Combining double sampling and bounds to address non-ignorable missing outcomes in randomized experiments

Alexander Coppock, Alan S. Gerber, Donald P. Green, and Holger L. Kern*

September 10, 2016

*Alexander Coppock is Assistant Professor of Political Science, Yale University. Alan S. Gerber is Charles C. & Dorathea S. Dilley Professor of Political Science, Yale University. Donald P. Green is Professor of Political Science, Columbia University. Holger L. Kern is Assistant Professor of Political Science, Florida State University. The authors thank Sebastian Bauhoff, Bill Berry, Chris Blattman, Jake Bowers, Matias Cattaneo, Kosuke Imai, Molly Offer-Westort, Rocio Titiunik, and participants of the 2013 Joint Statistical Meetings for very helpful comments on previous versions of this manuscript. We also thank Peter Aronow for his contributions to previous versions of this paper. This research was approved by the Columbia University IRB (Protocol AAAP1312) and the empirical analyses were preregistered at egap.org (ID: 20150702AA). Easy-to-use software for the statistical programming language R that implements the methods described in this paper is available at github.com/acoppock/attrition. The replication materials for all analyses reported here are available at http://dx.doi.org/10.7910/DVN/AQB4MP.

Abstract

Missing outcome data plague many randomized experiments. Common solutions rely on ignorability assumptions that may not be credible in all applications. We propose a method for confronting missing outcome data that makes fairly weak assumptions but can still yield informative bounds on the average treatment effect. Our approach is based on a combination of the double sampling design and non-parametric worst-case bounds. We derive a worst-case bounds estimator under double sampling and provide analytic expressions for variance estimators and confidence intervals. We also propose a method for covariate adjustment using post-stratification and a sensitivity analysis for non-ignorable missingness. Finally, we illustrate the utility of our approach using Monte Carlo simulations and a placebo-controlled randomized field experiment on the effects of persuasion on social attitudes with survey-based outcome measures.

1 Introduction

Over the last two decades, increased attention to causal identification has led to rapid growth in the number of randomized experiments conducted in political science, economics, and sociology. Although randomization in principle allows for the straightforward estimation of average treatment effects through a comparison of mean outcomes in the treatment and control groups, experimenters frequently encounter situations in which outcomes are missing for some subjects. Reasons for such missing outcome data include non-response in post-treatment surveys (e.g., Peress 2010; Frankel and Hillygus 2014; Si, Reiter, and Hillygus 2014), gaps in administrative records (e.g., Grogger 2012), and residential mobility and migration (e.g., Baird, Hamory, and Miguel 2008).[1] In contrast to missing covariate information, missing outcome data threaten the validity of experimental estimates because they in effect undo the random assignment to treatment and control groups. Treatment effect estimators that ignore this missingness and consider only subjects with observed outcomes will generally be biased and inconsistent.

Researchers often address the problem of missing outcome data by introducing additional assumptions. Selection models (e.g., Heckman 1979; Das, Newey, and Vella 2003; see also Freedman and Sekhon 2010) require either strong parametric assumptions or an instrumental variable that is related to missingness but has no direct effect on the outcome. Such assumptions are often hard to defend on empirical or theoretical grounds. Imputation (e.g., Little and Rubin 2002) and reweighting methods (e.g., Cassel, Särndal, and Wretman 1983) typically presuppose that missingness is “ignorable” (see Kang and Schafer 2007 for a review of these methods, including “doubly robust” estimators that combine imputation and reweighting). Ignorability implies that missingness does not depend on the unobserved values of the outcome variable itself after taking into account observables. However, researchers conduct randomized experiments with the goal of drawing credible causal inferences under a minimal set of assumptions and may be reluctant to impose strong assumptions about ignorable missingness.[2]

We propose a different approach to missing outcome data in randomized experiments.[3] In an effort to impose fairly weak assumptions, our approach is based on the combination of the double (or two-phase) sampling design and non-parametric worst-case bounds.[4] The double sampling design, originally due to Neyman (1938), was proposed as a solution to non-response in survey research by Hansen and Hurwitz (1946). Hansen and Hurwitz (1946) considered cases in

[1] In other cases, outcomes may not even be well-defined for some subjects, an example being wages for the non-working in randomized evaluations of active labor market programs (Imai 2008; McConnell, Stuart, and Devaney 2008; Lee 2010).

[2] There exist partial identification methods for situations in which missingness depends on the unobserved values of the outcome variable itself even after taking into account the information in the fully observed variables (see, for example, Imai 2009). Such techniques for dealing with non-ignorable missingness impose other identification assumptions that may or may not hold in any given application. Typically, these methods are tailored towards specific applications with specific reasons for non-ignorable missingness.

[3] The applicability of the framework developed here is broad. It covers not only survey and field experiments as well as laboratory experiments with temporally delayed outcome measurement but also naturally-occurring randomizations such a school and visa lotteries.

[4] Worst-case bounds impose weaker assumptions than bounds that presuppose a missing confounder or unknown elements of the sampling design; see, e.g., Robins, Rotnitzky and Scharfstein (1999) or Aronow and Lee (2013).

which an inexpensive initial survey encounters non-response and is then followed up with an intensive and more costly effort to interview everyone in a randomly selected subsample of initial non-respondents. Given the probability of non-response in the initial survey, the variances of the outcomes among initial respondents and non-respondents, and the costs of initial and follow-up data collection together with a budget constraint, Hansen and Hurwitz (1946) derived the optimal number of subjects in the initial and follow-up samples. Subsequent work has refined and extended Hansen and Hurwitz's (1946) approach but maintained their assumption of no missingness among follow-up subjects (Ericson 1967; Kaufman and King 1973; see also Lohr 2010 for a textbook discussion).

We relax this restrictive assumption to allow for missingness in both the initial sample and the follow-up sample. However, under our approach, the average treatment effect is no longer point identified. Nonetheless, as we will show below, fundamental uncertainty about the average treatment effect can be reduced by calculating worst-case bounds for missing outcomes in the follow-up sample. Such worst-case bounds have a long history in statistics, going back at least to Birnbaum and Sirken (1950a, 1950b) and Cochran, Mosteller, and Tukey (1954). Manski (1990, 1995) has popularized their use in economics. Recent econometric work on partial identification, that is, models where only bounds on parameters of interest are identified, includes Horowitz and Manski (1998, 2000), Manski (2007), and Lee (2010). Tamer (2010) provides a review of this literature. Mebane and Poast (2013) and Keele and Minozzi (2013) are recent contributions to the literature in political science.

Researchers have rarely drawn the connection between double sampling and worst-case bounds. In his classic textbook on survey sampling, Cochran (1977) discussed both double sampling designs and worst-case bounds as techniques for dealing with survey non-response but did not mention the possibility of combining these two approaches. We are aware of only two works that explicitly discuss the use of double sampling in combination with bounds. The first work is DiNardo, McCrary, and Sanbonmatsu's (2006) unpublished analysis of the *Moving to Opportunity* field experiment, which does not consider estimation, inference, or design implications for double sampling with worst-case bounds. The second work is the textbook by Gerber and Green (2012, ch. 7), which presents an intuitive discussion of how double sampling and worst-case bounds may be combined to produce a narrower identification region. Gerber and Green (2012) presents a description of this design and illustrates with empirical examples how the estimator may substantially improve upon worst-case bounds. However, Gerber and Green (2012) does not present a formal derivation of the estimator. It also does not provide analytic expressions for the sampling variance of the estimator or confidence intervals and does not discuss methods for covariate adjustment in this setting.

In the next section, we show how double sampling combined with worst-case bounds partially identifies the average treatment effect from experimental data. The double sampling approach shrinks the width of the worst-case bounds, reducing fundamental uncertainty about the average treatment effect. Our paper thus contributes to the literature on partial identification. However, in contrast to much of this literature, which introduces additional assumptions to tighten worst-case bounds at the analysis stage, our approach modifies the data collection stage in order to shrink worst-case bounds on average treatment effects. We present an estimator for bounds on average treatment effects, variance estimators, confidence intervals, and an extension that allows for covariate adjustment via post-stratification. We also propose a sensitivity analysis. We then explore the performance of our approach with a small simulation study before applying our method to a replication of a survey experiment (Levendusky and Malhotra 2016). These examples illustrate how double sampling combined with worst-case bounds can reduce the uncertainty associated with the analysis of experiments with potentially non-ignorable missing outcomes under fairly weak assumptions.

The approach proposed here applies directly to experiments in which the collection of subjects does not represent a random sample from a well-defined population, so that the sample average treatment effect is the causal estimand of interest (see Imai, King, and Stuart (2008) for a discussion of causal estimands). This would be the case, for example, in experiments conducted in convenience samples or experimental samples that are unrepresentative of the target population because of self-selection into the sample. We should note, however, that our proposed methodology is not designed to solve the problems involved in generalizing inferences from such samples to a target population.[5]

2 Partial identification and estimation

In this section, we show how double sampling combined with worst-case bounds partially identifies the average treatment effect (ATE).

