Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis

Kosuke Imai Princeton University
Teppei Yamamoto Princeton University

Political scientists have long been concerned about the validity of survey measurements. Although many have studied classical measurement error in linear regression models where the error is assumed to arise completely at random, in a number of situations the error may be correlated with the outcome. We analyze the impact of differential measurement error on causal estimation. The proposed nonparametric identification analysis avoids arbitrary modeling decisions and formally characterizes the roles of different assumptions. We show the serious consequences of differential misclassification and offer a new sensitivity analysis that allows researchers to evaluate the robustness of their conclusions. Our methods are motivated by a field experiment on democratic deliberations, in which one set of estimates potentially suffers from differential misclassification. We show that an analysis ignoring differential measurement error may considerably overestimate the causal effects. This finding contrasts with the case of classical measurement error, which always yields attenuation bias.

Political scientists have long been concerned about measurement error. In particular, various consequences of measurement error have been extensively studied in the context of survey research (e.g., Achen 1975; Asher 1974; Bartels 1993; Zaller and Feldman 1992). However, the existing research has either completely ignored the problem or exclusively focused on *classical* measurement error in linear regression models where the error is assumed to arise completely at random. In this article, we formally analyze the impact of nonclassical measurement error on the estimation of causal effects. Given the increasing use of randomized experiments in the discipline (Druckman et al. 2006; Horiuchi, Imai, and Taniguchi 2007), measurement error represents a threat to causal inference. Indeed, many experiments use surveys to measure treatment and outcome variables, which introduces the possibility of measurement error.

The methodological literature on measurement error is also immense (see Carroll et al. 2006), and yet statisticians and econometricians are only beginning to address measurement error problems explicitly in the formal statistical framework of causal inference (e.g., Lewbel 2007). Furthermore, much of the previous work has focused on *nondifferential* measurement error where the error is assumed to be independent of the outcome.[1] Nevertheless, *differential* measurement error frequently occurs in retrospective studies where measurements are taken after the

Kosuke Imai is Assistant Professor, Department of Politics, Princeton University, Princeton, NJ 08544 (kimai@princeton.edu; URL: http://imai.princeton.edu). Teppei Yamamoto is Ph.D. student, Department of Politics, Princeton University, Princeton, NJ 08544 (tyamamot@princeton.edu; URL: http://www.princeton.edu/∼tyamamot/).

We are grateful to Macartan Humphreys for generously providing data and constructive comments. We thank Chris Achen, Larry Bartels, Jake Bowers, Josh Clinton, Tasos Kalandrakis, Adam Meirowitz, Jas Sekhon, Curt Signorino, and Dustin Tingley for helpful comments and seminar participants at Northwestern, Princeton, and Rochester for stimulating discussions. An earlier version of this article was presented at the 2008 annual meeting of the Midwest Political Science Association and the 2008 Summer Political Methodology Meeting. Financial support from the National Science Foundation (SES-0752050 and SES-0849715) and the Princeton University Committee on Research in the Humanities and Social Sciences is acknowledged. The replication materials of the empirical results given in this article are available as Imai and Yamamoto (2010) on the Dataverse Network.

[1] For example, in an authoritative monograph of this field, Carroll et al. (2006) explain that "Most of this book focuses on nondifferential measurement error models" (37). Note that unlike classical measurement error studied in political science, nondifferential error may depend on the true value of the mismeasured variable.

American Journal of Political Science, Vol. 54, No. 2, April 2010, Pp. 543–560

ISSN 0092-5853

outcome is realized and thus the error could be correlated with the outcome.[2] Differential measurement error can also arise in prospective studies if, for example, an unobserved covariate is correlated with both the outcome and the measurement error.

In this article, we study the nonparametric identification of the average treatment effect (ATE) when a binary treatment variable is measured with differential error. Contributing to the methodological literature about nonparametric identification in causal inference (e.g., Balke and Pearl 1997; Imai 2008; Manski 1995), we derive for the first time the sharp (i.e., best possible) bounds of the ATE while explicitly allowing for the possibility of differential misclassification of the treatment.

Identification Analysis. As advocated by Manski (1995, 2007), the goal of *nonparametric identification analysis* is to establish the domain of consensus among researchers regarding what can be learned about causal quantities of interest from the data alone. In many situations, including the one we consider in this article, the quantities of interest cannot be consistently estimated without additional assumptions. In such cases, the identification analysis only yields the *bounds* rather than the point estimates of causal effects. The width of these bounds reveals the limitations of the research design employed for a particular study regardless of the sample size. Thus, identification problems must be addressed before the problems of statistical inference, which concerns the estimation based on a finite sample. In addition, the identification analysis can formally characterize the roles of different assumptions by comparing the identification region under alternative sets of assumptions. We believe that nonparametric identification analysis offers political scientists a way to evaluate the extent to which their conclusions depend on statistical assumptions rather than empirical evidence.[3]

The result of our identification analysis reveals that, under the assumption that the mismeasured treatment is positively correlated with the true treatment, the sharp bounds are informative even in the presence of differential measurement error. Unfortunately, the resulting bounds are wide, and contrary to the conclusion under the nondifferential misclassification settings, the bounds always include zero. Thus, additional assumptions are required to further narrow the bounds. We introduce such assumptions and show how to integrate them into the identification analysis. In addition, we characterize the identification region as a function of unknown treatment assignment probability so that researchers can incorporate their qualitative knowledge (whenever available) and obtain the range of plausible values of the ATE that is more informative than the bounds under a minimum set of assumptions. Our analysis highlights the significant role played by such auxiliary information when the measurement error is differential.

Sensitivity Analysis. Another methodological contribution of this article is to propose a new *sensitivity analysis* for differential measurement error. Sensitivity analysis is a common strategy to assess the robustness of empirical evidence by examining how one's estimate varies when a key assumption is relaxed (e.g., Rosenbaum 2002).[4] In our analysis, we derive the largest amount of misclassification that could occur without altering the original conclusions. If this sensitivity parameter turns out to be relatively large, we may conclude that differential measurement error does not threaten the validity of one's findings. Therefore, the proposed sensitivity analysis represents another useful tool for evaluating the credibility of empirical findings in the possible presence of differential measurement error.

Road Map. In the next section, we discuss several examples in political science where differential measurement error may arise. Then, we describe a randomized field experiment on democratic deliberations, which motivates our formal identification analysis, and discuss the nature of the methodological challenges posed by the experiment. In the following section, we briefly summarize the methodological literature and show that measurement error is often assumed to be independent of the outcome variable. We then introduce differential measurement error and formalize credible assumptions. We show how to obtain the sharp bounds of the ATE under these assumptions and also propose a new sensitivity analysis. Then, we illustrate our proposed methods by applying them to the democratic deliberations experiment and report the findings. Finally, we summarize our main theoretical and empirical findings.

[2] A related problem is that of endogeneity (or reverse causality). We later discuss the key differences between endogeneity and differential measurement error briefly.

[3] Nonparametric identification analysis is rarely used in the discipline. Exceptions include applications to ecological inference (Duncan and Davis 1953), voter registration laws (Hanmer 2007), and suicide terrorism (Ashworth et al. 2008).

[4] A recent application of this method in political science is Quinn (2008).

Differential Measurement Error in Political Science

As we formally discuss later, measurement error in the treatment variable is differential if it is not conditionally independent of outcome given observed covariates. This type of mismeasurement is common in political science, especially when measurements are self-reported via survey. The methodology we develop in this article can be applied to various situations where differential mismeasurement may exist, including the examples we discuss in this section.

A typical example of differential measurement error is found in retrospective studies, where a causal variable of interest (i.e., treatment) is measured after the outcome already occurred. In such a case, respondents' propensity of misreporting may be directly affected by and thus correlated with the outcome variable. For example, students of political participation are often interested in the effect of (pre-election) political knowledge on voting behavior (see Galston 2001 for a review). In many large-scale surveys, including American National Election Studies and British Election Studies, however, factual knowledge questions only appear in post-election surveys. Because participation in an election could increase respondents' political interests and thus their level of political knowledge, the answers to those questions may suffer from differential measurement error. This suggests that inferences based on these responses may be invalid. For example, Mondak (1999) regresses respondents' turnout on their post-election political knowledge and discusses its implications. However, due to differential measurement error, the reported regression coefficient will be a biased estimate of causal effect.

Even if measurement is taken before outcome is observed, differential measurement error could still arise if outcome and measurement error are both correlated with an unobserved variable. This possibility is of major concern in survey research, where for many variables of interest the degree of mismeasurement is known to be correlated with respondent characteristics. For example, Prior (2009) shows via a survey experiment that the magnitude of overreporting media exposure is associated with various demographic and attitudinal characteristics, such as education and level of political interest. If these characteristics are also correlated with the outcome of interest (e.g., voting) but are not included in the analysis, then the measurement error may become differential.

In the above examples, researchers may be able to minimize differential measurement error via the use of appropriate research design. In other situations, differential error is difficult to avoid. For example, in the literature on racial cues, experimental studies have shown that respondents with strong racial predispositions tend to report characteristic views on policy issues, such as crime and government welfare spending, but only when such predispositions are primed by implicit racial cues (e.g., Mendelberg 2001; Valentino, Hutchings, and White 2002). In these studies, racial predispositions are measured after the experimental manipulations, since asking such questions might reveal researchers' intentions and thus nullify the implicitness of the cues. However, because exposure to experimental manipulations that contain racial messages potentially affects respondents' attitude toward other racial groups, measuring racial predispositions after exposure may induce differential measurement error. A recent study by Huber and Lapinski (2006) found both explicit and implicit racial cues to be equally effective when racial resentments are measured before exposure to the cues. Of course, this finding is difficult to interpret because it could be either due to the nullification of implicit cues or that of differential measurement error. This dilemma raises the question of when to best measure racial predispositions in racial cue studies. (This question is directly addressed in our separate ongoing project.)

