

- Bowling, Ann. 2005. “Just one question: If one question works, why ask several?” *Journal of Epidemiology and Community Health* 59(5):342.
- Bradburn, Norman M. 1983. Response Effects. In *Handbook of Survey Research*, ed. Peter H. Rossi, James D. Wright and Andy B. Anderson. New York, NY: Academic Press.
- Bradburn, Norman M., Seymour Sudman and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design*. San Francisco: Jossey-Bass.
- Buckley, Jack. 2008. “Survey Context Effects in Anchoring Vignettes.” <http://polmeth.wustl.edu/workingpapers.php>.
- Craig, Stephen C., Richard G. Niemi and Glenn E. Silver. 1990. “Political efficacy and trust: A report on the NES pilot study items.” *Political Behavior* 12(3):289–314.
- Damacena, G.N., M.T.L. Vasconcellos and C.L. Szwarcwald. 2005. “Perception of health state and the use of vignettes to calibrate for socioeconomic status: results of the World Health Survey in Brazil, 2003.” *Cadernos de Saúde Pública* 21:65–77.
- DiPasquale, Denise and Edward L. Glaeser. 1999. “Incentives and Social Capital: Are Homeowners Better Citizens?” *Journal of Urban Economics* 45(2):354–384.
- Finkel, Steven E. 1985. “Reciprocal Effects of Participation and Political Efficacy: A Panel Analysis.” *American Journal of Political Science* 29(4):891–913.
- Finkel, Steven E. 1987. “The Effects of Participation on Political Efficacy and Political Support: Evidence from a West German Panel.” *Journal of Politics* 49(2):441–464.
- Fowler, Floyd J. 1995. *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications.
- Gerber, Eleanor R., Tracy R. Wellens and Catherine Keeley. 1996. Who lives here? The use of vignettes in household roster research. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*. pp. 962–967.
- Groves, Robert M., Mick P. Couper, James M. Lepkowski, Eleanor Singer and Roger

- Tourangeau. 2004. *Survey Methodology*. Hoboken, NJ: Wiley.
- Grusky, David and Jesper B. Sorensen. 1998. "Can Class Analysis Be Salvaged?" *American Journal of Sociology* 103(5):1187–1234.
- Grzymala-Busse, Anna. 2007. *Rebuilding Levithan: Party Competition and State Exploitation in Post-Communist Democracies*. New York: Cambridge University Press.
- Gupta, Nabanita Datta, Nicolai Kristensen and Dario Pozzoli. 2008. "External Validation of the Use of Vignettes in Cross-Country Health Studies." Danish National Centre for Social Research.
- Holland, Paul W. and Howard Wainer, eds. 1993. *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum.
- Hsee, Christopher K. and Judy Ningyu Tang. 2007. "Sun and Water: On a Modulus-Based Measurement of Happiness." *Emotion* 7:213–218.
- Jackman, Mary R. and Robert W. Jackman. 1991. *Class Awareness in the United States*. Berkeley, CA: University of California Press.
- Javaras, Kristin, Harrison Pope, Justine Lalonde, Jacqueline Roberts, Yael Nillni, Nan Laird, Cynthia Bulik, Scott Crow, Susan McElroy, B. Timothy Walsh et al. 2008. "Co-occurrence of Binge Eating Disorder with Psychiatric and Medical Disorders." *The Journal of Clinical Psychiatry* 69(2):266–273.
- Javaras, Kristin N. and Brian D. Ripley. 2007. "An 'Unfolding' Latent Variable Model for Likert Attitude Data: Drawing Inferences Adjusted for Response Style." *Journal of the American Statistical Association* 102(478):454–463.
- Kapteyn, Arie, James P. Smith and Arthur Soest. 2007. "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." *American Economic Review* 97(1):461–473.
- King, Gary, Christopher J.L. Murray, Joshua A. Salomon and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1, February):191–207.
- King, Gary and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses:

