Political Analysis (2025), **33**, 140–155
doi:10.1017/pan.2024.13

PA

ARTICLE

Countering Non-Ignorable Nonresponse in Survey Models with Randomized Response Instruments and Doubly Robust Estimation

Michael A. Bailey

Department of Government and McCourt School of Public Policy, Georgetown University, Washington, DC, USA.
Email: Michael.Bailey@georgetown.edu

(Received 3 October 2023; revised 29 March 2024; accepted 2 April 2024; published online 2 January 2025)

Abstract

Conventional survey tools such as weighting do not address non-ignorable nonresponse that occurs when nonresponse depends on the variable being measured. This paper describes non-ignorable nonresponse weighting and imputation models using randomized response instruments, which are variables that affect response but not the outcome of interest. This paper uses a doubly robust estimator that is valid if one, but not necessarily both, of the weighting and imputation models is correct. When applied to a national 2019 survey, these tools produce estimates that suggest there was nontrivial non-ignorable nonresponse related to turnout, and, for subgroups, Trump approval and policy questions. For example, the conventional MAR-based weighted estimates of Trump support in the Midwest were 10 percentage points lower than the MNAR-based estimates.

Keywords: survey nonresponse; doubly robust estimators; non-ignorable nonresponse

Edited by: Jeff Gill

Making inferences about populations from polls is extremely challenging in the modern polling environment. Probability-based samples have nonresponse rates that often exceed 90%, making it hard to justify analyzing them with tools based on random sampling theory. Non-probability samples—a rapidly increasing share of polls (Kennedy, Popky, and Keeter 2023)—are even harder to connect to populations given opaque sampling mechanisms.

Most contemporary tools assume that nonresponse is "ignorable" (or, equivalently, "missing at random" [MAR]). This assumption is that respondents are representative conditional on covariates. The raw sample may have too many college-educated people, but as long as the college and noncollege respondents are representative of their subgroups, the data can be re-weighted in a way that will produce an accurate estimate of the population. Nearest neighbor imputation and multilevel regression with post-stratification also assume the data are MAR.

It is not clear that survey nonresponse is always ignorable because unobserved attributes such as social trust could affect response and predict political opinions. It is also possible that the outcome being measured directly affects response propensity; supporters of former President Donald Trump may have been less inclined to answer polls in 2016 and 2020, for example. In these cases, survey nonresponse is non-ignorable and data are "missing not at random" (MNAR). Estimating MNAR models is challenging. A survey may produce a 40% approval rating when approval is 60% and the most approving are less inclined to respond or when approval is 20% and the most approving are more inclined to respond.

Response instruments are useful to identify MNAR models. They are variables that affect the propensity to respond but do not directly affect the attribute being measured. Ideally, a response instrument is randomized, which makes it much easier to defend the claim that it does not directly affect the outcome being measured and also increases statistical power. Given an assumption that the effect of the treatment does not interact with the outcome of interest, such instruments can identify MNAR models (Sun *et al.* 2018).

There are two modeling strategies when working with a randomized response instrument. First, one can re-weight observed data based on the propensity of response. Second, one can impute the values of the outcome variable for non-respondents. The MAR versions of these estimators do not work in the MNAR context, but following Sun *et al.* (2018) estimation is feasible with a response instrument. I focus in this paper on a doubly robust estimator developed by Sun *et al.* (2018) that combines these two approaches to produce estimates that are consistent if either (but not necessarily both) of the specifications of the MNAR weighting and imputation models is correct.

I apply these methods to a 2019 survey of U.S. adults. MNAR estimates using a randomized response instrument differ from conventionally weighted analyses of the same data. For example, MNAR-based methods suggest that turnout is exaggerated in the MAR-weighted sample. I also show how MNAR models produce estimates of Trump support in the Midwest in 2019 that are 10 percentage points higher than the MAR-weighted model. In addition, the MNAR models estimate less partisan polarization than MAR-weighted models. For example, MNAR models suggest more racial conservatism among Democrats and less support for tax cuts among Republicans than suggested by conventional weighting models.

This paper proceeds as follows. Part 1 explains response instruments in MNAR models. Part 2 discusses MNAR weighting, and Part 3 discusses MNAR imputation. Part 4 explains the doubly robust methods used here. Part 5 analyzes 2019 survey data.