A Motivating Example and the Framework of Causal Inference

In this section, we briefly describe the randomized field experiment that serves as both a motivation and an illustration for our proposed identification and sensitivity analysis. A more complete description of the study is given by Humphreys, Masters, and Sandbu (2006), who originally analyzed the experiment. This pioneering study also holds important implications for major theoretical debates in the fast-growing literature of democratic deliberations (see, for example, Mendelberg 2002). As explained below, the differential measurement error problem arises in this experiment primarily because the treatment variable was measured after the outcome has realized.

Background

The experiment was conducted in the Democratic Republic of São Tomé and Príncipe, an island state off the western coast of Africa. In 2004, the Santomean government held a national forum, in which citizens gathered in small groups moderated by randomly assigned discussion leaders and discussed economic policy priorities for the country. The forum took place as a result of a military coup and ensuing diplomatic interventions. From a theoretical perspective, the Santomean national forum provides an interesting case. Many political theorists and practitioners in civil society have advocated such participatory processes, on the ground that participation and deliberation lead to better, more rational collective decisions (e.g., Dryzek 1990, 2000; Habermas 1985). However, others have argued that such participatory decision making is susceptible to the undue influence of authoritative figures in the deliberation process (e.g., Sanders 1997; Young 2002). Thus, it is an open question how popular participation in decision-making processes affects the practice of democracy, and the national forum experiment provides a unique opportunity to empirically test these competing theoretical claims in real-world settings.

From a methodological perspective, the fact that discussion leaders were randomly assigned makes this national forum an ideal *randomized field experiment* where external validity can be improved without compromising much internal validity. For each of the 56 forum sites throughout the country, three to four discussion leaders were randomly selected from the pool of potential leaders, which themselves were selected from government services and civil society organizations. On the meeting day, following a plenary session led by one of the leaders on each site, participants were divided into smaller deliberation groups of about 15 to 20 people in order to discuss the country's expenditure priorities. Then, each of the discussion leaders was assigned to a deliberation group at random. The groups then discussed a set of predetermined questions, which we describe in more detail later in the article. Finally, the outcomes of the discussion were recorded, and the leaders were asked their own preferences about those questions about one week after the meetings. As explained below, this timing of measurement led to possible differential measurement error. In total, this procedure yielded 148 discussion groups, which represent our units of observation.

Causal Quantities of Interest and Immutable Characteristics

The main substantive motivation of this field experiment was to "examine the extent to which participatory processes of this form are in practice vulnerable to manipulation by political elites" (Humphreys, Masters, and Sandbu 2006, 583–84). The authors first provide evidence that the presence of leaders had a significant effect on the outcome of the group decisions. To explore why this is the case, Humphreys, Masters, and Sandbu (2006) propose the hypothesis that discussion leaders can manipulate group decision outcomes toward their own preferences. They compute the observed correlation between leaders' policy preferences and discussion outcomes (see Table 6), and they examine whether this correlation is positive.

While this is a well-defined causal quantity, we emphasize that it does *not* measure the causal effect of leaders' preferences alone on group discussion outcomes because policy preferences cannot be randomly assigned to leaders. As a consequence, leaders' preferences may be correlated with other observed and unobserved characteristics of their own, making it difficult to isolate the causal effect that can be attributed to leaders' preferences alone. For example, those with higher education may have different spending priorities and also be more persuasive as discussion leaders.

A similar problem frequently arises in social science randomized experiments. In the statistics literature on causal inference, estimating the causal effects of immutable characteristics is recognized as a challenge because it is impossible to manipulate a defining feature such as gender or race. Yet, these characteristics are associated systematically with other attributes such as income, education, or beliefs. This led some to contend "no causation without manipulation" (Holland 1986, 959). In these situations, however, one may still be able to make an inference about a redefined causal quantity.

For example, Chattopadhyay and Duflo (2004) use a randomized natural experiment to examine the causal effect of politicians' gender on policy outcomes where randomly selected local village councils were required to reserve certain policy-making positions for women. In this case, female politicians differ from their male counterparts in various characteristics other than their gender, and so the differences in observed policy outcomes cannot be solely attributed to policy makers' gender differences. Other factors such as education could be confounding factors for evaluating the effect of gender. Thus, we cannot distinguish whether it is their "femaleness" or the kind of life experience of a woman who has chosen to become a politician. Nevertheless, the study *can* estimate the effect of having a female politician.

Similarly, in the democratic deliberations experiment analyzed in this article, the random assignment of leaders does not permit definitive inferences about the causal effect of leaders' preferences per se. Therefore, we instead analyze the effect of having deliberative discussions moderated by a leader with particular policy preferences. Although we will occasionally refer to this randomized treatment simply as "leader preferences" for the sake of brevity, readers should keep in mind that our quantity of interest is the causal effect of having a leader with particular preferences rather than the effect of leaders' preferences themselves.

Along with the problem of immutable characteristics, the statistical analysis of leader preferences in the national forum experiment encounters a potential problem of measurement error. It is this methodological challenge that we will discuss below at length.[5]

The Problem of Mismeasured Treatment

As noted above, to explore how discussion leaders influence deliberation, Humphreys, Masters, and Sandbu (2006) examine whether the decision outcomes of a group are more likely to resemble its leader's policy preferences. The problem is that leaders' preferences were measured *after* the meetings had taken place and group decisions had been made. In other words, leaders' policy preferences were measured with error that is possibly correlated with the outcome.[6]

As Humphreys, Masters, and Sandbu (2006) correctly point out, this means that "we cannot discount the possibility that the preferences of the leaders are a result of, rather than a determinant of, the outcomes of the discussions" (598). For example, discussion leaders who failed to influence discussion may have an incentive to report the group decision outcome as their own policy preference so that they can be viewed as effective leaders. Another possibility is that discussion leaders may be persuaded by groups during the deliberation. If the reported preferences of leaders were influenced by the outcomes of group discussion, then the direct comparison of the discussion outcomes and reported leaders' (post-deliberation) preferences would yield biased estimates of the causal effects of leaders' (pre-deliberation) preferences on discussion outcomes. The authors undertook an informal diagnostic analysis to explore this issue. In this article, we address their concern about misclassification by conducting a formal analysis.

In particular, we study the nonparametric identification of causal effects in the presence of differential measurement error. We begin by defining the causal quantity of interest and formally stating the measurement error problem using the potential outcomes notation (e.g., Holland 1986). Let $Z_i^* \in \{0, 1\}$ be a binary treatment variable which indicates the unmeasured pre-deliberation preference of the leader randomly assigned to group i about a public policy (e.g., whether the government should spend more on primary education, $Z_i^* = 1$, or not, $Z_i^* = 0$). In addition, we use $Y_i(z^*) \in \{0, 1\}$ to define the (potential) binary outcome of group i's discussion, which is a function of the pre-deliberation preference of its discussion leader, $Z_i^* = z^*$. The observed group discussion outcome is then equal to $Y_i = Z_i^* Y_i(1) + (1 - Z_i^*) Y_i(0)$ or equivalently $Y_i = Y_i(Z_i^*)$.

Throughout this article, we assume that the true treatment assignment (i.e., pre-deliberation preferences of leaders) is unconfounded and a common support of covariate distributions exists between the treatment and control groups. This assumption is called strong ignorability in the statistics literature (Rosenbaum and Rubin 1983). Let X be a vector of observed covariates and $\mathcal{X}$ be the support of X. Then, the assumption can be written as

Assumption 1 (Strong Ignorability)

$$Z_i^* \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i = x, \quad \text{and}$$
$$0 < \Pr(Z_i^* = 1 \mid X_i = x) < 1 \quad \text{for all } x \in \mathcal{X}.$$

Assumption 1, which we maintain throughout the article, is automatically satisfied in randomized experiments, while it needs to be carefully examined in observational studies. Our proposed methods are applicable to certain observational studies where this assumption holds as well as any randomized experiments. In the context of the deliberation experiment we analyze, the second part of the assumption implies the heterogeneity of leaders' pre-deliberation preferences.

In this article, we focus on the ATE of leaders' preferences on the group discussion outcome given the observed covariates. The ATE is defined as

$$\begin{aligned}\tau^*(x) &\equiv \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i = x)\\ &= \Pr(Y_i = 1 \mid Z_i^* = 1, X_i = x)\\ &\quad - \Pr(Y_i = 1 \mid Z_i^* = 0, X_i = x),\end{aligned}$$

where the equality follows from Assumption 1 and the binary nature of Y_i. In the democratic deliberation experiment, a positive (negative) effect implies that leaders can influence the outcome of their group discussions in the same (opposite) direction as their true preferences.

[5] An alternative quantity of interest estimated in the Humphreys, Masters, and Sandbu (2006) study (see Section V), which we do not examine in this article, is the causal effect of leaders' *presence* on group discussion outcomes, rather than that of leaders' *preferences*. Here, there is no problem with immutable characteristics, and causal inference could be straightforward (though the fact that the control group was lacking in this experiment introduces another identification problem). While we cannot directly randomize preferences, it is possible to randomize whether or not each group has a discussion leader and estimate the effect of the presence of a leader on group discussion outcomes.

[6] Since leader's presence can be measured without error, the original analysis regarding the causal effects of leader's presence does not suffer from the differential measurement error problem we study in this article.

Next, let $Z_i \in \{0, 1\}$ be the leader's reported (post-deliberation) policy preference that was measured after the experiment was conducted. Since it represents the mismeasured treatment variable, Z_i is in general not equal to Z_i^*. If we ignore the measurement error problem entirely and assume $Z_i = Z_i^*$ for all i, we estimate the following quantity:

$$\begin{aligned}\tau(x) &\equiv \mathbb{E}(Y_i \mid Z_i = 1, X_i = x) \\ &\quad - \mathbb{E}(Y_i \mid Z_i = 0, X_i = x) \\ &= \Pr(Y_i = 1 \mid Z_i = 1, X_i = x) \\ &\quad - \Pr(Y_i = 1 \mid Z_i = 0, X_i = x). \qquad (1)\end{aligned}$$

Clearly, a naïve comparison of this kind would lead to a biased estimate of the quantity of interest (i.e., $\tau(x) \neq \tau^*(x)$). The causal effect would be overestimated if, for example, the leaders' own involvement in the deliberation process made their opinions closer to those of their group members.