2.1 Notation

Consider a completely randomized experiment with a binary treatment conducted in a sample of subjects possibly drawn at random from a large population.[6] Let D_i denote subject i's treatment status, so that D_i is equal to 1 when subject i is assigned to treatment and equal to 0 when subject i is assigned to control. We assume that compliance with treatment assignment is unproblematic, so that the assigned treatment is the same as the received treatment for all subjects (see Gerber, Green, Kaplan, and Kern 2010 for a discussion of non-compliance). Each subject has two potential outcomes, denoted $Y_i(d)$ for $d \in \{0,1\}$, where $Y_i(0)$ indicates the outcome under control and $Y_i(1)$ the outcome under treatment.[7] We observe at most one of these two potential outcomes, depending on whether subject i receives the treatment or the control, so that the outcome $Y_i = Y_i(1)D_i + Y_i(0)(1-D_i)$. When there is no attrition, the outcome Y_i is observed for all subjects. From random assignment of D_i and the definition of Y_i, it follows that the ATE equals $\mathrm{E}[Y_i(1) - Y_i(0)] = \mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0]$. We can unbiasedly estimate the ATE by the sample analogue of the second expression, the difference in mean outcomes between the treatment and control groups. As noted above, depending on whether the set of experimental subjects can be considered a random sample

[5] For more on this, see Kern, Stuart, Hill, and Green 2016 and Hartman, Grieve, Ramsahai, and Sekhon Forthcoming.

[6] Our approach immediately extends to experiments with multiple treatments.

[7] Throughout, we make the stable unit treatment value assumption (SUTVA), which states that there is no interference between subjects and no unobserved variation in the nature of the treatments (Imbens and Rubin 2015: 9–12).

from a well-defined population, the ATE corresponds to the sample average treatment effect or the population average treatment effect.

Some additional notation allows us to discuss the consequences of missing outcomes for the identification of the ATE. Let $R_i(d)$ denote whether subject i's outcome is observed when subject i is assigned to treatment d, with $R_i(d)$ equal to 1 [0] if subject i's outcome is observed [not observed] given treatment assignment d. For example, $\{R_i(1) = 1, R_i(0) = 0\}$ indicates that subject i's outcome would be observed if subject i were assigned to treatment but would not be observed if subject i were assigned to control. Since $R_i(d)$ depends on treatment assignment, we observe either $R_i(0)$ or $R_i(1)$ but never both. The observed value of $R_i(d)$ can be written as $R_i = R_i(1)(D_i) + R_i(0)(1 - D_i)$. Y_i is only observed when $R_i(d) = 1$; thus, the marginal distributions of $Y_i(1)$ and $Y_i(0)$ can only be observed conditional on $R_i(1) = 1$ and $R_i(0) = 1$, respectively.

Attrition may lead to bias in a simple comparison of treatment and control means when missingness is related to potential outcomes $Y_i(0)$ and $Y_i(1)$. To see this, note that the expected outcome in the treatment group can be written as a weighted average, as in

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1] &= \mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + \\ &\quad (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 0]. \end{aligned} \tag{1}$$

Likewise, the expected outcome in the control group can be written as

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 0] &= \mathrm{E}[R_i \mid D_i = 0] \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 1] + \\ &\quad (1 - \mathrm{E}[R_i \mid D_i = 0]) \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 0]. \end{aligned} \tag{2}$$

These two expressions are not point identified since the data are not informative about $\mathrm{E}[Y_i \mid D_i = 1, R_i = 0]$ and $\mathrm{E}[Y_i \mid D_i = 0, R_i = 0]$. However, suppose $Y_i(D_i = d) \perp\!\!\!\perp R_i(D_i = d)$ for $d \in \{0, 1\}$, so that missingness is independent of potential outcomes. (As usual, $\perp\!\!\!\perp$ denotes statistical independence.) Then the equality

$$\mathrm{E}[Y_i \mid D_i = d] = \mathrm{E}[Y_i \mid D_i = d, R_i = 1] = \mathrm{E}[Y_i \mid D_i = d, R_i = 0] = \mathrm{E}[Y_i(d)]$$

holds, in which case the ATE is point identified and can be estimated by comparing the means of the observed outcomes in the treatment and control groups.[8]

2.2 Worst-case bounds

Suppose we are unwilling to assume that missingness is independent of potential outcomes (even conditional on covariates). A fallback approach is to put worst-case bounds on the ATE. These bounds are sharp, i.e., they are the narrowest bounds possible unless additional assumptions are introduced (Manski 2007). Assume that the support of the outcome variable is bounded, with

[8] This ignorability assumption can be further relaxed by conditioning on covariates, as in e.g., Assumption 1 in Imai (2009). Moreover, in rare cases biases in the estimation of treatment and control group means could be largely offsetting, leading to approximately unbiased estimation of the ATE (we thank an anonymous reviewer for stressing this point.) Generally, however, some form of ignorability assumption is necessary to point identify the ATE when outcomes are missing.

y^L denoting the lower bound and y^U denoting the upper bound. It is then possible to bound the expected outcome in the treated group, $\mathrm{E}[Y_i \mid D_i = 1]$, by assuming that the missing outcomes take the smallest possible value y^L or the largest possible value y^U:

$$\begin{aligned}\mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot y^L \\ \leq \mathrm{E}[Y_i \mid D_i = 1] = \mathrm{E}[Y_i(1)] \leq \\ \mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot y^U. \end{aligned} \tag{3}$$

Bounds for the expected outcome in the control group are formed the same way, so that

$$\begin{aligned}\mathrm{E}[R_i \mid D_i = 0] \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 0]) \cdot y^L \\ \leq \mathrm{E}[Y_i \mid D_i = 0] = \mathrm{E}[Y_i(0)] \leq \\ \mathrm{E}[R_i \mid D_i = 0] \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 0]) \cdot y^U. \end{aligned} \tag{4}$$

We then form the lower bound on the ATE by subtracting the upper bound on the expected outcome under control from the lower bound on the expected outcome under treatment. The upper bound on the ATE is formed by subtracting the lower bound on the expected outcome under control from the upper bound on the expected outcome under treatment:

$$\begin{aligned}\mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot y^L - \\ \mathrm{E}[R_i \mid D_i = 0] \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 0]) \cdot y^U \\ \leq \mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0] = \mathrm{E}[Y_i(1) - Y_i(0)] \leq \\ \mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot y^U - \\ \mathrm{E}[R_i \mid D_i = 0] \cdot \mathrm{E}[Y_i \mid D_i = 0, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 0]) \cdot y^L. \end{aligned} \tag{5}$$

It is easy to show that the bounds on the expected outcomes in the control and treatment groups have widths equal to $y^U - y^L$ times the fractions of missing outcomes. The width of the bound on the ATE is equal to the sum of the widths of these two bounds. When the fraction of missing outcomes is appreciable, worst-case bounds on the ATE thus tend to be wide and not very informative. However, as we will show in the next section of the paper, we can often dramatically reduce the width of the identification region by combining worst-case bounds with the double sampling design.

2.3 Double sampling

The basic insight behind the double sampling design is that we can eliminate or at least reduce the adverse consequences of missing outcome data by directing an intensive follow-up effort at a randomly selected subsample of subjects with initially missing outcomes.[9] Let F_i denote whether subject i is randomly assigned to the follow-up sample, with F_i equal to 1 [0] if subject i is assigned [not assigned] to the follow-up sample. Because only subjects with initially missing outcomes are assigned to the follow-up sample, $\Pr(F_i = 1 \mid R_i = 1) = 0$ for all i.

[9] See An et al. 2009, Jenkins et al. 2009, and Fraser and Yan 2007 for applications of double sampling in a survey context that are able to achieve relatively high response rates in follow-up samples.

The double sampling design rests on the assumption that potential outcomes are invariant across data collection rounds, so that it does not matter whether we observe outcomes in the initial sample or the follow-up sample. We call this assumption *outcome stability.* Importantly, note that this assumption allows for arbitrary differences between observations with initially missing and initially observed outcomes. We do not have to assume that, perhaps conditional on covariates, outcomes for the latter group can be used to impute outcomes for the former group.

Outcome stability can be violated in different ways. For example, if the mode of data collection in the initial sample is different from the mode of data collection in the follow-up sample, outcome stability does not hold if the mode of data collection itself affects outcomes. In Hansen and Hurwitz (1946), the implicit assumption is that how subjects with initially missing outcomes respond when interviewed in the follow-up round (face-to-face) is identical to how the same subjects would have responded when surveyed in the initial round (by mail). Outcome stability is also violated if the follow-up data collection offers monetary incentives to increase responsiveness and these incentives affect subjects' responses. Finally, outcome stability fails to hold if treatment effects decay (or increase) over time and there exists an appreciable time gap between the initial outcome data collection and the follow-up data collection.