1. The Vital Role of Response Instruments in MNAR Models

There are two approaches to estimating population values in the face of missing data (Chen, Li, and Wu 2020):

1. **Weighting** is the most widely used method to deal with nonresponse. "Inverse propensity weighting" (IPW) weights observations by the inverse of the probability of being observed conditional on a set of covariates X. Let $R_i = 1$ if individual i responds to the survey. In a MAR context, $\hat{\pi} = Pr(R = 1|X)$. An IPW estimate of the population mean is

$$\hat{\bar{Y}}_{IPW} = \sum_i \frac{R_i Y_i}{\hat{\pi}_i}.$$

2. **Imputation** estimates the expected value of Y for individuals who did not respond (Vermeulen and Vansteelandt 2016). There are many possible ways to impute missing outcome data, ranging from OLS to multilevel regression with post-stratification to nearest neighbor imputation (Chen *et al.* 2020). Assuming that Y is dichotomous, we can specify that value of Y for an individual who did not respond as $\hat{Y}_i^0 = Pr(Y_i = 1|X_i, R_i = 0)$. In a MAR context, $Pr(Y_i = 1|X_i, R_i = 0) = Pr(Y_i = 1|X_i, R_i = 1)$. An imputation estimate of the population mean is

$$\hat{\bar{Y}}_{Imp} = \frac{1}{N}\sum_i [R_i Y_i + (1 - R_i)\hat{Y}_i^0].$$

These approaches are straightforward if the data are MAR. Weights are generated via a model predicting response as a function of X; imputed values of Y for missing data are generated as the fitted values of Y based on observed data.

There are two ways that the MAR assumption may be violated. First, an unmeasured factor may affect both response propensity and Y even after controlling for observed covariates. For example, unmeasured social trust may predict response and disapproval of Donald Trump. If true, conventional weighting will produce a sample less supportive of Trump than the population, even after weighting on standard demographics (Clinton *et al.* 2021). Second, Y may directly affect response propensity. It could be that supporting Trump directly predicts nonresponse even after controlling for the weighting variables. This paper focuses on models in which Y directly affects R; a model with common unobserved factors in response and outcome can be subsumed into this model.

If survey data are MNAR, bias can be large. Meng (2018) derives a general equation that characterizes the sampling error for any sample, be it random or nonrandom:

$$\overline{Y}_n - \overline{Y}_N = \underbrace{\rho_{R,Y}}_{\text{data defect correlation}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{data quantity}} \times \underbrace{\sigma_Y}_{\text{data difficulty}} . \tag{1}$$

Equation (1) shows that sampling error is a function of the "data defect correlation" which is the correlation of R and Y, a data quantity term and a data difficulty term.[1] The data defect correlation characterizes the degree of non-ignorability in the sample. If it is zero, then there is no error. (Since the correlation is unlikely to be literally zero even for a random sample, there will still be sampling error in random samples.) The key insight from Meng's equation is that the non-ignorability term interacts with a data quantity term that is a function of population size. This comes as a bit of a shock for survey researchers steeped in random sampling theory, but for nonrandom sampling even a small amount of non-ignorability (via the $\rho_{R,Y}$ term) can induce large sampling errors if the population is large. We have known that large samples can go awry since the *Literary Digest* polling fiasco of 1936 (Lusinchi 2012).

Recent advances have expanded our toolkit for dealing with non-ignorable nonresponse. Hartman and Huang (2023) use sensitivity analysis to assess when non-ignorable nonresponse is a threat. Isakov and Kuriwaki (2020) and Bradley *et al.* (2021) combine the Meng equation with known benchmarks to present evidence that even seemingly small levels of non-ignorability can lead to large biases. Such sensitivity-based methods do not provide estimates of the influence of non-ignorability based on the data and models, however.

In this paper, I model nonresponse by allowing the response variable to directly affect the probability an individual responds. Suppose, for example, that

$$Pr(R = 1|X, Y) = g(\gamma_0 + \gamma_X X + \gamma_Y Y). \tag{2}$$

The data generating process is MNAR if $\gamma_Y \neq 0$. We cannot use standard methods of estimating the probability of response because such a model needs observations for which R = 1 and R = 0 and we will have missing data for Y for all individuals with R = 0. We also cannot implement MAR-type imputation because $Pr(Y_i = 1|X_i, R_i = 0) \neq Pr(Y_i = 1|X_i, R_i = 1)$ when Y and R are jointly determined.

While some believe that MNAR data are inherently beyond the scope of modeling, there is a vast literature on how to estimate such models. One theme is that these models should incorporate a response instrument. A response instrument

- influences response, conditional on covariates and
- is not directly related to the outcome of interest in the population, conditional on covariates.