Before moving on, we address one common misconception. While the difference between Z_i^* and Z_i could be attributed to the presence of reverse causality, it should be stressed that a differential measurement error problem is distinct from an endogeneity problem. Endogeneity implies that the causal variable of interest is itself correlated with the potential outcomes due to the lack of strong ignorability, i.e., Assumption 1. In contrast, our treatment variable (i.e., the pre-deliberation preference of leaders) does not suffer from this identification problem because the treatment is randomized and causally precedent to the outcome (no post-treatment bias). Thus, the possibility of reverse causality only affects the measurement of the causal variable but not the variable itself.[7]

In what follows, we formally characterize the identification region of the true ATE by deriving the sharp (i.e., best possible) bounds under various assumptions. Our identification analysis establishes the exact degree to which the data-generating mechanism, when combined with additional assumptions, is informative about the true ATE in the presence of possible misclassification of treatment. We also offer a new sensitivity analysis that allows researchers to evaluate the robustness of their conclusions. Because our analysis is nonparametric, the results do not rely on distributional or other parametric assumptions of regression models commonly used in the political science research.

The Proposed Methodology

In this section, we first briefly review the related statistics and econometrics literature on measurement error models (see Carroll et al. 2006 and Fuller 1987 for comprehensive monographs). We show that the existing theoretical results are not applicable to the democratic deliberation experiment and other studies that suffer from a similar measurement error problem. We then derive the sharp bounds on the ATE and develop a new sensitivity analysis by relaxing the key assumption of the previous literature.

Nondifferential Measurement Error: A Brief Review of Literature

In *classical* error-in-variables models, the measurement errors are assumed to be independent of their true value. In such models, measurement error generally leads to attenuation bias. For example, it is well known that a linear least-squares regression will underestimate the coefficient when an explanatory variable is subject to the classical measurement error. This attenuation bias arises even when there are control variables in the regression which are correlated with the true values of the mismeasured explanatory variable of interest (e.g., Wooldridge 2002, Section 4.4.2). A large number of existing studies examine the identification and estimation problems in the presence of classical measurement error, based on the existence of auxiliary information, such as a repeat measurement and other instrumental variables (e.g., Buzas and Stefanski 1996; Carroll et al. 2004; Hausman et al. 1991).

However, this classical errors-in-variables assumption is necessarily violated for binary variables because errors are always negatively correlated with their correct values (see Cochran 1968). Thus, in the econometrics literature on measurement error models for binary treatment variables, researchers instead assume that the measurement error is conditionally independent of the outcome given the true value as well as the observed

[7] In contrast, the endogeneity problem does affect the identification of the reverse causal effects, i.e., the causal effect of the group decision on leaders' post-deliberation preferences. Although this quantity is not studied either in this article or in the original study of Humphreys, Masters, and Sandbu (2006), it may be of interest to researchers who would like to know the degree to which groups can influence leaders' preferences rather than vice versa. The difficulty of identifying this quantity originates in the fact that group decisions are not randomly determined. This is not a measurement error problem, since both leaders' post-deliberation preferences (the outcome variable) and group decisions (the treatment variable) are directly measured. Rather, it is an endogeneity (or selection) problem where the treatment variable is not randomized. This distinction is important because the identification results already exist in the literature for the situation of endogeneity (see, e.g., Manski 1990).

covariates (e.g., Aigner 1973; Bollinger 1996; Klepper 1988; Lewbel 2007). In this case, the measurement error is said to be *nondifferential*, and the assumption can be written as follows.

Assumption 2 (Nondifferential Measurement Error)

$$Z_i \perp\!\!\!\perp Y_i \mid Z_i^*, \quad X_i = x \quad \text{for all } x \in \mathcal{X}.$$

The assumption is also equivalent to the statement that Z_i is a *surrogate* as defined in the statistics literature (Carroll et al. 2006, Section 2.5).

Another critical assumption made in the literature is that the degree of measurement error is not "too large" in the following sense.

Assumption 3 (Restriction on the Degree of Measurement Error)

$$\begin{aligned}&\Pr(Z_i = 0 \mid Z_i^* = 1, X_i = x)\\ &\quad + \Pr(Z_i = 1 \mid Z_i^* = 0, X_i = x)\\ &< 1 \quad \text{for all } x \in \mathcal{X}.\end{aligned}$$

The first and second terms in the left-hand side equal the probabilities of misclassifying $Z_i^* = 1$ as $Z_i = 0$ and $Z_i^* = 0$ as $Z_i = 1$, respectively. Since the sum of these probabilities represents the total probability of misclassification, this assumption implies that the observed treatment status is at least informative about the true treatment status. In particular, it can be shown that this assumption implies a positive correlation between the true and mismeasured treatment variables (though the converse does not necessarily hold).

Although Assumptions 1–3 are not strong enough to identify the true ATE or $\tau^*(x)$, one can derive its sharp bounds. In particular, Lewbel (2007) shows that under these three assumptions, the naïve estimator based on the mismeasured treatment variable, $\tau(x)$, equals the sharp lower bound of the true ATE. Prior to Lewbel (2007), Bollinger (1996) shows (in the context of the linear regression model) that under the additional assumption that the outcome variable has a finite variance, the ATE also has a finite sharp upper bound. In addition, recent studies have explored additional assumptions and auxiliary information that can be used to achieve *point-identification* (e.g., Black, Berger, and Scott 2000; Mahajan 2006).

Unfortunately, these existing identification results in the literature are not applicable to the deliberation experiment or other studies which suffer from a similar differential measurement error problem. While Assumption 3 is reasonable, the assumption of nondifferential measurement error is unlikely to hold in our setting. Indeed, Assumption 2 implies that the conditional distribution of the self-reported leader preferences given the true preferences does not depend on group discussion outcomes, i.e., $\Pr(Z_i = 1 \mid Y_i, Z_i^*, X_i) = \Pr(Z_i = 1 \mid Z_i^*, X_i)$. Yet, as explained in the previous section, the leaders' preferences may have been influenced by their involvement in the discussion. Thus, we relax this assumption below by allowing the measurement error to directly depend on the discussion outcomes. Our analysis fills the gap in the methodological literature where the identification of causal effects under the assumption of differential measurement error has not been systematically studied.

Limited Identification Power under Differential Measurement Error

Next, we study the implications of relaxing Assumption 2 for the identification of the ATE. We begin by considering the implications of the assumptions that are generally applicable and fairly weak when the treatment status is measured with differential error. First, the following proposition shows that Assumptions 1 and 3 have only limited identification power in the presence of differential measurement error (note that Assumption 1 alone has no identification power; see Appendix A.1). Although the resulting sharp bounds are always narrower than the original bounds $[-1, 1]$, only either the upper or lower bound can be informative.

Proposition 1 (Informativeness of Assumptions 1 and 3). *Let* $[\alpha, \beta]$ *denote the sharp bounds of the average treatment effect under Assumptions 1 and 3. Then,*

1. $\alpha = -1$ *if and only if* $\Pr(Z_i = 1 \mid Y_i = 1) < \Pr(Z_i = 1 \mid Y_i = 0)$,
2. $\beta = 1$ *if and only if* $\Pr(Z_i = 1 \mid Y_i = 1) > \Pr(Z_i = 1 \mid Y_i = 0)$.

Thus, the bounds are always informative.

A proof is given in Appendix A.2. The proposition states that if we use $[\alpha, \beta]$ to denote the sharp bounds under Assumptions 1 and 3, we will have either $\alpha > -1$ or $\beta < 1$, but neither $-1 < \alpha < \beta < 1$ nor $-\alpha = \beta = 1$ holds. Thus, either the upper or lower bound is always informative but both cannot be informative simultaneously.

Proposition 1 suggests that under fairly weak and generally applicable assumptions about the magnitude of differential misclassification, the sharp bounds are always informative but only to a limited degree. This contrasts with the previous results in the literature, which are only applicable to nondifferential measurement error of treatment. In the presence of differential misclassification, informative inference is difficult to make unless researchers invoke additional assumptions that are specific to and credible within their own study. As an illustration, we now consider such assumptions in the context of the democratic deliberation experiment. Throughout the rest of the article, we maintain Assumptions 1 and 3.

Additional Assumptions for More Informative Inference

Below, we consider two kinds of additional assumptions. The first set of assumptions is concerned about the nature of measurement error. Although they are introduced in the context of the democratic deliberation experiment, similar assumptions may be applicable to other studies. To generate credible assumptions, it is often fruitful to consider the mechanisms underlying the misclassification. As explained above, the authors of the original analysis are concerned about "the possibility that the [reported] preferences of the leaders are a result of, rather than a determinant of, the outcomes of the discussions" (Humphreys, Masters, and Sandbu 2006, 598). Our assumptions directly address this concern.

The first assumption about the nature of measurement error is related to the possibility that some groups can persuade leaders while others fail to do so. In particular, we assume that groups who would always follow their leader's (true) preferences lack the ability to persuade leaders. To formalize this idea, we stratify groups into four mutually exclusive "types" based on the potential values of the outcome given their treatment status. We use $S_i \in \{c, a, n, d\}$ to indicate group i's type. This formulation, called *principal stratification* (Frangakis and Rubin 2002), is particularly useful for formally incorporating assumptions about potential outcomes. Type c groups represent those who $\underline{c}$omply with their leader by yielding the same discussion outcome as the leader's preference, i.e., $(Y_i(1), Y_i(0)) = (1, 0)$, whereas type a groups would $\underline{a}$lways favor the given policy regardless of what their leaders prefer, i.e., $(Y_i(1), Y_i(0)) = (1, 1)$. Type n groups would $\underline{n}$ever favor the given policy, i.e., $(Y_i(1), Y_i(0)) = (0, 0)$, and type d groups $\underline{d}$efy their leader by always coming to the decision opposite to their leader's preference, i.e., $(Y_i(1), Y_i(0)) = (0, 1)$.

Given this notation, our first assumption can be formalized as follows.