Researchers using a double sampling design to address missing outcome data in randomized experiments should explicitly discuss the plausibility of outcome stability in light of their specific application, in particular with respect to differences between data collection procedures in the initial sample and the follow-up sample and the time gap between initial and follow-up data collection.

In order to see the intuition behind double sampling under outcome stability, consider the following case. Assume for the moment, following Hansen and Hurwitz (1946), that follow-up data collection is entirely successful, so that we obtain outcomes for every subject in the follow-up sample. Also assume that the follow-up sample contains both treated and control subjects, which is easily assured through random sampling within strata defined by treatment assignment. Since the follow-up sample is a stratified random sample from the subgroups of control and treated subjects with initially missing outcomes, $\{i : D_i = 0, R_i = 0\}$ and $\{i : D_i = 1, R_i = 0\}$, it is clear that the outcomes observed in the follow-up sample are sufficient to identify $\mathrm{E}[Y_i \mid D_i = 0, R_i = 0]$ and $\mathrm{E}[Y_i \mid D_i = 1, R_i = 0]$ when outcome stability holds. That is, when outcomes in the follow-up sample are fully observed and outcome stability holds, all terms in (1) and (2) are point identified either from the initial sample or the follow-up sample, so that the sample analogues of (1) and (2) can be used to estimate the ATE.

2.4 Double sampling and bounds combined

To deal with situations in which we fail to observe outcomes for every subject in the follow-up sample, let $S_i(d)$ denote whether subject i's outcome is observed in the follow-up sample when subject i is assigned to treatment d, with $S_i(d)$ equal to 1 [0] if subject i's outcome is observed [not observed] in the follow-up sample given treatment assignment d. For example, $\{S_i(1) = 1, S_i(0) = 0\}$ indicates that subject i's outcome would be observed in the follow-up sample if subject i were assigned to treatment but would not be observed if subject i were assigned to control. Of course,

$S_i = S_i(1)D_i + S_i(0)(1 - D_i)$ is only observed for subjects randomly assigned to the follow-up sample, that is, subjects for which $F_i = 1$.

Recall from (1) that we can write the expected outcome in the treatment group as a weighted average and that the sampling process identifies all expectations except $\mathrm{E}[Y_i \mid D_i = 1, R_i = 0]$. If outcome stability holds, $\mathrm{E}[Y_i \mid D_i = 1, R_i = 0] = \mathrm{E}[Y_i \mid D_i = 1, R_i = 0, F_i = 1] = \mathrm{E}[Y_i \mid D_i = 1, F_i = 1]$, where the first equality follows from stratified random sampling, which implies $Y_i \perp\!\!\!\perp F_i \mid D_i = 1, R_i = 0$, and the second equality follows from the definition of F_i. We can write

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1, F_i = 1] &= \mathrm{E}[S_i \mid D_i = 1, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 1] + \\ &\quad (1 - \mathrm{E}[S_i \mid D_i = 1, F_i = 1]) \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 0]. \qquad (6) \end{aligned}$$

All terms in expression (6) are point identified with the exception of the expected outcome in the follow-up treatment group among subjects for which follow-up outcomes are missing, $\mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 0]$. Again, we can bound this term by substituting in y^L and y^U to get the lower and upper bounds, respectively: The lower bound equals

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1, F_i = 1]^L &= \mathrm{E}[S_i \mid D_i = 1, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 1] + \\ &\quad (1 - \mathrm{E}[S_i \mid D_i = 1, F_i = 1]) \cdot y^L. \qquad (7) \end{aligned}$$

The upper bound equals

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1, F_i = 1]^U &= \mathrm{E}[S_i \mid D_i = 1, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 1] + \\ &\quad (1 - \mathrm{E}[S_i \mid D_i = 1, F_i = 1]) \cdot y^U. \qquad (8) \end{aligned}$$

Bounds on $\mathrm{E}[Y_i \mid D_i = 0, F_i = 1]$ are constructed the same way, yielding

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 0, F_i = 1]^L &= \mathrm{E}[S_i \mid D_i = 0, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 0, F_i = 1, S_i = 1] + \\ &\quad (1 - \mathrm{E}[S_i \mid D_i = 0, F_i = 1]) \cdot y^L \qquad (9) \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 0, F_i = 1]^U &= \mathrm{E}[S_i \mid D_i = 0, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 0, F_i = 1, S_i = 1] + \\ &\quad (1 - \mathrm{E}[S_i \mid D_i = 0, F_i = 1]) \cdot y^U. \qquad (10) \end{aligned}$$

Substituting these four bounds into (1) and (2) then yields bounds on the expected outcomes in the treatment and control groups based on the double sampling design. The lower bound for the expected outcome in the treatment group equals

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1]^L &= \mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1]^L \quad (11) \\ &= \mathrm{E}[R_i \mid D_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, R_i = 1] + (1 - \mathrm{E}[R_i \mid D_i = 1]) \cdot \\ &\quad \Big(\mathrm{E}[S_i \mid D_i = 1, F_i = 1] \cdot \mathrm{E}[Y_i \mid D_i = 1, F_i = 1, S_i = 1] + (1 - \mathrm{E}[S_i \mid D_i = 1, F_i = 1]) \cdot y^L\Big). \end{aligned}$$

The upper bound equals

$$\begin{aligned}
\mathrm{E}\left[Y_i \mid D_i = 1\right]^U &= \mathrm{E}\left[R_i \mid D_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 1\right]\right) \cdot \mathrm{E}\left[Y_i \mid D_i = 1, F_i = 1\right]^U \qquad (12)\\
&= \mathrm{E}\left[R_i \mid D_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 1\right]\right) \cdot \\
&\quad \left(\mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right]\right) \cdot y^U\right).
\end{aligned}$$

Analogous expressions hold for the control group, yielding

$$\begin{aligned}
\mathrm{E}\left[Y_i \mid D_i = 0\right]^L &= \mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1\right]^L \qquad (13)\\
&= \mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \\
&\quad \left(\mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right]\right) \cdot y^L\right).
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[Y_i \mid D_i = 0\right]^U &= \mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1\right]^U \qquad (14)\\
&= \mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \\
&\quad \left(\mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right]\right) \cdot y^U\right)
\end{aligned}$$

The bounds on the ATE are formed as in (5). To compute the lower bound on the ATE, we subtract the upper bound on the expected outcome under control (eq. (14)) from the lower bound on the expected outcome under treatment (eq. (11)). To compute the upper bound on the ATE, we subtract the lower bound on the expected outcome under control (eq. (13)) from the upper bound on the expected outcome under treatment (eq. (12)). Formally, the bounds on the ATE are given by

$$\begin{aligned}
&\mathrm{E}\left[R_i \mid D_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 1\right]\right) \cdot \\
&\left(\mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right]\right) \cdot y^L\right) - \\
&\mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \\
&\left(\mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right]\right) \cdot y^U\right) \\[1em]
&\leq \mathrm{E}\left[Y_i \mid D_i = 1\right] - \mathrm{E}\left[Y_i \mid D_i = 0\right] = \mathrm{E}\left[Y_i(1) - Y_i(0)\right] \leq \\[1em]
&\mathrm{E}\left[R_i \mid D_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 1\right]\right) \cdot \\
&\left(\mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 1, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 1, F_i = 1\right]\right) \cdot y^U\right) - \\
&\mathrm{E}\left[R_i \mid D_i = 0\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, R_i = 1\right] + \left(1 - \mathrm{E}\left[R_i \mid D_i = 0\right]\right) \cdot \\
&\left(\mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right] \cdot \mathrm{E}\left[Y_i \mid D_i = 0, F_i = 1, S_i = 1\right] + \left(1 - \mathrm{E}\left[S_i \mid D_i = 0, F_i = 1\right]\right) \cdot y^L\right). \qquad (15)
\end{aligned}$$

Let τ denote the ATE, let τ^L denote the lower bound on τ, and let τ^U denote the upper bound on τ (all given in (15)), so that

$$\tau^L \leq \tau \leq \tau^U. \tag{16}$$

The identification region for the ATE is denoted by $[\tau^L, \tau^U]$. If outcome stability holds, the ATE must lie in this interval, although it cannot be pinpointed any further unless we are willing to introduce additional assumptions. See Manski (1990, 1995, 2007), Zhang and Rubin 2004, and Lee (2010) for a discussion of the use of additional assumptions to further narrow worst-case bounds on quantities of interest and Manski and Nagin (1998), Manski and Pepper (2011), Mebane and Poast (2013), and Keele and Minozzi (2013) for empirical examples.