- New Tools for Anchoring Vignettes.” *Political Analysis* 15(1, Winter):46–66.
- Krieger, N., D.R. Williams and M.E. Moss. 1997. “Measuring Social Class in U.S. Public Health Research: Concepts, Methodologies, and Guidelines.” *Annual Reviews in Public Health* 18(1):341–378.
- Kristensen, Nicolai and Edvard Johansson. 2008. “New evidence on cross-country differences in job satisfaction using anchoring vignettes.” *Labour Economics* 15(1):96–117.
- Krosnick, Jon A. 1999. “Survey Research.” *Annual Review of Psychology* 50(1):537–567.
- McClendon, McKee J. and David J. O’Brien. 1988. “Question-Order Effects on the Determinants of Subjective Well-Being.” *Public Opinion Quarterly* 52(3):351–364.
- Niemi, Richard G., Stephen C. Craig and Franco Mattei. 1991. “Measuring Internal Political Efficacy in the 1988 National Election Study.” *American Political Science Review* 85(4):1407–1413.
- Oliver, J. Eric. 2001. *Democracy in Suburbia*. Princeton, NJ: Princeton University Press.
- Payne, Stanley L. 1951. *The Art of Asking Questions*. Princeton, NJ: Princeton University Press.
- Salomon, Joshua A., Ajay Tandon and Christopher J.L. Murray. 2004. “Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes.” *British Medical Journal* 328(7434):258.
- Schuman, Howard and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. Thousand Oaks, CA: Sage.
- Schwarz, Norbert. 1999. “Self-Reports: How the Questions Shape the Answers.” *American Psychologist* 54(2):93–105.
- Schwarz, Norbert, Fritz Strack and Hans-Peter Mai. 1991. “Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis.” *Public Opinion Quarterly* 55(1):3–23.
- Schwarz, Norbert, Hans J. Hippler, Bridget Deutsch and Fritz Strack. 1985. “Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments.” *Public Opinion Quarterly* 49(3):388–395.

- Schwarz, Norbert, Hans J. Hippler and Elisabeth Noelle-Neumann. 1991. A Cognitive Model of Response-Order Effects. In *Contextual Effects in Social and Psychological Research*, ed. Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag pp. 187–202.
- Smith, Tom. 1991. Thoughts on the Nature of Context Effects. In *Contextual Effects in Social and Psychological Research*, ed. Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag pp. 163–186.
- Soest, Arthur Van, Liam Delaney, Colm P. Harmon, Arie Kapteyn and James P. Smith. 2007. “Validating the Use of Vignettes for Subjective Threshold Scales.” UCD Geary Institute Working Paper.
- Sorensen, Aage B. 2000. “Toward a Sounder Basis for Class Analysis.” *American Journal of Sociology* 105(6):1523–1558.
- Stewart, Marianne C., Allan Kornberg, Harold D. Clarke and Alan Acock. 1992. “Arenas and Attitudes: A Note on Political Efficacy in a Federal System.” *Journal of Politics* 54(1):179–196.
- Strack, Fritz. 1991. ‘Order Effects’ in Survey Research: Activation and Information Functions of Preceding Questions. In *Contextual Effects in Social and Psychological Research*, ed. Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag pp. 23–34.
- Sudman, Seymour, Norman M. Bradburn and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass Publishers.
- Tourangeau, Roger. 1991. Context Effects on Responses to Attitude Questions: Attitudes as Memory Structures. In *Contextual Effects in Social and Psychological Research*, ed. Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag pp. 35–48.
- Tourangeau, Roger. 2004. “Survey Research and Societal Change.” *Annual Review of Psychology* 55(1):775–801.
- Verba, Sidney, Kay Lehman Schlozman and Henry E. Brady. 1995. *Voice and Equality:*

*Civic Volunteerism in American Politics*. Cambridge, MA: Harvard University Press.

Wand, Jonathan. 2007. "Credible Comparisons Using Interpersonally Incomparable Data: Ranking self-evaluations relative to anchoring vignettes or other common survey questions." <http://wand.stanford.edu>.

Weisberg, Herbert F., Jon A. Krosnick and Bruce D. Bowen. 1996. *An introduction to survey research, polling, and data analysis*. Thousand Oaks, CA: Sage Publications.

Willits, Fern K. and Bin Ke. 1995. "Part-Whole Question Order Effects: Views of Rural-ity." *Public Opinion Quarterly* 59(3):392–403.