Assumption 4 (No Persuasion by Compliant Groups)

$$\Pr(Z_i = z \mid S_i = c, Z_i^* = z) = 1, \quad \text{for } z \in \{0, 1\}.$$

The assumption implies that no misclassification occurs for leaders who are assigned to compliant groups, which make the decision in agreement with their leader's (true) preferences whatever they may be. This seems plausible as long as one assumes that the groups' potential responses well approximate their capability of persuading the leaders into new preferences. Although these group types are defined with respect to potential outcomes and thus not directly observable, they serve as a useful device to express our substantive assumptions in a form that can be directly incorporated into our analytical framework. Also, this formulation can easily incorporate weaker versions of Assumption 4 using an inequality rather than equality, e.g., $\Pr(Z_i = z \mid S_i = c, Z_i^* = z) > \Pr(Z_i = 1 - z \mid S_i = c, Z_i^* = z)$ for $z \in \{0, 1\}$.

The second assumption concerns leaders' incentives to misreport their preferences in order to appear effective in influencing group decisions. This scenario leads to the assumption that leaders do not misreport their true preferences if the actual (rather than potential) group decision outcome agrees with their true preferences. Then, unlike Assumption 4, this assumption can be expressed as a constraint on the distribution of realized (but not necessarily observed) variables rather than that of principal strata.

Assumption 5 (Leaders' Incentives)

$$\Pr(Z_i = z \mid Y_i = z, Z_i^* = z) = 1, \quad \text{for } z \in \{0, 1\}.$$

Mathematically, the difference between Assumptions 4 and 5 is that the former conditions on the principal strata with respect to potential outcomes while the latter conditions on the observed outcome. In substantive terms, this assumption seems plausible given the role such "leaders" are expected to play in many cultural settings. However, we need additional information about their socioeconomic and cultural background to fully assess the plausibility of this assumption.

Although Assumptions 4 and 5 are formulated differently, they both address the concern about the nature of differential measurement error where groups can influence leaders' reported preferences. Indeed, these assumptions are mathematically related. In particular, Assumption 5 implies Assumption 4 (but not vice versa) because Assumption 5 is equivalent to $\Pr(Z_i = z \mid S_i = c, Z_i^* = z) = \Pr(Z_i = 1 \mid S_i = a, Z_i^* = 1) = \Pr(Z_i = 0 \mid S_i = n, Z_i^* = 0) = 1$ for $z \in \{0, 1\}$. Thus, Assumption 5 puts restrictions on the possibility of differential measurement error for the groups belonging to $S_i \in \{a, n\}$ as well as the compliant groups with $S_i = c$ (i.e., Assumption 4).

In different studies, various substantive knowledge can be brought into one's analysis. Applied researchers can often express these assumptions as restrictions on the distributions of either potential outcomes or realized variables in a manner similar to the two assumptions described here.

Derivation of the Sharp Bounds under Additional Assumptions

Having introduced the assumptions, we now show how to obtain the sharp bounds on the true ATE under different combinations of these assumptions. Throughout this subsection and for the rest of the article, we maintain Assumptions 1 and 3. Recall that these two assumptions imply that the treatment assignment is strongly ignorable and that the correlation between the true and mismeasured treatment status is positive.

The analytical derivation of the sharp bounds under several assumptions can be a difficult task. Our strategy is to formulate the problem as one of constrained linear optimization. The advantage of this approach is that the sharp bounds can be easily found (whenever they exist) once all the assumptions are expressed in the form of linear equality or inequality constraints. Balke and Pearl (1997) used this strategy to derive the sharp bounds on the ATE in randomized experiments with noncompliance. Following the literature, we focus on the derivation of large sample bounds. The confidence intervals for the bounds can be calculated using a bootstrap method of Beran (1988; see Imai and Soneji 2007 for details).

Sharp Bounds under Assumptions 1, 3, and 5. To illustrate our approach, we first provide the exact analytical solution for the most informative case. In particular, we assume that Assumptions 1, 3, and 5 are satisfied simultaneously. (Recall that Assumption 5 implies Assumption 4.) We begin by letting $P_{yz} \equiv \Pr(Y_i = y, Z_i = z)$ represent the probabilities of observed strata where $y, z \in \{0, 1\}$. Since both Y_i and Z_i, the realizations of the discussion outcome and self-reported leader preference, are observed, the data-generating mechanism alone identifies these probabilities. Next, we denote the probability of the true treatment status by $Q \equiv \Pr(Z_i^* = 1)$. This quantity is not directly identifiable from the observed data, but Assumption 1 guarantees $0 < Q < 1$. Furthermore, we define $\psi_{yz} = \Pr(Y_i = y, Z_i = z \mid Z_i^* = 1)$ and $\phi_{yz} = \Pr(Y_i = y, Z_i = z \mid Z_i^* = 0)$ for $y, z \in \{0, 1\}$. Then, we can rewrite the probabilities of observed strata in terms of Q, ψ_{yz}, and ϕ_{yz},

$$P_{yz} = (1 - Q)\phi_{yz} + Q\psi_{yz}. \qquad (2)$$

This represents the relationship between the observable joint probabilities and the unobservable conditional probabilities that must be satisfied. In other words, it can be interpreted as a set of baseline constraints when inferring the true ATE in the presence of differential misclassification.

Next, Assumption 3 is equivalent to

$$\sum_{y=0}^{1}(\psi_{y0} + \phi_{y1}) < 1. \qquad (3)$$

Finally, using the Bayes theorem, Assumption 5 can be equivalently expressed as

$$\phi_{01} = \psi_{10} = 0. \qquad (4)$$

Now, under Assumption 1 the true ATE equals the following expression:

$$\tau^* = \sum_{z=0}^{1} \psi_{1z} - \sum_{z=0}^{1} \phi_{1z}. \qquad (5)$$

Thus, under Assumptions 1, 3, and 5, the identification region of the ATE can be obtained by solving the linear programming problem where the objective function is equation (5) and the constraints are given by equations (2), (3), and (4). The following proposition characterizes the identification region of the true ATE as a function of the (unidentifiable) treatment assignment probability Q as well as identifiable quantities. Using this result, we derive the sharp bounds of the ATE.

Proposition 2 (Identification of the Average Treatment Effect). *Suppose that Assumptions 1, 3, and 5 hold. Then, the following results are obtained.*

1. *The identification region of τ^* can be expressed as a function of Q and the identifiable quantities in the following manner.*

$$\max\left(-\frac{P_{10} + P_{11}}{1 - Q}, -\frac{P_{01}}{Q} - \frac{P_{10}}{1 - Q}, -\frac{P_{00} + P_{01}}{Q}\right) \leq \tau^* \leq \min\left(\frac{P_{00}}{1 - Q} - \frac{P_{01}}{Q}, \frac{P_{11}}{Q} - \frac{P_{10}}{1 - Q}\right).$$

2. *The sharp upper and lower bounds are given by*

$$\max\left\{-1, \min\left(P_{00} - \frac{P_{01}P_{10}}{P_{11}} - 1, P_{11} - \frac{P_{01}P_{10}}{P_{00}} - 1\right)\right\} \leq \tau^* \leq \tau.$$

A proof is given in Appendix A.3.[8] Proposition 2 implies that without the knowledge of Q, the sharp upper bound equals the naïve estimate of the ATE, which

[8] To be precise, these and other bounds in this article correspond to the supremum and infinimum of the identification region rather than its maximum and minimum. This slight abuse of notation does not alter the interpretation of the bounds.

ignores the measurement error problem altogether, i.e., $Q = P_{01} + P_{11}$. On the other hand, the sharp lower bound of the ATE never exceeds zero. This result stands in contrast to the case of nondifferential measurement error reviewed earlier in this section where measurement error is found to induce an attenuation bias. Here, we find that the bias goes in the opposite direction when measurement error is differential. Indeed, Proposition 2 suggests that the differential misclassification of treatment can seriously hamper one's causal inference.

The expression of the identification region given in Proposition 2 is useful because it is written as a function of Q, which is the true treatment assignment probability. Here, we describe two analytical approaches that exploit this feature. First, as is the case in the democratic deliberation experiment, researchers often have auxiliary information about the range of possible values of Q. Even though the value of Z_i^* itself can never be observed for any i, such auxiliary information might help tighten the sharp lower and upper bounds and thus significantly improve the identification result. Second, Proposition 2 implicitly provides the sharp bounds on Q, which is given by

$$Q \in [P_{01}, 1 - P_{10}]. \qquad (6)$$

Thus, with some knowledge about Q, we can assess the plausibility of Assumptions 1, 3, and 5 by letting Q take different values within this range and observing how the bounds change. If one believes that the plausible values of Q lie outside of this range, then those assumptions may not be satisfied simultaneously. Later, we illustrate how to conduct analyses like these with the actual data from the deliberation experiment.

Sharp Bounds under Other Assumptions. Next, we show how to obtain the sharp bounds under Assumptions 1, 3, and 4. Unlike the case considered above, Assumption 4 places a restriction on the distribution of principal strata. Thus, it is necessary to represent the constraints in terms of principal strata. Let π_{sz} and η_{sz} denote $\Pr(S_i = s, Z_i = z \mid Z_i^* = 1)$ and $\Pr(S_i = s, Z_i = z \mid Z_i^* = 0)$ for $s \in \{c, a, n, d\}$ and $z \in \{0, 1\}$, respectively. Although these probabilities are defined with respect to unobservable principal strata, they can be re-expressed in terms of outcomes because a respondent's type represents her potential outcome given the treatment she received. For example, $\Pr(Y_i = 1, Z_i = z \mid Z_i^* = 1) = \Pr(S_i = c, Z_i = z \mid Z_i^* = 1) + \Pr(S_i = a, Z_i = z \mid Z_i^* = 1) = \pi_{cz} + \pi_{az}$.