2.5 Estimation and inference

Let $n_{(c)1}$ be the number of control subjects for whom we attempt to measure outcomes initially and let $p_{(c)1}$ be the proportion of control subjects for which we observe outcomes in the initial sample. From among the $(1-p_{(c)1})n_{(c)1}$ control subjects for which we initially fail to observe outcomes, a random follow-up sample of size $n_{(c)2}$ is taken, with $2 \leq n_{(c)2} \leq (1-p_{(c)1})n_{(c)1}$. Let $p_{(c)2}$ be the proportion of control subjects for which we observe outcomes in the follow-up sample. Let $s^2_{(c)1}$ denote the sample variance of the outcome variable among control subjects with observed outcomes in the initial sample. Let $s^2_{(c)2L}$ denote the sample variance of the outcome variable among control subjects in the follow-up sample with missing outcomes substituted by y^L and $s^2_{(c)2U}$ the sample variance of the outcome variable among control subjects in the follow-up sample with missing outcomes substituted by y^U. Finally, let $\bar{y}_{(c)1}$ be the sample mean among control subjects with observed outcomes in the initial sample, let $\bar{y}_{(c)2}$ be the sample mean among control subjects with observed outcomes in the follow-up sample, and let $\bar{y}_{(c)2L}$ and $\bar{y}_{(c)2U}$ denote the lower and upper bounds on the sample mean among control subjects in the follow-up sample after substituting y^L and y^U for missing outcomes. For the treatment group, $n_{(t)1}$, $p_{(t)1}$, $n_{(t)2}$, $p_{(t)2}$, $s^2_{(t)1}$, $s^2_{(t)2L}$, $s^2_{(t)2U}$, $\bar{y}_{(t)1}$, $\bar{y}_{(t)2}$, $\bar{y}_{(t)2L}$, and $\bar{y}_{(t)2U}$ are defined analogously.

We can use the sample analogues of the expressions in (15) to estimate lower and upper bounds on the ATE. Call these estimated lower and upper bounds $\widehat{\tau}^L$ and $\widehat{\tau}^U$, respectively. From (15), we have

$$\begin{aligned} \widehat{\tau}^L \;=\; & p_{(t)1} \cdot \bar{y}_{(t)1} + (1-p_{(t)1}) \cdot \Big(p_{(t)2} \cdot \bar{y}_{(t)2} + (1-p_{(t)2}) \cdot y^L\Big) - \\ & p_{(c)1} \cdot \bar{y}_{(c)1} + (1-p_{(c)1}) \cdot \Big(p_{(c)2} \cdot \bar{y}_{(c)2} + (1-p_{(c)2}) \cdot y^U\Big) \end{aligned} \tag{17}$$

and

$$\begin{aligned} \widehat{\tau}^U \;=\; & p_{(t)1} \cdot \bar{y}_{(t)1} + (1-p_{(t)1}) \cdot \Big(p_{(t)2} \cdot \bar{y}_{(t)2} + (1-p_{(t)2}) \cdot y^U\Big) - \\ & p_{(c)1} \cdot \bar{y}_{(c)1} + (1-p_{(c)1}) \cdot \Big(p_{(c)2} \cdot \bar{y}_{(c)2} + (1-p_{(c)2}) \cdot y^L\Big). \end{aligned} \tag{18}$$

The asymptotic variance of the double sampling estimator is given in Glynn, Laird, and Rubin (1993, equation 2) (see also Rao 1973 and Cochran 1977). Their assumptions are identical to ours; we do omit a finite-population correction factor. Applying this formula yields expressions for the variances of the lower and upper bounds on the mean outcomes in the control and treatment groups:

$$\widehat{\sigma}^2_{(c)_L} = \frac{p_{(c)1} s^2_{(c)1}}{n_{(c)1}} + \frac{(1-p_{(c)1})^2 s^2_{(c)2L}}{n_{(c)2}} + \frac{(1-p_{(c)1}) p_{(c)1} (\bar{y}_{(c)_{2L}} - \bar{y}_{(c)1})^2}{n_{(c)1}}, \tag{19}$$

$$\widehat{\sigma}^2_{(c)_U} = \frac{p_{(c)1} s^2_{(c)1}}{n_{(c)1}} + \frac{(1-p_{(c)1})^2 s^2_{(c)2U}}{n_{(c)2}} + \frac{(1-p_{(c)1}) p_{(c)1} (\bar{y}_{(c)_{2U}} - \bar{y}_{(c)1})^2}{n_{(c)1}}, \tag{20}$$

$$\widehat{\sigma}^2_{(t)_L} = \frac{p_{(t)1} s^2_{(t)1}}{n_{(t)1}} + \frac{(1-p_{(t)1})^2 s^2_{(t)2L}}{n_{(t)2}} + \frac{(1-p_{(t)1}) p_{(t)1} (\bar{y}_{(t)_{2L}} - \bar{y}_{(t)1})^2}{n_{(t)1}}, \tag{21}$$

$$\widehat{\sigma}^2_{(t)_U} = \frac{p_{(t)1} s^2_{(t)1}}{n_{(t)1}} + \frac{(1-p_{(t)1})^2 s^2_{(t)2U}}{n_{(t)2}} + \frac{(1-p_{(t)1}) p_{(t)1} (\bar{y}_{(t)_{2U}} - \bar{y}_{(t)1})^2}{n_{(t)1}}. \tag{22}$$

Using Neyman's estimator for the variance of the treatment effect in completely randomized experiments, except that we estimate group variances based on the double sampling estimators given in (19)–(22), the standard error of the lower bound on the ATE is equal to $\sqrt{\widehat{\sigma}^2_{(t)_L} + \widehat{\sigma}^2_{(c)_U}}$ and the standard error of the upper bound is equal to $\sqrt{\widehat{\sigma}^2_{(t)_U} + \widehat{\sigma}^2_{(c)_L}}$. If our inferential target is the sample average treatment effect these standard errors are conservative (biased upwards), at least in large samples, unless treatment effects are constant. If our inferential target is the population average treatment effect the standard errors are unbiased in large samples (Freedman, Pisani, and Purves 2007: ch. 27; Imbens and Rubin 2015: ch. 6).

We consider a confidence interval with asymptotically proper coverage for the ATE. Following Imbens and Manski (2004) and Lee (2010) (see also Stoye 2009), a confidence interval that will asymptotically contain the ATE with at least $1-\alpha$ probability is given by

$$\mathrm{CI}_\alpha = \left[\widehat{\tau}^L - c_\alpha \sqrt{\widehat{\sigma}^2_{(t)_L} + \widehat{\sigma}^2_{(c)_U}},\ \ \widehat{\tau}^U + c_\alpha \sqrt{\widehat{\sigma}^2_{(t)_U} + \widehat{\sigma}^2_{(c)_L}} \right], \tag{23}$$

where c_α solves

$$\Phi\left(c_\alpha + \left(\widehat{\tau}^U - \widehat{\tau}^L\right) / \max\left\{ \sqrt{\widehat{\sigma}^2_{(t)_L} + \widehat{\sigma}^2_{(c)_U}},\ \sqrt{\widehat{\sigma}^2_{(t)_U} + \widehat{\sigma}^2_{(c)_L}} \right\} \right) - \Phi(-c_\alpha) = 1 - \alpha, \tag{24}$$

with Φ denoting the standard normal distribution function. Numerically, the value of c_α that solves (24) can easily be found using a line search.

2.6 Covariate adjustment with post-stratification

When discrete covariates prognostic of outcomes are available, this information can be used to improve precision.[10] Consider the case of a categorical variable B_i with support $\{1,\dots,K\}$ and note the following decomposition:

$$\mathrm{E}[Y_i \mid D_i = d] = \sum_{k=1}^{K} \Pr(B_i = k \mid D_i = d) \cdot \mathrm{E}[Y_i \mid D_i = d, B_i = k] \tag{25}$$

for $d \in \{0,1\}$ and $k \in \{1,\dots,K\}$. When $B_i \perp\!\!\!\perp D_i$, as is ensured by random assignment of D_i when B_i is a covariate, $\Pr(B_i = k \mid D_i = d) = \Pr(B_i = k)$. We therefore have

$$\mathrm{E}[Y_i \mid D_i = d] = \sum_{k=1}^{K} \Pr(B_i = k) \cdot \mathrm{E}[Y_i \mid D_i = d, B_i = k]. \tag{26}$$

This decomposition suggests a superior plug-in estimator since $\Pr(B_i = k)$ can be more precisely estimated than $\Pr(B_i = k \mid D_i = d)$. We refer to the plug-in estimator of bounds associated with (26) as post-stratification estimator.