While it is not strictly necessary to have a response instrument (because, e.g., some parametric models can be estimated without one), the field recognizes the poor performance of MNAR models when there is no response instrument (see, e.g., Peress 2010).

[1] This equation does not condition on covariates. Meng (2018) provides a weighted version. One can also condition on covariates by using this equation for subgroups defined by the covariates.

While a response instrument can in theory be observational, a response instrument is ideally based on a randomization treatment that affects response propensity and does not directly affect Y. Such an instrument can be created by randomly assigning individuals to a treatment protocol that increases the response probability. For example, *New York Times*/Sienna pollsters randomly assigned adults in Wisconsin in 2022 to a treatment in which they received $30 for responding (Cohn 2022). Randomization makes it easier to accept the condition that the instrument has no direct effect on Y; randomization also increases statistical power by reducing dependence of the instrument on covariates.

Sun *et al.* (2018) explain how to use an instrument to estimate MNAR models. They formulate their models in terms of the following selection bias function that "quantifies the degree of association between Y and R given (X, Z) on the log odds scale":

$$\eta(x,y,z) = log[\frac{Pr(R=1|x,y,z)}{Pr(R=0|x,y,z)} / \frac{Pr(R=1|x,Y=0,z)}{Pr(R=0|x,Y=0,z)}]. \tag{3}$$

The Supplementary Material shows that if we use a logit function for the response equation (and impose the condition described momentarily), then η is simply the coefficient on Y in the response equation (Equation (2)). I use a logit model in what follows as doing so simplifies presentation and implementation; the model can be conceptualized and implemented with other specifications.

Sun *et al.* (2018, 6) derive the conditions under which MNAR models are identified when one has a correct model for the selection bias and one assumes (as in MAR models) that every (X,Y,Z) has a nonzero probability of responding.[2] The identification conditions can be summarized with reference to a sufficient condition: the MNAR models discussed below are identified as long as Y and Z do not interact in the response equation.

To illustrate the identification assumption of an MNAR model as compared to a MAR model, consider the following model in which response depends on a response instrument Z, covariates X, the outcome Y, and all linear interactions.

$$Pr(R=1|Z,X,Y) = g(\gamma_0 + \gamma_Z Z + \gamma_x X + \gamma_Y Y + \gamma_{ZX} ZX + \gamma_{XY} XY + \gamma_{ZY} ZY).$$

- Assumptions in MAR-model: $\gamma_Y = \gamma_{XY} = \gamma_{ZY} = 0$.
- Assumptions in MNAR-model: $\gamma_{ZY} = 0$.

In a MAR model, Y does not affect response, whether or not this effect interacts with an X variable that also affects Y or an instrument that only affects R. In other words, MAR models assume that conditional on covariates, the people who respond do not have systematically different values of Y than the people who do not respond.

MNAR models allow Y to affect response. The effect of Y can vary depending on X; the effect of Z can also depend on X. In the data analysis below, I estimate MNAR models by subgroup, which allows us to measure the effects of Y and Z on response for selected values of X.

The MNAR models do have an important assumption, however: that $\gamma_{ZY} = 0$. This assumption is violated if the treatment interacts with the Y in affecting response propensity as would happen, for example, in a survey about support for President Biden if the treatment had a different effect on those who approved of Biden. In the extreme, a response instrument that pulled in only Biden supporters could be taken by an MNAR model as indicating that the survey underestimated Biden support. This could lead to a biased estimate—and potentially a worse estimate than a MAR-based model.

[2] Assuming that Z has no direct effect on Y, the necessary and sufficient condition to identify the joint distribution of the full data is that $\frac{P_{\theta_1}(R=1|z,y)}{P_{\theta_2}(R=1|z,y)} \neq \frac{P_{\xi_2}(y)}{P_{\xi_1}(y)}$. This means that for any set of parameters θ and ξ that define response and outcome equations, it is not possible to find another set of parameters that produce the same equality of ratios. In a corollary and examples, Sun *et al.* (2018) show that a sufficient (but not necessary) condition for identifiability is that the effects of Z and Y on R are separable (e.g., $Pr(R=1|Z,Y) = logit[q(Z) + h(Y)]$ for unknown functions $q()$ and $h()$). Conditioning on X does not change these results.

A researcher using an MNAR model therefore needs to pay considerable attention to the design and defense of the instrument. In a political context, the instrument should be depoliticized and reflect the kinds of steps pollsters generically use to affect response propensity.