Then, we can rewrite the probabilities of observed strata in terms of Q, π_{jk}, and η_{sz},

$$P_{0z} = (1 - Q)(\eta_{cz} + \eta_{nz}) + Q(\pi_{nz} + \pi_{dz}), \qquad (7)$$

$$P_{1z} = (1 - Q)(\eta_{az} + \eta_{dz}) + Q(\pi_{cz} + \pi_{az}), \qquad (8)$$

for $z \in \{0, 1\}$. Similarly, Assumption 3 is equivalent to

$$\sum_{j \in \{c,a,n,d\}} (\eta_{j1} + \pi_{j0}) < 1. \qquad (9)$$

Finally, using the Bayes theorem, Assumption 4 can be equivalently expressed as

$$\begin{aligned} &\frac{\Pr(Y_i = 1, Z_i = 0 \mid Z_i^* = 1)}{\Pr(Y_i = 1 \mid Z_i^* = 1)} \\ &\quad = \frac{\Pr(Y_i = 0, Z_i = 1 \mid Z_i^* = 0)}{\Pr(Y_i = 0 \mid Z_i^* = 0)} \\ &\quad = 0 \iff \pi_{c0} = \pi_{a0} = \eta_{c1} = \eta_{n1} = 0. \end{aligned}$$

Thus, Assumption 4 provides additional restrictions on the ATE, which can now be written as

$$\tau^* = \pi_{c1} + \pi_{a1} - (\eta_{a1} + \eta_{d1} + \eta_{a0} + \eta_{d0}).$$

The sharp bounds on the ATE can now be derived numerically using the standard algorithm for linear programming problems subject to the restrictions given in equations (7), (8), and (9). Likewise, even when only Assumptions 1 and 3 hold, the problem can be formulated in the form of linear programming and the standard algorithm can be applied. Unfortunately, unlike the case considered in Proposition 2, it is difficult to obtain exact analytical expressions for the bounds in these cases. In the next section, we report the sharp bounds under these two sets of assumptions obtained by this methodology, along with the analytically derived bounds given in Proposition 2.

Sensitivity Analysis

The results given in Proposition 2 demonstrate the serious consequences of differential measurement error. Even under the strongest set of assumptions we consider, the bounds are wide and always contain zero unless researchers possess auxiliary information about the true treatment assignment probability. This illustrates that when differential misclassification is present, adding assumptions may prove insufficient to draw a substantively interesting conclusion.

More importantly, these substantive assumptions themselves may be controversial since they cannot be tested directly from empirical data. For example, leaders may have an opposite incentive of *concealing* their influence by reporting policy preferences that are contrary to the group decision outcomes. This alternative scenario is plausible if the leaders are concerned about being viewed as violating their supposed role of facilitating deliberations. If this is the case, we may assume $\Pr(Z_i = z \mid Y_i = z, Z_i^* = z) = 0$, rather than the equality given in Assumption 5. Obviously, this would lead to bounds that are very different from those given in Proposition 2.

Here, we offer a new *sensitivity analysis* that can be used in the situations where strong assumptions such as Assumption 4 and 5 are unavailable. Under such circumstances, we propose to derive the conditions under which the original conclusions based on the mismeasured treatment variable still hold. That is, we ask the question, "Can the experiment be saved in the presence of differential measurement error?"

More specifically, we vary the magnitude of differential measurement error and observe how the bounds of the ATE change as a function of this parameter. Then, we calculate the maximum size of total misclassification probability that can be present while still guaranteeing the conclusions based on the mismeasured treatment. This is done by replacing Assumption 3 with the following:

$$\begin{aligned}&\Pr(Z_i = 0 \mid Z_i^* = 1, X_i = x)\\&\quad + \Pr(Z_i = 1 \mid Z_i^* = 0, X_i = x) \leq \rho\\&\quad \text{for all} \quad x \in \mathcal{X},\end{aligned} \tag{10}$$

where $0 \leq \rho \leq 1$, and then deriving the bounds of the ATE using various values of ρ. Since $\rho = 0$ implies no measurement error, the maximum value of this error parameter represents the maximum magnitude of measurement error that can exist without contradicting the conclusions obtained by incorrectly ignoring the measurement error. Put more simply, our sensitivity analysis asks how much measurement error we could accommodate in order for the original conclusion to remain valid.

The analytical strategies used for the derivation of the sharp bounds are directly applicable to the proposed sensitivity analysis under different sets of additional assumptions. Namely, we calculate the sharp bounds while fixing ρ to various nonnegative values. This can be done by solving the linear optimization problem as before with the following additional constraint:

$$\sum_{y=0}^{1}(\psi_{y0} + \phi_{y1}) \leq \rho\,. \tag{11}$$

We can then obtain the maximum value of ρ which yields the sharp bounds that do not overlap with zero. A large value of this maximum value implies that the conclusions obtained from the mismeasured treatment are robust to the possible existence of differential measurement error.

Empirical Results

To illustrate our proposed methods, we now apply our identification results to the actual data from the randomized field experiment on democratic deliberations described earlier.

Data

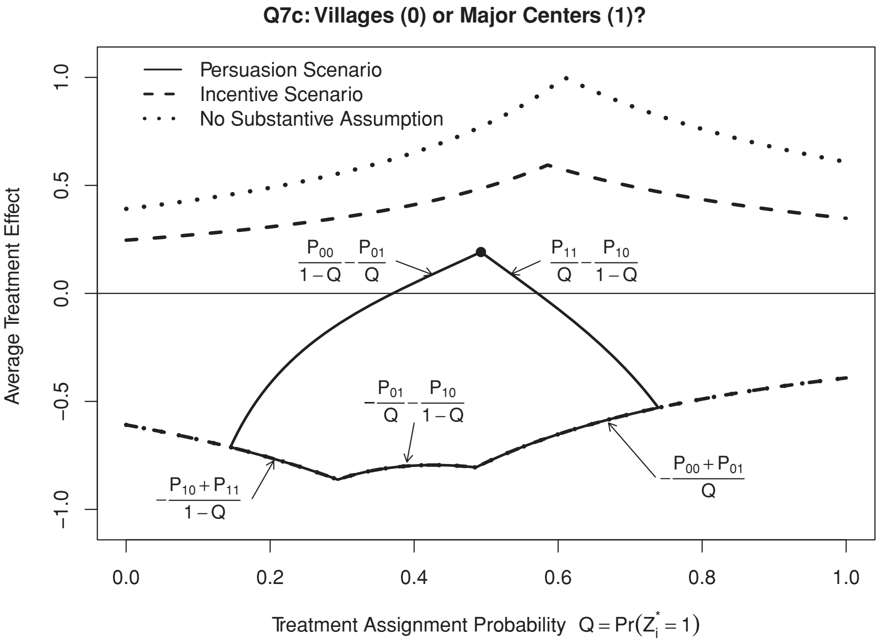
Of the 12 items on the questionnaire examined by Humphreys, Masters, and Sandbu (2006), we focus on five and examine them in detail. The first four questions concern the government's spending priorities: whether resources should be spent on nearby health clinics or full reliable hospitals (Q3), primary or secondary education (Q4a), road improvement or mass transportation (Q7b), and village roads or roads between centers (Q7c). The remaining question asks whether participants prefer to spend or invest the money they obtained as a windfall in their hypothetical community savings account (Q11a). For the actual texts used in the questionnaire, see Humphreys, Masters, and Sandbu (2006). The data set contains the outcomes of group discussions as well as the self-reported preference of the discussion leaders for each question. As explained earlier in the article, the latter is likely to be measured with differential error because the measurement was taken after the group discussions were completed.

Identification Analysis

We begin by examining one of the five questions in detail. Figure 1 represents our identification results regarding the ATE for Question 7c, which asks whether the government should spend extra resources on village roads ($Y_i = 0$) or roads between major centers ($Y_i = 1$). The vertical axis indicates the ATE of leaders' preferences on group discussion outcomes. The horizontal axis represents the (unknown) treatment assignment probability, i.e., the true proportion of leaders who prefer village roads to roads between centers. Since we do not observe the sample proportion corresponding to this probability Q, we analyze how the sharp bounds vary as functions of Q.

In Figure 1, the area between the dotted lines represents the identification region under Assumptions 1 and 3, i.e., two assumptions that are fairly weak and generally applicable. The other lines represent the sharp bounds under the sets of assumptions that include the two substantive assumptions analyzed in the previous section. Specifically, Assumptions 1, 3, and 5 correspond to the

FIGURE 1 Estimated Sharp Bounds on the Average Treatment Effect (ATE) under Various Combinations of Assumptions for Question 7c



The figure shows the estimated sharp bounds under Assumptions 1, 3, and 5 (solid lines), Assumptions 1, 3, and 4 (dashed lines) and Assumptions 1 and 3 (dotted lines) as functions of the true treatment assignment probability, i.e., $Q = \Pr(Z_i^* = 1)$ (the horizontal axis). For this question, Q represents the true population fraction of leaders who prefer to build roads between major centers ($Z_i^* = 1$) rather than village roads ($Z_i^* = 0$). The binary outcome variable represents whether a group decides that the resources should be spent on roads between centers ($Y_i = 1$) rather than village roads ($Y_i = 0$). The solid circle represents the case where there is no measurement error and the naïve estimator given in equation (1) is unbiased for the true ATE.

"persuasion scenario" and Assumptions 1, 3, and 4 to the "incentive scenario" in the legend. The solid circle located at the intersection of the two lines indicates the situation where there is no measurement error and hence the naïve estimator provided in equation (1) is unbiased for the true ATE.

The figure shows that the sharp lower bound is greater than -1 for any value of the treatment assignment probability under Assumptions 1 and 3. This verifies Proposition 1, which states that either the upper or lower bound (but never both) is always guaranteed to be informative under these two assumptions. In the present case, it is the lower bound that turns out to be informative, as the condition $\Pr(Y_i = 1 \mid Z_i = 1) > \Pr(Y_i = 1 \mid Z_i = 0)$ is satisfied $(0.706 > 0.514)$.

As our analytical results indicated, little can be learned under these two general assumptions for this question as demonstrated by the wide estimated bounds of the ATE, $[-0.862, 1]$. In contrast, under the strongest set of assumptions, the estimated upper bound decreases to 0.192, which occurs when the true treatment assignment probability is approximately 0.493. The estimated bounds under a slightly weaker set of assumptions (Assumptions 1, 3, and 4) are indicated by the dashed lines. Notice that in this case additional assumptions greatly improve the upper bound. However, they fail to improve the lower bound, and the bounds are still wide overall. As shown in Proposition 2, even with the relatively strong assumptions, the bounds always contain zero, suggesting that differential measurement error has serious consequences in this case.