We derive an approximate variance estimator for the post-stratification estimator by analogy to estimators following blocked randomization. The difference-in-means estimator under blocking may be represented as the population-weighted mean of independent estimates across blocks. Denote the feature θ for the kth block with θ^k. Then, by the asymptotic equivalence of the blocking and post-stratification estimators (e.g., Miratrix, Sekhon, and Yu 2013), we have a consistent estimator of the variance of the post-stratification estimator of the lower bound on the outcome in the control group,

$$\widehat{\sigma^R_{(c)_L}}^2 = \sum_{k=1}^{K} \left(\frac{n^k_{(c)1} + n^k_{(t)1}}{n_{(c)1} + n_{(t)1}} \right)^2 \left[\frac{p^k_{(c)1} {s^k_{(c)1}}^2}{n^k_{(c)1}} + \frac{(1-p^k_{(c)1})^2 {s^k_{(c)2L}}^2}{n^k_{(c)2}} + \frac{(1-p^k_{(c)1}) p^k_{(c)1} (\bar{y}^k_{(c)_{2L}} - \bar{y}^k_{(c)1})^2}{n^k_{(c)1}} \right]. \tag{27}$$

Analogous expressions are straightforward to derive for $\widehat{\sigma^R_{(c)_U}}^2$, $\widehat{\sigma^R_{(t)_L}}^2$, and $\widehat{\sigma^R_{(t)_U}}^2$. As a consequence of the law of total variance, the asymptotic variance of the post-stratification estimator is guaranteed to be no larger than that of the unadjusted estimator.

2.7 Sensitivity analysis

Instead of constructing worst-case bounds that impute y^L and y^U for missing follow-up outcomes, we can also conduct an analysis that investigates the sensitivity of our estimates to limited violations of the ignorable missingness assumption in the follow-up sample.[11] Following a suggestion

[10] As usual, covariates are variables that are not affected by the treatment. Continuous covariates need to be coarsened for post-stratification.

[11] See Rosenbaum and Rubin (1983), Rosenbaum (2002), Robins, Rotnitzky, and Scharfstein (1999), Little (2008), and Imai (2009) for different approaches to sensitivity analysis for unmeasured confounding and missing data.

by Little (2008: 428) in the context of sensitivity analysis for likelihood-based methods for data missing not at random, we use a simple mixture model to set up our sensitivity analysis. In a given experimental group, we assume that for a fraction $\delta \in [0,1]$ of subjects with missing follow-up outcomes, ignorability does not hold in that the expected mean outcome for these subjects differs from the expected mean outcome among subjects with observed follow-up outcomes. For the remaining $1-\delta$ fraction of subjects with missing follow-up outcomes we assume that their expected mean outcome is the same as the expected mean outcome among follow-up subjects with observed outcomes. Follow-up subjects with missing outcomes are thus assumed to be drawn from two distributions with mixing (or sensitivity) parameter δ. Subjects for which ignorability does not hold are drawn from a degenerate distribution centered at y^L or y^U. Subjects for which ignorability holds are drawn from an unknown distribution with mean and variance equal to the mean and variance of the outcome distribution for subjects with observed follow-up outcomes.[12]

Given the observed data and sensitivity parameter δ, we can construct bounds on the ATE as in (17)–(18), except that instead of imputing y^L or y^U for follow-up subjects with missing outcomes, we impute a weighted average of y^L or y^U and the observed follow-up group mean, with weights equal to δ and $1-\delta$, respectively:

$$\begin{aligned}\tilde{\hat{\tau}}^L = {} & p_{(t)1} \cdot \bar{y}_{(t)1} + (1-p_{(t)1}) \cdot \Big(p_{(t)2} \cdot \bar{y}_{(t)2} + (1-p_{(t)2})\delta \cdot y^L + (1-p_{(t)2})(1-\delta) \cdot \bar{y}_{(t)2} \Big) - \\ & p_{(c)1} \cdot \bar{y}_{(c)1} + (1-p_{(c)1}) \cdot \Big(p_{(c)2} \cdot \bar{y}_{(c)2} + (1-p_{(c)2})\delta \cdot y^U + (1-p_{(c)2})(1-\delta) \cdot \bar{y}_{(c)2} \Big) \qquad (28)\end{aligned}$$

and

$$\begin{aligned}\tilde{\hat{\tau}}^U = {} & p_{(t)1} \cdot \bar{y}_{(t)1} + (1-p_{(t)1}) \cdot \Big(p_{(t)2} \cdot \bar{y}_{(t)2} + (1-p_{(t)2})\delta \cdot y^U + (1-p_{(t)2})(1-\delta) \cdot \bar{y}_{(t)2} \Big) - \\ & p_{(c)1} \cdot \bar{y}_{(c)1} + (1-p_{(c)1}) \cdot \Big(p_{(c)2} \cdot \bar{y}_{(c)2} + (1-p_{(c)2})\delta \cdot y^L + (1-p_{(c)2})(1-\delta) \cdot \bar{y}_{(c)2} \Big). \qquad (29)\end{aligned}$$

Setting $\delta = 0$ implies that ignorability holds for all follow-up subjects with missing outcomes and so the ATE is point-identified (i.e., $\tilde{\hat{\tau}}^L = \tilde{\hat{\tau}}^U$ since all terms involving y^L or y^U drop out). We call this estimate the naive estimate. Setting $\delta = 1$ implies that ignorability does not hold for any follow-up subject with missing outcomes and (28) and (29) simplify to the worst-case bounds under double sampling derived in (17) and (18). Values of δ strictly between zero and one imply identification regions that are strictly smaller than the worst-case bounds.

As before, we can construct a confidence interval around the identification region implied by any given value of δ and the observed data.[13] In cases in which the confidence interval around the naive estimate (i.e., setting $\delta = 0$) does not include zero but the confidence interval around the

[12] For binary outcomes, equality of means implies equality of variances. For non-binary outcomes, we could drop the equality of variances assumption and use Popoviciu's inequality to bound the variance (Sharma, Gupta, and Kapoor 2010). We do not pursue this approach here.

[13] Let $s^2_{(c)2}$ be the variance of the observed outcomes in the follow-up control group and $s^2_{(t)2}$ be the variance of the observed outcomes in the follow-up treatment group. Since we assumed that follow-up subjects with ignorably missing outcomes are drawn from a distribution with the same mean and variance as subjects with observed follow-up outcomes, our mixture distribution consists of two components with mixture weight $w = p_{(c)2} + (1-p_{(c)2})(1-\delta)$, the proportion of follow-up subjects with observed outcomes plus the proportion of follow-up subjects with ignorably

worst-case bounds (i.e., setting $\delta = 1$) does, we can perform a line search to find the value of δ at which the confidence interval begins to contain zero. Call this value δ^*. When δ^* is close to zero, inferences are very sensitive to even small violations of the ignorability assumption; when δ^* is close to one, inferences are not very sensitive to even large violations of the ignorability assumption.

The sensitivity analysis proposed here assists researchers as they assess the consequences of varying degrees of violation of the ignorability assumption. We provide an illustration of its usefulness in the application section below.

3 Simulation Study

To illustrate the implications of the initial sample and follow-up sample response rates ($p_{(c)1}$, $p_{(t)1}$, $p_{(c)2}$, $p_{(t)2}$) for the width of the identification region and 95% confidence intervals, we conducted a small simulation study. The parameters are as follows: 800 units are split evenly between treatment and control. For simplicity, we assume the same response rates in each condition across both waves ($p_{(c)1} = p_{(t)1} = p_1$ and $p_{(c)2} = p_{(t)2} = p_2$). We vary the proportion p_1 that responds in the initial sample in each group from 0.25 to 0.75. The average outcome in the initial sample is 0.5 in both groups and the variance is 0.25. We sample exactly 50 of the missing control units and 50 of the missing treatment units in the follow-up sample; we vary the proportion p_2 that responds in the follow-up sample from 0 to 1. The average outcome among those who respond in the follow-up sample is set to 0.5. The variances, as in the initial sample, are 0.25 for each group. In our simulation, the true ATE is equal to 0.

[Figure 1 about here.]

Figure 1 displays the results of our simulation. The left panel shows how the identification region and the 95% confidence interval for the ATE respond to changes in p_2. When p_2 is equal to zero, the intervals are the same as the worst-case intervals that would be obtained in the absence of double sampling. When p_2 is equal to 1—all subjects respond in the follow-up sample—the ATE is point-identified. We consider three scenarios, corresponding to an initial sample response rate of 0.25, 0.50, or 0.75. As the initial sample response rate increases, the identification region gets smaller. This pattern is more easily seen in the right panel, which plots the width of the identification region and associated confidence interval as a function of p_2 for each p_1 scenario. The lower the initial sample response rate, the wider the identification region. The width of this region shrinks to 0 as the follow-up sample response rate increases.[14]

missing outcomes. In turn, $1-w$ represents the weight given to subjects with missing follow-up observations for which ignorability does not hold, which are drawn from a degenerate distribution centered at y^L or y^U. From the law of total variance, we have $s^2_{(c)2L} = w \cdot s^2_{(c)2} + w(1-w) \cdot \left(\bar{y}_{(c)2} - y^L\right)^2$, with $s^2_{(c)2U}$, $s^2_{(t)2L}$, and $s^2_{(t)2U}$ defined analogously. When conducting a sensitivity analysis we plug these expressions into (19)–(22). Naturally, we also substitute the bounds on the follow-up group means given in (28)–(29) for the corresponding worst-case bounds on the follow-up group means in (19)–(22). For example, when computing $\widehat{\sigma}^2_{(c)_L}$ we substitute $p_{(c)2} \cdot \bar{y}_{(c)2} + (1-p_{(c)2})\delta \cdot y^L + (1-p_{(c)2})(1-\delta) \cdot \bar{y}_{(c)2}$ for $\bar{y}_{(c)_{2L}}$.