Another important feature of Figure 1 is that it shows the estimated range of the true treatment assignment probability that is consistent with Assumptions 1, 3, and 5. This range is calculated as $[0.145, 0.739]$ using equation (6). As we explained earlier briefly, if we have some auxiliary information about the range of possible values of Q, we can judge the plausibility of our assumptions by examining whether these two ranges overlap with one another. For weaker sets of assumptions, auxiliary information about Q is still helpful because such information generally leads to narrower bounds of the ATE.

In fact, in the democratic deliberations experiment, there exists potentially useful auxiliary information from a survey the original research team conducted on a random sample from all the citizens (see Humphreys, Masters, and Sandbu 2006, 594–95 for more details). Although the primary purpose of this portion of the study was to examine differences between pre-forum and post-forum individual attitudes, we may exploit the fact that the survey asked roughly the same set of questions as in the national forum. For example, according to Humphreys, Masters, and Sandbu (2006), of the total of 266 individuals who were interviewed prior to the forum meetings, 19% answered that they preferred improving major roads as opposed to village roads. While the average preferences of discussion leaders might differ from those of survey respondents, we may still use this percentage as a rough estimate of Q assuming that these two types of population are similar in their preferences.[9] In this case, the estimated bounds of the ATE are sharpened to $[-0.751, -0.459]$, which no longer contains zero. Alternatively, we can assume that the leaders' average preferences are within, say, 5% of those of the survey respondents. In that case, the ATE falls between the range of $[-0.801, -0.300]$, which is somewhat wider but still does not contain zero. These results suggest that the effect of leaders' preferences on group decision outcomes might have been negative for this question. This is an opposite conclusion from the one based on the naïve estimate.

Figure 2 presents the results of the similar identification analysis for the other four questions. First of all, it is evident from the figure that inferences based only on self-reported leader preferences could be misleading. For example, the upper left panel shows that about 81.4% of the discussion leaders stated they preferred spending government resources on reliable hospitals as opposed to local clinics. The ATE would be about as high as 0.495 if this was the case. However, our analysis reveals that even under the strongest set of assumptions, the true proportion of such leaders (i.e., Q) could have been anywhere between 0.286 and 0.971 and the true ATE as low as -0.858. As shown empirically in Figure 2 and is formally characterized in equation (6), this range is wide in all of the questions we have analyzed (0.081 to 0.730 in Question 4a, 0.172 to 0.895 in Question 7b, and 0.352 to 0.879 in Question 11a).

Not surprisingly, the possible range of the ATE is also wide and always contains zero even under the strongest assumptions we consider. For instance, the range is estimated to be -0.754 to 0.349 in Question 4a, which asks whether government should spend more resources on primary schools or secondary schools (upper right panel). For the other two questions, however, the bounds on the ATE are somewhat narrower and are suggestive about the direction of the ATE. In the question about road improvement versus public transportation (Question 7b, lower left panel), the possible range of the ATE is estimated to be $[-0.999, 0.002]$, and the range turns out to be $[-0.945, 0.092]$ for the question about the use of hypothetical community resource obtained as a windfall (Question 11a, lower right panel). However, even in these two cases, the ATE is still estimated to be positive for some range of Q that includes the observed treatment probability (0.197 and 0.704, respectively).

Again, auxiliary information about Q can be helpful to sharpen inference. For example, in the pre-forum survey, the proportion of the respondents who preferred hospitals over local clinics was approximately 58% (Question 3), and the proportion of those who preferred primary schools to secondary schools was about 28% (Question 4a). Similarly, about 46% of the respondents preferred to invest windfalls rather than receive them now (Question 11a). If we use this information as estimates of Q in each question, the sharp bounds of the ATE under Assumptions 1, 3, and 5 are significantly tightened to $[-0.561, -0.118]$ in Question 3, $[-0.665, 0.180]$ in Question 4a, and $[-0.875, -0.439]$ in Question 11a.

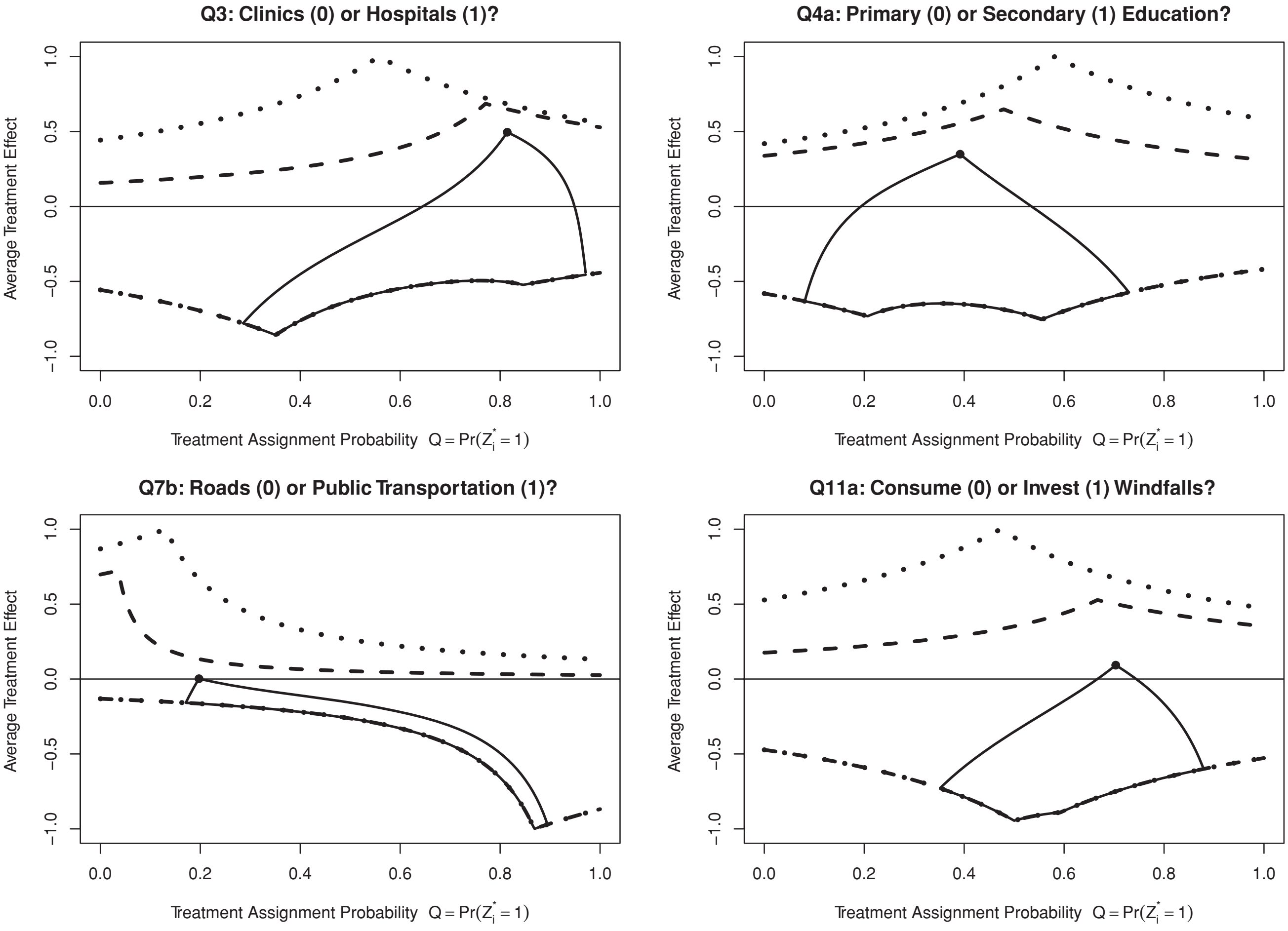
In sum, our identification analysis of the democratic deliberation experiment shows that the risk of differential measurement error makes it difficult to draw a definitive conclusion about the causal effects of leaders' preferences on group decision outcomes. The bounds on the ATE are wide and always contain zero even under the strongest set of assumptions we consider. Thus, to make informative inferences, researchers must rely on available auxiliary information.

Sensitivity Analysis

As demonstrated above, differential measurement error leads to serious identification problems in the democratic deliberation experiment, even after adding substantive assumptions about the error-generating mechanism. At the same time, these assumptions are strong and may lack

[9] Although we recognize that this is a rather tenuous assumption, the analysis presented here serves as an illustration about how one might exploit the existence of auxiliary information about Q.

FIGURE 2 Estimated Sharp Upper and Lower Bounds on the Average Treatment Effect for the Other Four Questions



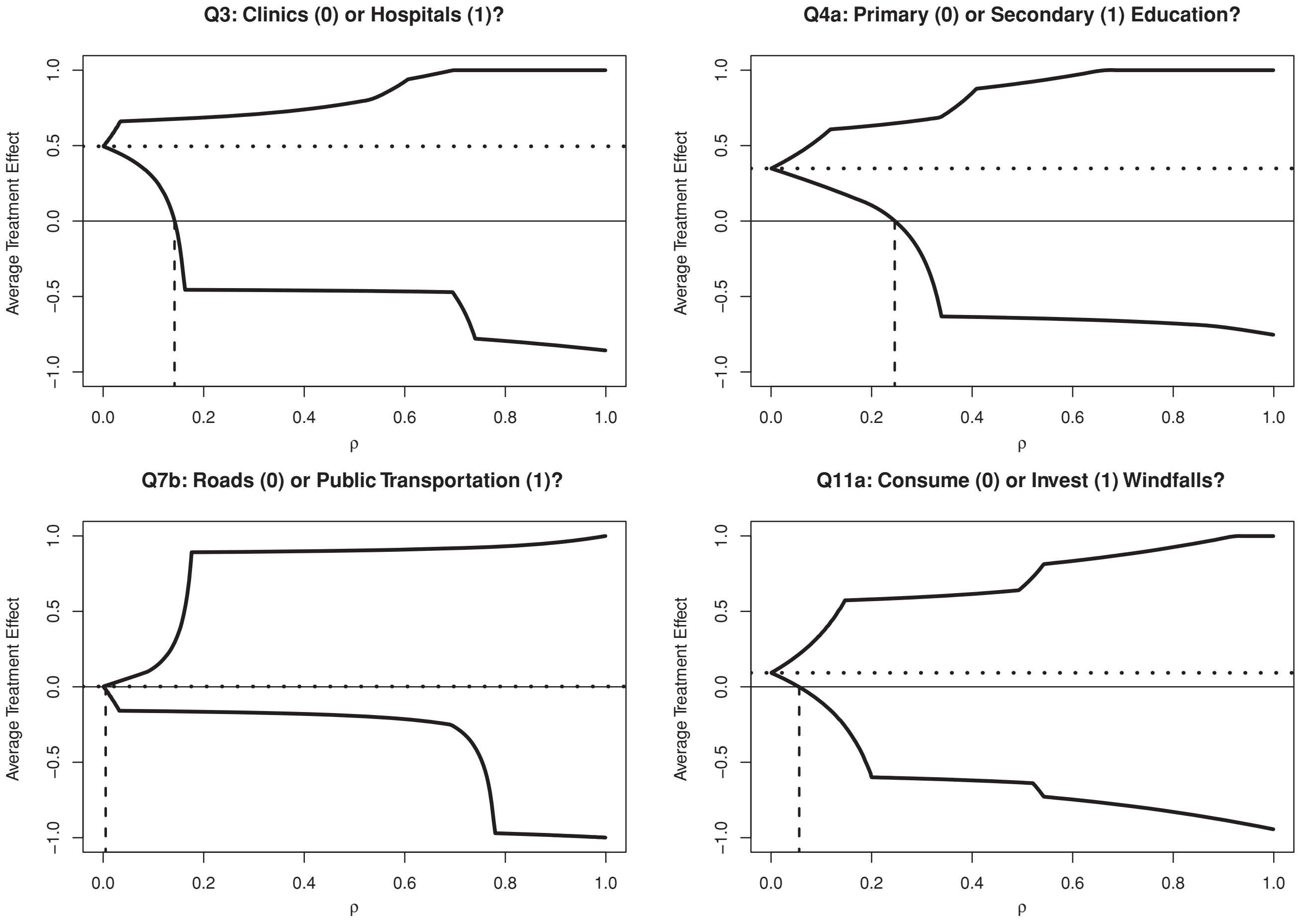
Each question is concerned about whether a leader (or a group after deliberation) prefers to spend resources on [Q3] clinics (coded as 0) or hospitals (coded as 1); [Q4a] primary (coded as 0) or secondary (coded as 1) education; [Q7b] roads (coded as 0) or public transportation (coded as 1); and [Q11a] to spend (coded as 0) or invest (coded as 1) windfalls. See the caption of Figure 1 for the interpretation of each plot.

credibility. As an alternative approach, we now turn to our proposed sensitivity analysis described earlier that only maintains Assumptions 1 and 3 and does not rely either on additional assumptions such as Assumptions 4 and 5[10] or on the availability of auxiliary information. We apply this analysis to the democratic deliberation experiment by deriving the weakest condition under which the conclusion based on the mismeasured variable remains valid. In particular, we obtain the maximum probability of misclassification that could exist in order for the leader effect to be always positive. As already explained, this can be done by deriving the bounds of the ATE for various values of ρ defined in equation (10).

The results of this sensitivity analysis are shown in Figure 3 . In each panel, the bounds of the ATE of leaders' preferences (the vertical axis) are presented as a function of the maximum magnitude of measurement error that may exist (the horizontal axis). Recall that ρ represents the maximal value of the total misclassification error probability. Thus, when $\rho = 0$, there is no measurement error and the true ATE equals the expected value of its naïve estimate given by the estimated value of τ (the dotted horizontal line). As we move toward the right along the horizontal axis, however, we allow in calculating the bounds the leaders' recorded preferences to differ from their unobserved pre-deliberation preferences. Since differential misclassification can become more frequent

[10] We can also conduct the proposed sensitivity analysis under these two additional assumptions. However, it turns out that the estimated lower bounds, which we will focus on in the analysis below, do not change in this experiment even if we incorporate them.

FIGURE 3 Sensitivity Analyses



Each panel presents the estimated sharp upper and lower bounds on the average treatment effect (the vertical axis) as a function of ρ in equation (10), which represents the maximum probability of misclassification that is assumed to exist (the horizontal axis) for each question item. The dotted horizontal line in each panel represents the naïve estimate of the ATE given in equation (1). The dashed vertical line in the lower half of each panel indicates the maximum probability of misclassification that could exist between Z_i^* and Z_i in order for the conclusion based on the mismeasured variable to be guaranteed valid.

as ρ increases, the bounds are widened toward the right side of each panel. When $\rho = 1$, they coincide with the sharp bounds for Assumptions 1 and 3 reported in Figure 2.

Since our primary interest is in the maximum value of ρ that allows the true ATE to remain positive, we focus on the lower bound for each question and examine where it crosses the zero line. This maximum value of ρ is indicated by the dashed vertical line in the lower half of each panel. For example, in the question about local clinics versus reliable hospitals (Q3, upper left panel), the leaders' recorded preferences must not differ from their true pre-forum preferences more than 14.2% of the time in order to conclude that they influenced the discussion outcomes in the direction they prefer.

These analyses give us another way to judge how plausible the conclusions based on the mismeasured treatment variables are in the presence of differential measurement error. For instance, the maximum probability of misclassification that may be present for Question 4a (upper right panel) turns out to be somewhat high (24.6%) compared to the other questions. This finding suggests that leaders are likely to have influenced discussion outcomes for this particular question when compared with the other questions. In particular, the leaders' influence on outcomes is substantially more likely to be positive for Question 4a than for Question 3, despite the fact that the naïve estimate of the effect is *smaller* for Question 4a (0.35) than Question 3 (0.50). This illustrates that our sensitivity analysis gives an insight beyond the relative magnitude of causal estimate based on no measurement error assumption (though clearly both are closely related). In contrast, for Questions 7b and 11a (bottom panels), the maximum misclassification probability that may exist for the leader influence being positive is as low as 0.2% and 5.6%, respectively. Thus, for these questions, we reject the presence of leaders' influence with much higher certainty than the others. In sum, the proposed sensitivity analysis can be used to formally assess the robustness of the original conclusions.

Concluding Remarks

Previous methodological literature on measurement error has almost exclusively focused on the cases of nondifferential measurement error where the error is assumed to be independent of the outcome. Yet, differential measurement error is common in retrospective studies where measurements are taken after the outcome is realized. Even in prospective studies, differential measurement error may arise if unobserved covariates are correlated with both the outcome and the error.

Unfortunately, causal inference with the differential misclassification of treatment is highly problematic because, as shown in this article, little can be learned without reliance on strong assumptions or auxiliary information. We show that under minimal assumptions the sharp bounds of the ATE are informative but their width is large and they always contain zero. Hence, further progress to narrow the bounds requires additional assumptions that are based on researchers' substantive knowledge. We demonstrate how to formulate such assumptions and derive the sharp bounds under a fully nonparametric, distribution-free setting. We characterize the identification region as a function of the unknown treatment assignment probability. This will allow researchers to utilize auxiliary information about this probability when it is available.

Another methodological contribution of this article is the new sensitivity analysis we propose. Given the serious identification problem caused by differential measurement error, our sensitivity analysis directly investigates a weakest condition under which the conclusions based on the mismeasured treatment variable remain valid. In particular, we offer a method to identify the maximum frequency of misclassification that may exist in order to identify the sign of the ATE. Such an analysis should help researchers evaluate the robustness of their conclusions in the presence of differential measurement error.

Political scientists have long used parametric regression models to analyze their data. The problem of this approach is that these commonly used regression models rely on statistical assumptions that are not necessarily based on researchers' substantive knowledge and are often difficult to verify from the observed data. It is well known that inferences based on such modeling assumptions are necessarily sensitive and likely to yield unreliable conclusions (e.g., Ho et al. 2007). To directly address this issue, the nonparametric identification analysis advocated by Manski (1995, 2007) and others has been widely applied across disciplines. Such an analysis aims to first establish what can be learned from the data alone and then clarify the role each additional assumption plays in identifying the quantities of interest. We contribute to this broad methodological literature by deriving the identification region of the ATE in the presence of differential measurement error and proposing a new sensitivity analysis. We also show here its potential for applications in political science by reexamining an important field experiment. We believe that this kind of analysis allows political scientists to recognize the degree to which the debates in this discipline depend on extraneous assumptions, rather than on data themselves.

Mathematical Appendix

A.1 No Identification Power of Assumption 1

Proposition 3 (Uninformativeness of Assumption 1). *Suppose that Assumption 1 holds. Then,*

1. *The sharp bounds of the average treatment effect are* $[-1, 1]$.
2. *The upper bound equals 1 if and only if* $\Pr(Z_i^* = 1) = \Pr(Y_i = 1)$.
3. *The lower bound equals* -1 *if and only if* $\Pr(Z_i^* = 1) = \Pr(Y_i = 0)$.