[14] In the online appendix, we also consider the case in which the total sample size increases by a factor of 100.

4 Application

We replicate a survey experiment originally conducted by Levendusky and Malhotra (2016) that tested the effects of news stories on subjects' policy views.[15] We conducted our replication on Amazon's Mechanical Turk. Subjects could be exposed to one of three treatments: a "moderate" condition, in which subjects read an article describing the electorate as focused on finding common ground, a "polarized" condition, in which the article describes politics as contentious and the electorate as sharply divided, or a placebo condition (not analyzed here). The finding from the original experiment was that, compared to subjects in the moderate condition, subjects in the polarized condition reported more centrist policy positions but perceived typical partisans as holding more divergent views.

Our replication uses the identical experimental stimuli and outcome variables.[16] While the original experiment only collected immediate outcomes, our replications collected twice: immediately in a Wave 1 survey and again in a Wave 2 survey conducted 10 days later. The dependent variable, *Perceived Polarization*, is built from subjects' responses to a series of policy questions. After giving their response to each question in the list below, subjects were asked how they think a "typical Democratic voter" and a "typical Republican voter" would respond to each question. The outcome variable is the average of the absolute values of the differences in subjects' Democratic and Republican Responses.[17]

1. The tax rates on the profits people make from selling stocks and bonds, called capital gains taxes, are currently lower than the income tax rates many people pay. Do you think that capital gains tax rates should be increased, decreased, or kept about the same? (7 point scale. Decreased a lot: -3; Increased a lot: 3)

2. There is some debate about whether or not undocumented immigrants who were brought to this country illegally as children should be deported. Which of the following positions on the scale below best represents your position on this issue? (7 point scale. Very strongly oppose deportation: -3; Very strongly support deportation: 3)

3. The United States is currently considering signing additional free trade agreements with Central American, South American, and Asian countries. The Democratic Party wants to make it more difficult for the U.S. to enter into such agreements. The Republican Party wants to make it easier to do so. What do you think? Do you support or oppose the United States signing more free trade agreements with Central American, South American, and

As expected, the width of the 95% confidence interval shrinks with the increased sample size, but the width of the identification region does not. This illustrates the fact that larger samples do not reduce fundamental uncertainty, but higher response rates do.

[15] The replication materials for analyses reported here are available at http://dx.doi.org/10.7910/DVN/AQB4MP.

[16] See the online appendix for the full text of the treatments. We obtained the wording of the stimuli and outcome variables from the publicly available archive hosted on the Time-sharing Experiments for the Social Sciences website. We made small aesthetic changes to the charts that accompanied the treatment text.

[17] See the online appendix for an analysis of the effects of the treatments on the *Extremity* of subjects' own policy positions.

Asian countries? (7 point scale. Very strongly oppose free trade: -3; Very strongly support free trade: 3)

4. Public financing of elections is when the government pays for the cost of campaigning for various offices, rather than the campaigns relying on donations from the general public, corporations, or unions. Democrats typically support public financing plans while Republicans have wanted to eliminate them. What do you think? Do you support or oppose the government paying for the public financing of elections? (7 point scale. Very strongly oppose public financing: -3; Very strongly support public financing: 3)

The estimand in this application is the sample average treatment effect in wave 2 among those who participated in wave 1 of the experiment.

4.1 Attrition and double sampling

We encounter attrition when subjects who responded in Wave 1 fail to do so again in Wave 2. The "initial sample" refers to the first round of data collection in Wave 2 and the "follow-up sample" refers to the data collection effort among those selected for double sampling. Our experiment comprises 1,980 subjects, of which 1,444 had non-missing responses in the initial sample.

We assigned exactly 50 non-respondents at random from each treatment group to the follow-up sample. While the initial compensation for the Wave 2 survey was $1.00, we offered selected non-respondents $4.00 to participate. Of these 100 subjects, 72 completed the Wave 2 survey. Table 1 summarizes the number of subjects in each condition.

[Table 1 about here.]

4.2 Results

Table 2 reports the average outcomes in each condition for subjects who initially responded in Wave 2 and for those who responded when offered additional incentives. A naive analysis (using the difference-in-means estimator with sampling weights) would conclude that the effect of the *polarized* condition was to raise average outcomes by 0.159 (SE: 0.087, two-sided *p*-value: 0.069). However, the naive estimator is only consistent if missingness is indeed ignorable. Table 2 also shows that the average outcomes in the initial and follow-up samples do not differ dramatically.

[Table 2 about here.]

Table 3 illustrates what happens when we do not assume that missing outcomes are ignorable and use worst-case bounds instead. To estimate a lower bound on the ATE using responses only from Wave 2 without any double sampling, we replace missing outcomes in the Polarized group with the lowest possible values (0) and missing outcomes in the Moderate group with the highest possible values (6). To obtain an upper bound on the ATE, we reverse the procedure. The bounds, as expected, are so wide as to be of little probative value. The estimated bounds are $(-1.54, 1.71)$ and the associated 95% confidence interval equals $(-1.67, 1.84)$.

[Table 3 about here.]

Our proposed method, double sampling combined with worst-case bounds, leads to dramatic reductions in the widths of the worst-case bounds and the associated 95% confidence intervals. The relatively high response rates in the follow-up samples imply that fewer subjects receive worst-case imputations for their missing outcomes, so that worst-case bounds on the ATE are much more informative. The double sampling identification region for the ATE is estimated to be $(-0.34, 0.57)$. Adjusting for party identification by post-stratification yields estimated bounds of $(-0.34, 0.53)$ with an associated 95% confidence interval equal to $(-0.53, 0.70)$. Whereas the width of the 95% confidence interval using only the initial sample equals 3.50, double sampling with post-stratification reduces it to 1.23, for a 65 percent reduction in confidence interval width. This large decrease in interval width attests to the sharp reduction in fundamental uncertainty achieved by the double sampling design.

4.3 Sensitivity

We implemented the sensitivity analysis described in Section 2.7. As shown in Figure 2, we obtain a δ^* of 0.07, implying that the statistically significant (at the 10% level) naive estimate is very sensitive to even small violations of ignorability. If we conduct our sensitivity analysis at the 5% level of statistical significance, δ^* does not exist, because the 95% confidence interval around the naive estimate contains zero.

[Figure 2 about here.]

5 Conclusion

We have proposed an approach for addressing missing outcome data in completely randomized experiments that makes fairly weak assumptions but may still yield substantively informative bounds on the average treatment effect. Our approach is based on a combination of the double sampling design (Neyman 1938; Hansen and Hurwitz 1946) and non-parametric worst-case bounds (Birnbaum and Sirken 1950a,b; Manski 1990). We have shown that the resulting bounds on the ATE can be much narrower than the usual worst-case bounds from a single round of data collection. At the same time, our approach avoids the assumption that missingness is ignorable, at least after conditioning on observables. This ignorability assumption is credible in some but not all cases. The outcome stability assumption required by our approach provides an alternate identification assumption for researchers concerned about missing outcome data in their experiments. Finally, our approach can also complement the multiple imputation of missing outcomes, perhaps under the assumption that initial missingness is non-ignorable but follow-up missingness is ignorable conditional on observables (see Little and Rubin 2002: ch. 15).

The advantages of our method are greatest when the initial outcome data collection generates inexpensive outcomes for a substantial fraction of experimental subjects and the follow-up data collection obtains outcomes for a large fraction of a random subsample of subjects with initially missing outcomes. As illustrated in our empirical application, fairly high success rates in the follow-up sample can often be achieved by concentrating significant effort and resources on a small number of subjects (see also An et al. 2009, Jenkins et al. 2009, and Fraser and Yan 2007). It is conceptually straightforward to combine double sampling with additional assumptions to further narrow the identification region (for examples of such a strategy see Manski and Nagin 1998, Manski and Pepper 2011, Mebane and Poast 2013, and Keele and Minozzi 2013).

The approach proposed here can be generalized in various ways. For instance, it would be desirable to extend it to more complicated treatment assignment rules such as clustered random assignment, which introduce complications in variance estimation. It would also be useful to derive rules for optimal resource allocation along the lines of Hansen and Hurwitz (1946) for experimental designs that combine double sampling and worst-case bounds.

References

An, Ming-Wen, Constantine E. Frangakis, Beverly S. Musick, and Constantin T. Yiannoutsos. 2009. “The need for double-sampling designs in survival studies: An application to monitor PEPFAR.” *Biometrics* 65 (1): 301–306.