Proof: Suppose that the ATE is equal to 1, or equivalently $\Pr(Y_i = 1 \mid Z_i^* = 1) = 1$ and $\Pr(Y_i = 1 \mid Z_i^* = 0) = 0$. Then, by law of total probability, we have $\Pr(Y_i = y, Z_i = z) = \Pr(Z_i^* = y, Z_i = z)$ for $y, z \in \{0, 1\}$, and the assumption implies no other restriction on the joint probability. Thus, from this, $\Pr(Y_i = 1) = \Pr(Z_i^* = 1)$ follows. Conversely, when $\Pr(Y_i = 1) = \Pr(Z_i^* = 1)$, we have $\Pr(Z_i^* = 1) = \Pr(Y_i = 1 \mid Z_i^* = 1)\Pr(Z_i^* = 1) + \Pr(Y_i = 1 \mid Z_i^* = 0)(1 - \Pr(Z_i^* = 1))$. This equation is satisfied when $\Pr(Y_i = 1 \mid Z_i^* = 1) = 1$ and $\Pr(Y_i = 1 \mid Z_i^* = 0) = 0$ or equivalently $\tau^* = 1$. A proof for the lower bound is similar. □

Remark: This result is intuitive. For example, the ATE equals 1 only if all units with $Y_i = 1 (Y_i = 0)$ actually belong to the treatment (control) group, i.e., $Y_i = Z_i^*$. Thus, unless researchers have auxiliary information that $\Pr(Z_i^* = 1) \neq \Pr(Y_i = 1)$ or $\Pr(Z_i^* = 1) \neq \Pr(Y_i = 0)$, Assumption 1 alone has no identifying power. □

A.2 Proof of Proposition 1

By Proposition 3, the sharp lower bound is uninformative if and only if $\Pr(Z_i^* = 1) = \Pr(Y_i = 0)$. Therefore, under this condition, we have, by law of total probability, $\Pr(Y_i = 1, Z_i = 1) = \Pr(Z_i = 1 \mid Z_i^* = 0)\Pr(Y_i = 1)$, where the equality follows from the fact that $\Pr(Y_i = 1 \mid Z_i^* = 1) = 0$, $\Pr(Y_i = 1 \mid Z_i^* = 0) = 1$, and $\Pr(Z_i^* = 0) = \Pr(Y_i = 1)$. By rearrangement, we obtain $\Pr(Z_i = 1 \mid Z_i^* = 0) = \Pr(Y_i = 1, Z_i = 1)/\Pr(Y_i = 1) = \Pr(Z_i = 1 \mid Y_i = 1)$. A similar calculation yields $\Pr(Z_i = 1 \mid Z_i^* = 1) = \Pr(Z_i = 1 \mid Y_i = 0)$. Therefore, Assumption 3 is equivalent to $1 - \Pr(Z_i = 1 \mid Y_i = 0) + \Pr(Z_i = 1 \mid Y_i = 1) < 1$ or $\Pr(Z_i = 1 \mid Y_i = 1) < \Pr(Z_i = 1 \mid Y_i = 0)$. This and Proposition 3 together establish the first part of Proposition 1. The second part can be proved by starting from the condition $\Pr(Z_i^* = 1) = \Pr(Y_i = 1)$ and following similar steps. □

A.3 Proof of Proposition 2

The sharp upper and lower bounds on τ^* under Assumptions 1, 3, and 5 can be obtained as functions of Q by solving the linear programming problem described in the text. Rearranging the constraints and objective functions given by equations (2), (3), (4) and (5), the problem can be expressed in the following simpler form:

$$\text{maximize/minimize} \quad \tau^* = \psi_{11} - \phi_{11} - \frac{P_{10}}{1-Q},$$

$$\text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \ \boldsymbol{x} \geq \boldsymbol{0},$$

$$\text{where} \quad \boldsymbol{A} = \begin{bmatrix} 1-Q & 0 & Q & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{x} = \begin{bmatrix} \phi_{00} \\ \phi_{11} \\ \psi_{00} \\ \psi_{11} \\ \eta \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} P_{00} \\ 1 - \dfrac{P_{01}}{Q} \\ 1 - \dfrac{P_{10}}{1-Q} \\ 1 \end{bmatrix},$$

and η denotes a slack variable such that $\eta \geq 0$. Note that we modified the problem slightly by changing the strict inequality in Assumption 3 to a weak inequality so that the resulting bounds correspond to the supremum and infinimum of the identification region rather than its maximum and minimum. We solve the problem by enumerating all vertexes of the constraint polygon. □

References

Achen, C. H. 1975. “Mass Political Attitudes and the Survey Response.” *American Political Science Review* 69(4): 1218–31.

Aigner, D. J. 1973. “Regression with a Binary Independent Variable Subject to Errors of Observation.” *Journal of Econometrics* 1: 49–60.

Asher, H. B. 1974. “Some Consequences of Measurement Error in Survey Data.” *American Journal of Political Science* 18(2): 469–85.

Ashworth, S., Clinton, J., Meirowitz, A., and Ramsay, K. W. 2008. “Design, Inference, and the Strategic Logic of Suicide Terrorism.” *American Political Science Review* 102(2): 269–73.

Balke, A., and Pearl, J. 1997. “Bounds on Treatment Effects from Studies with Imperfect Compliance.” *Journal of the American Statistical Association* 92: 1171–76.

Bartels, L. M. 1993. “Messages Received: The Political Impact of Media Exposure.” *American Political Science Review* 87(2): 267–85.

Beran, R. 1988. “Balanced Simultaneous Confidence Sets.” *Journal of the American Statistical Association* 83(403): 679–86.

Black, D. A., Berger, M. C., and Scott, F. A. 2000. “Bounding Parameter Estimates with Nonclassical Measurement Error.” *Journal of the American Statistical Association* 95(451): 739–48.

Bollinger, C. R. 1996. “Bounding Mean Regressions When a Binary Regressor Is Mismeasured.” *Journal of Econometrics* 73: 387–99.

Buzas, J., and Stefanski, L. A. 1996. “Instrumental Variable Estimation in Generalized Linear Measurement Error Models.” *Journal of the American Statistical Association* 91(435): 999–1006.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Tosteson, T. D., and Karagas, M. R. 2004. “Nonlinear and Nonparametric Regression and Instrumental Variables.” *Journal of the American Statistical Association* 99(467): 736–50.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. London: Chapman & Hall.

Chattopadhyay, R., and Duflo, E. 2004. “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India.” *Econometrica* 72(5): 1409–43.

Cochran, W. G. 1968. “Errors of Measurement in Statistics.” *Technometrics* 10(4): 637–66.

Druckman, J. N., Green, D. P., Kuklinski, J. H., and Lupia, A. 2006. “The Growth and Development of Experimental

Research in Political Science." *American Political Science Review* 100(4): 627–35.

Dryzek, J. S. 1990. *Discursive Democracy: Politics, Policy, and Political Science*. Cambridge: Cambridge University Press.

Dryzek, J. S. 2000. *Deliberative Democracy and Beyond: Liberals, Critiques, Contestations*. Oxford: Oxford University Press.

Duncan, O. D., and Davis, B. 1953. "An Alternative to Ecological Correlation." *American Sociological Review* 18(6): 665–66.

Frangakis, C. E., and Rubin, D. B. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1): 21–29.

Fuller, W. A. 1987. *Measurement Error Models*. New York: John Wiley & Sons.

Galston, W. A. 2001. "Political Knowledge, Political Engagement, and Civic Education." *Annual Review of Political Science* 4: 217–34.

Habermas, J. 1985. *Theory of Communicative Action, Vol. 1, Reason and the Rationalization of Society*. Boston: Beacon Press.

Hanmer, M. J. 2007. "An Alternative Approach to Estimating Who Is Most Likely to Respond to Changes in Registration Laws." *Political Behavior* 29(1): 1–30.

Hausman, J. A., Newey, W. K., Ichimura, H., and Powell, J. L. 1991. "Identification and Estimation of Polynomial Errors-in-Variables Models." *Journal of Econometrics* 50(3): 273–95.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Holland, P. W. 1986. "Statistics and Causal Inference (with Discussion)." *Journal of the American Statistical Association* 81: 945–60.

Horiuchi, Y., Imai, K., and Taniguchi, N. 2007. "Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment." *American Journal of Political Science* 51(3): 669–87.

Huber, G. A., and Lapinski, J. S. 2006. "The 'Race Card' Revisited: Assessing Racial Priming in Policy Contests." *American Journal of Political Science* 50(2): 421–40.

Humphreys, M., Masters, W. A., and Sandbu, M. E. 2006. "The Role of Leaders in Democratic Deliberations: Results from a Field Experiment in São Tomé and Príncipe." *World Politics* 58(4): 583–622.

Imai, K. 2008. "Sharp Bounds on the Causal Effects in Randomized Experiments with Truncation-by-Death." *Statistics & Probability Letters* 78(2): 144–49.

Imai, K., and Soneji, S. 2007. "On the Estimation of Disability-Free Life Expectancy: Sullivan's Method and Its Extension." *Journal of the American Statistical Association* 102(480): 1199–1211.

Imai, K., and Yamamoto, T. 2010. "Replication Data for Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis." hdl:1902.1/14057.

Klepper, S. 1988. "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables." *Journal of Econometrics* 37: 343–59.

Lewbel, A. 2007. "Estimation of Average Treatment Effects with Misclassification." *Econometrica* 75(2): 537–51.

Mahajan, A. 2006. "Identification and Estimation of Regression Models with Misclassification." *Econometrica* 74(3): 631–65.

Manski, C. F. 1990. "Non-parametric Bounds on Treatment Effects." *American Economic Review, Papers and Proceedings* 80: 319–23.

Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Mendelberg, T. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton, NJ: Princeton University Press.

Mendelberg, T. 2002. "The Deliberative Citizen: Theory and Evidence." *Political Decision Making, Deliberation and Participation* 6: 151–93.

Mondak, J. J. 1999. "Reconsidering the Measurement of Political Knowledge." *Political Analysis* 8(1): 57–82.

Prior, M. 2009. "Improving Media Effects Research Through Better Measurement of News Exposure." *Journal of Politics* 71(3): 893–908.

Quinn, K. M. 2008. "What Can Be Learned from a Simple Table: Bayesian Inference and Sensitivity Analysis for Causal Effects from 2 × 2 and 2 × 2 × k Tables in the Presence of Unmeasured Confounding." Unpublished manuscript, Harvard University.

Rosenbaum, P. R. 2002. *Observational Studies*. 2nd ed. New York: Springer-Verlag.

Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Sanders, L. M. 1997. "Against Deliberation." *Political Theory* 25(3): 347–76.

Valentino, N. A., Hutchings, V. L., and White, I. K. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96(1): 75–90.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Young, I. M. 2002. *Inclusion and Democracy*. Oxford: Oxford University Press.

Zaller, J., and Feldman, S. 1992. "A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences." *American Journal of Political Science* 36(3): 579–616.