Aronow, Peter M. and Donald K. K. Lee. 2013. “Interval estimation of population means under unknown but bounded probabilities of sample selection.” *Biometrika* 100 (1): 235–240.

Baird, Sarah, Joan Hamory, and Edward Miguel. 2008. “Tracking, attrition and data quality in the Kenyan life panel survey round 1 (KLPS-1).” Working paper.

Birnbaum, Z. W. and Monroe G. Sirken. 1950a. “On the total error due to non-interview and to random sampling.” *International Journal of Opinion and Attitude Research* 4: 179–191.

Birnbaum, Z. W. and Monroe G. Sirken. 1950b. “Bias due to non-availability in sampling surveys.” *Journal of the American Statistical Association* 45 (249): 98–111.

Cassel, C. M., C. E. Särndal, and J. H. Wretman. 1983. “Some uses of statistical models in connection with the non-response problem.” In *Incomplete data in sample surveys III. Symposium on incomplete data, proceedings* (W. G. Madow and I. Olkin, eds.). Academic Press, pp. 143–160.

Cochran, William G., Frederick Mosteller, and John W. Tukey. 1954. *Statistical problems of the Kinsey report on sexual behavior in the human male.* American Statistical Association.

Cochran, William G. 1977. *Sampling techniques.* Third edition. John Wiley & Sons.

Coppock, Alexander, Alan S. Gerber, Donald P. Green, Holder L. Kern. 2016. *Replication Data for: Combining double sampling and bounds to address non-ignorable missing outcomes in randomized experiments*. Harvard Dataverse. http://dx.doi.org/10.7910/DVN/AQB4MP

Das, Mitali, Whitney K. Newey, and Francis Vella. 2003. “Nonparametric estimation of sample selection models.” *Review of Economic Studies* 70 (1): 33–58.

DiNardo, John, Justin McCrary, and Lisa Sanbonmatsu. 2006. “Constructive proposals for dealing with attrition: An empirical example.” Working paper.

Ericson, W. A. 1967. “Optimal sample design with nonresponse.” *Journal of the American Statistical Association* 62 (317): 63–78.

Fraser, Gary and Ru Yan. 2007. “Guided multiple imputation of missing data.” *Epidemiology* 18 (2): 246–252.

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics.* Fourth edition. W. W. Norton & Company.

Freedman, David A. and Jasjeet S. Sekhon. 2010. “Endogeneity in probit response models.” *Political Analysis* 18 (2): 138–150.

Frankel, Laura Lazarus and D. Sunshine Hillygus. 2014. “Looking beyond demographics: Panel attrition in the ANES and GSS.” *Political Analysis* 22 (3): 336–353.

Alan S. Gerber, Donald P. Green, Edward H. Kaplan, and Holger L. Kern. 2010. “Baseline, placebo, and treatment: Efficient estimation for three-group experiments.” *Political Analysis* 18 (3): 297–315.

Gerber, Alan S. and Donald P. Green. 2012. *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.

Glynn, Robert J., Nan M. Laird, and Donald B. Rubin. 1993. “Multiple imputation in mixture models for non-ignorable nonresponse with follow-ups.” *Journal of the American Statistical Association* 88 (423): 984–993.

Grogger, Jeffrey. 2012. “Bounding the effects of social experiments: Accounting for attrition in administrative data.” *Evaluation Review* 36 (6): 449–474.

Hansen, Morris H. and William N. Hurwitz. 1946. “The problem of non-response in sample surveys.” *Journal of the American Statistical Association* 41 (236): 517–529.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S. Sekhon. Forthcoming. “From SATE to PATT: Combining experimental with observational studies to estimate population treatment effects.” *Journal of the Royal Statistical Society, Series A.*

Heckman, James J. 1979. “Sample selection bias as a specification error.” *Econometrica* 47 (1): 153–161.

Horowitz, Joel L. and Charles F. Manski. 1998. “Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations.” *Journal of Econometrics* 84 (1): 37–58.

Horowitz, Joel L. and Charles F. Manski. 2000. “Nonparametric analysis of randomized experiments with missing covariate and outcome data.” *Journal of the American Statistical Association* 95 (449): 77–84.

Imai, Kosuke. 2008. “Sharp bounds on the causal effects in randomized experiments with ‘truncation-by-death’.” *Statistics & Probability Letters* 78 (1): 144–149.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. “Misunderstandings between experimentalists and observationalists about causal inference.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171 (2): 481–502.

Imai, Kosuke. 2009. “Statistical analysis of randomized experiments with non-ignorable missing binary outcomes: an application to a voting experiment.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58 (1): 83–104.

Imbens, Guido W. and Charles F. Manski. 2004. “Confidence intervals for partially identified parameters.” *Econometrica* 72 (6): 1845–1857.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction.* Cambridge University Press.

Jenkins, Paul, Charles Scheim, Jen-Ting Wang, Roberta Reed, and Allan Green. 2004. “Assessment of coverage rates and bias using double sampling methodology.” *Journal of Clinical Epidemiology* 57 (2): 123–130.

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. “Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.” *Statistical Science* 22 (4): 523–539.

Kaufman, G. M. and Benjamin King. 1973. “A Bayesian analysis of nonresponse in dichotomous processes.” *Journal of the American Statistical Association* 68 (343): 670–678.

Keele, Luke and William Minozzi. 2013. “How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data.” *Political Analysis* 21 (2): 193–216.

Kern, Holger L., Elizabeth A. Stuart, Jennifer Hill, and Donald P. Green. 2016. “Assessing methods for generalizing experimental impact estimates to target populations.” *Journal of Research on Educational Effectiveness* 9 (1): 103–127.

Lee, David S. 2010. “Training, wages, and sample selection: Estimating sharp bounds on treatment effects.” *Review of Economic Studies* 76 (3): 1071–1102.

Levendusky, Matthew and Neil Malhotra. 2016. “Does media coverage of partisan polarization affect political attitudes?” *Political Communication* 33 (2): 283–301.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical analysis with missing data.* Second edition. Wiley-Interscience.

Little, Roderick J. A. 2008. Selection and pattern-mixture models. In Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs (eds.), *Longitudinal data analysis.* Chapman & Hall/CRC, ch. 18.

Lohr, Sharon L. 2010. *Sampling: Design and analysis.* Second edition. Brooks/Cole.

Manski, Charles F. 1990. “Nonparametric bounds on treatment effects.” *American Economic Review Papers and Proceedings* 80 (2): 319–323.

Manski, Charles F. 1995. *Identification problems in the social sciences.* Harvard University Press.

Manski, Charles F. 2007. *Identification for prediction and decision.* Harvard University Press.

Manski, Charles F. and Daniel S. Nagin. “Bounding disagreements about treatment effects: A case study of sentencing and recidivism.” *Sociological Methodology* 28 (1): 99–137.

Manski, Charles F. and John V. Pepper 2011. “Deterrence and the death penalty: Partial identification analysis using repeated cross sections.” NBER working paper 17455.

McConnell, Sheena, Elizabeth A. Stuart, and Barbara Devaney. 2008. “The truncation-by-death problem: What to do in an experimental evaluation when the outcome is not always defined.” *Evaluation Review* 32 (2): 157–186.

Mebane, Walter R., Jr. and Paul Poast. 2013. “Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data.” *Political Analysis* 21 (2): 233–251.

Miratrix, Luke W., Jasjeet S. Sekhon, and Bin Yu. 2013. “Adjusting treatment effect estimates by post-stratification in randomized experiments.” *Journal of the Royal Statistical Society, Series B* 75 (2): 369–396.

Neyman, J. 1938. “Contribution to the theory of sampling human populations.” *Journal of the American Statistical Association* 33 (201): 101–116.

Peress, Michael. 2010. “Correcting for survey nonresponse using variable response propensity.” *Journal of the American Statistical Association* 105 (492): 1418–1430.

Rao, J. N. K. 1973. “On double sampling for stratification and analytical surveys.” *Biometrika* 60 (1): 125–133.

Robins, James M., Andrea Rotnitzky, and Daniel O. Scharfstein. 1999. “Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models.” In *Statistical models in epidemiology* (M. E. Halloran and D. Berry (eds.). Springer, pp. 1–92.

Rosenbaum, Paul R. 2002. *Observational studies.* Second edition. Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome.” *Journal of the Royal Statistical Society, Series B* 45 (2): 212–218.

Sharma, R., M. Gupta, and G. Kapoor. 2010. “Some better bounds on the variance with applications.” *Journal of Mathematical Inequalities* 4 (3): 355–363.

Si, Yajuan, Jerome P. Reiter, and D. Sunshine Hillygus. 2014. “Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples.” *Political Analysis* 23 (1): 92–112.

Stoye, Jörg. 2009. “More on confidence intervals for partially identified parameters.” *Econometrica* 77 (4): 1299–1315.

Tamer, Elie. 2010. “Partial identification in econometrics.” *Annual Review of Economics* 2: 167–195.

Zhang, Junni L. and Donald B. Rubin. 2004. “Estimation of causal effects via principal stratification when some outcomes are truncated by ‘death’.” *Journal of Educational and Behavioral Statistics* 28 (4): 353–368.

6 Figures

Figure 1: Simulation Study

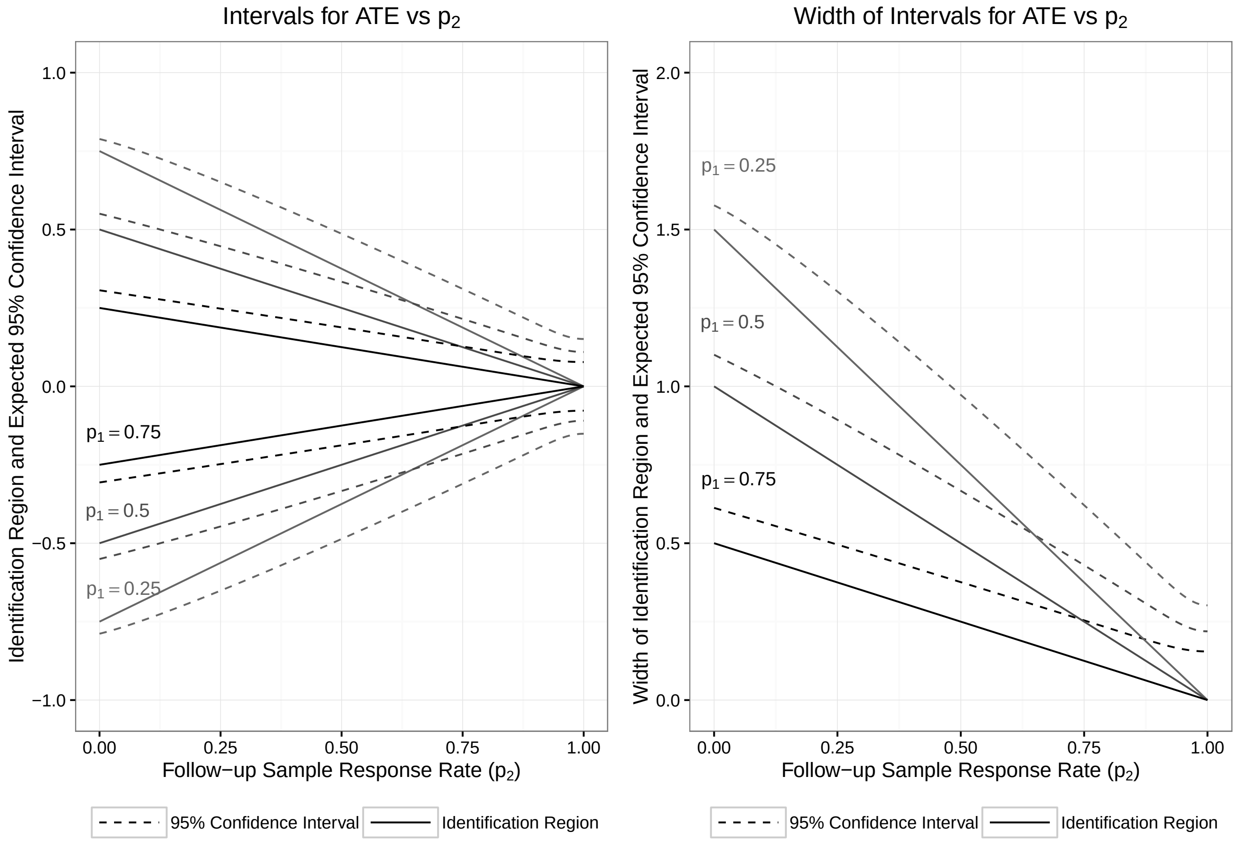
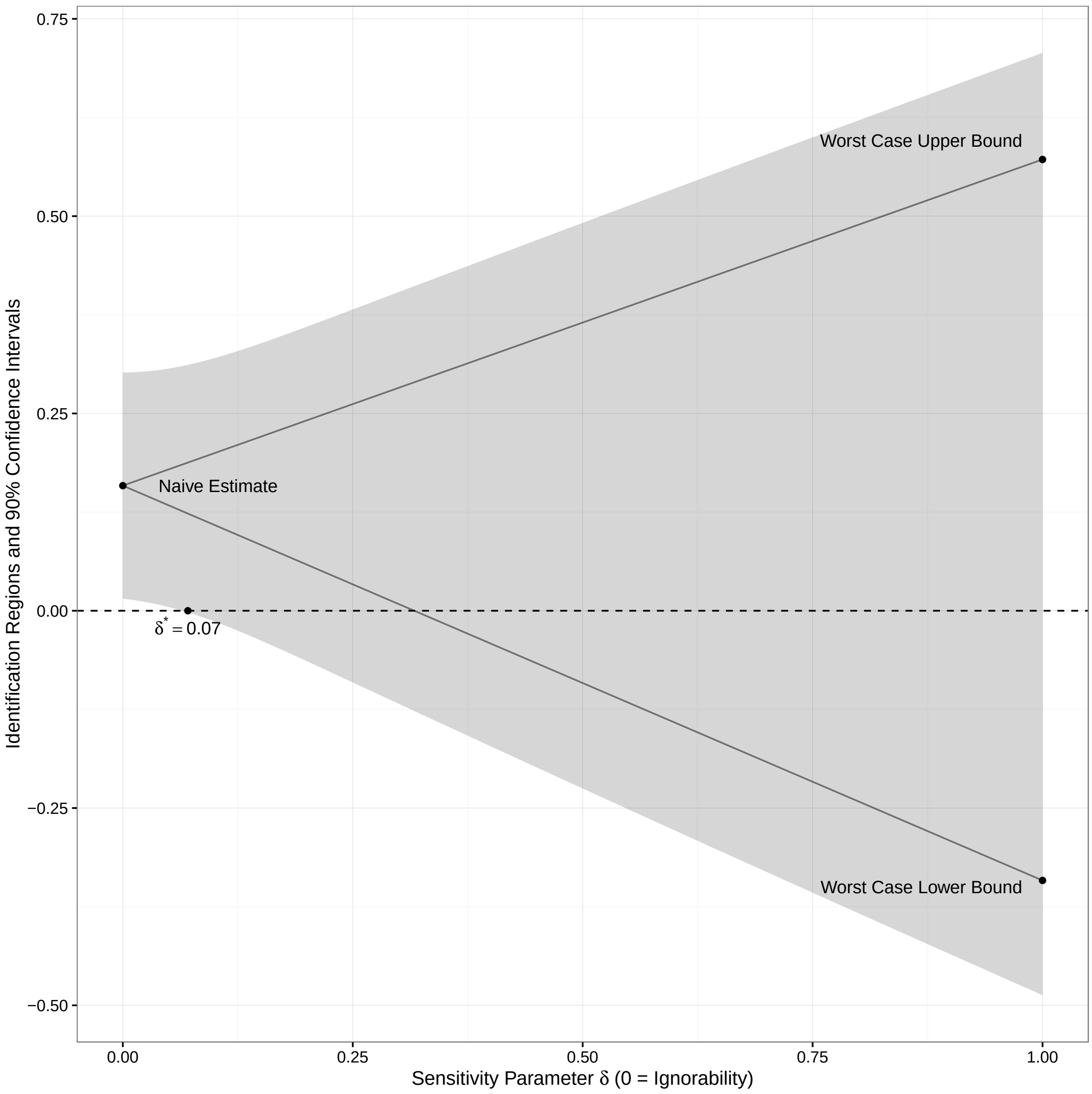


Figure 2: Sensitivity Analysis



7 Tables

Table 1: Attrition by Treatment Condition

	Initial Sample	Follow-up Sample			
	N	Responded	Did Not Respond	Invited	Responded to Invitation
Moderate	995	731	264	50	39
Polarized	985	713	272	50	33
Total	1980	1444	536	100	72

Table 2: Outcome Summaries by Condition and Response Type

	Initial Responders			Double Sampled Responders		
	Mean	SD	N	Mean	SD	N
Moderate	3.542	1.243	731	3.583	1.367	39
Polarized	3.668	1.270	713	3.826	1.313	33

Table 3: Replication Study Results

	DV: Perceived Polarization		
	Worst-Case Bounds	WCB + Double Sampling	WCB + DS + Post-stratification
95% CI Lower Bound	-1.6691	-0.5283	-0.5290
95% CI Upper Bound	1.8359	0.7452	0.6966
Worst-Case Bound: Low Estimate	-1.5391	-0.3417	-0.3444
Worst-Case Bound: High Estimate	1.7097	0.5718	0.5257
Variance of Low Estimate	0.0062	0.0129	0.0126
Variance of High Estimate	0.0059	0.0111	0.0108