Suppose there are two survey protocols A and B, with B having a higher response rate. An MNAR model will perform well if the underlying response propensity function is the same for both protocols, with the only difference being the threshold above which people respond is different under treatment. If protocol B induces a differential response that depends on Y, then this would be problematic for MNAR models. Note, however, that a survey firm using only protocol B would have non-ignorable nonresponse and would also have bias. In short, for MAR models, Y cannot affect R; for MNAR models, Y can affect R but needs to do so in the same way under control and treatment.[3]

MNAR models cannot be taken as comprehensive tests of non-ignorability. Finding a significant effect of Y on R in an MNAR model could be due to actual non-ignorability or due to an instrument that violates the identifying assumption of the model. MNAR models should instead be treated as an important part of the diagnostic toolkit. If one finds MNAR models diverging from MAR models—as we do below in turnout and other models, for example—it seems prudent to be transparent about assumptions underlying each set of estimates. If one believes the response in the overall survey is ignorable, the weighted estimates are best; if one believes that the effect of the treatment does not depend on Y, the MNAR estimates are best.

There are many ways to estimate MNAR models. In this paper, I focus on weighting and imputation. Other options include maximum likelihood (Heckman 1979; Tchetgen and Wirth 2017) and copula methods (Marra *et al.* 2017). Bailey (2024) argues that all these methods benefit from randomized response instruments.

2. MNAR Weighting

MAR-based weighting uses weights based on results from a first-stage model that predicts whether or not someone responds. If Y directly affects response, however, a conventional MAR-based model would require information that is unobserved for non-respondents.

MNAR-based weighting models proceed differently: they leverage the orthogonality condition implied by randomized response instruments to generate weights in which Y does affect response.

Consider the following MNAR model for response:

$$\pi = Pr(R = 1|Z,X,Y) = logit(\gamma_0 + \gamma_Z Z + \gamma_X X + \gamma_Y Y), \tag{4}$$

where Z is a randomized response instrument and X is a vector of covariates that affect response and outcome.

There are two key parameters in this model. The coefficient on Z (γ_Z) reflects the effect of the response instrument on response. This parameter needs to be nonzero for the MNAR models discussed here to be identified. The coefficient on Y in the response equation (γ_Y) reflects the effect of Y on response. If this parameter is nonzero, then Y directly affects response, which indicates a non-ignorable/MNAR data generating process. If $\gamma_Y = 0$, conventional weighting is better on efficiency grounds as I demonstrate in the Supplementary Material (see also Bailey 2024).

The theory of Sun *et al.* (2018) allows for semi- and non-parametric estimation of these values. In this paper, I use a logit link function that is of the form above. In the simulations, there are no interactions; in the application, I allow for Y and Z to interact with X by looking at subgroups based on X.

[3]Coppock *et al.* (2017) propose selecting a random set of non-respondents and doing whatever it takes to get them to respond. Doing so is appealing for several reasons, including that one could test whether Y had differential effects under control and treatment. The challenge with their approach is feasibility: even academic surveys in the field for months struggle to get 40% response rates. To the extent that the extra effort individuals respond at a higher—but not perfect—rate, we would have a randomized response instrument such as described here, with all its benefits and limitations.

An MNAR-IPW estimator is in the form

$$\begin{aligned}\hat{\overline{Y}} &= \sum_i \frac{R_i Y_i}{\hat{\pi}} \\ &= \sum_i \frac{R_i Y_i}{logit(\gamma_0 + \gamma_Z Z_i + \gamma_X X_i + \gamma_Y Y_i)}.\end{aligned} \tag{5}$$

As noted above, we cannot use the conventional weighting approach of estimating the response equation first because Y is missing for all $R = 0$ observations. We therefore estimate the model with estimating equations. These are similar to methods of moment estimators in which "population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population" (Cameron and Trivedi 2005, 135).

The estimating equations impose the following properties on the model: the weights sum to the sample size, the IPW weighted sums of the Z and X covariates are the same as the sums in the population and, critically, that Z conditional on X is orthogonal to the weighted Y which is the estimated Y in the full population. Given Equation (4), the estimating equations are

$$\begin{aligned}h[1]: &\ \sum_i \frac{R_i}{\hat{\pi}_i} = N, \\ h[2]: &\ \sum_i \frac{R_i}{\hat{\pi}_i} Z_i = \sum Z_i, \\ h[3]: &\ \sum_i \frac{R_i}{\hat{\pi}_i} X_i = \sum X_i, \\ h[4]: &\ \sum_i \frac{R_i Y_i}{\hat{\pi}_i} (Z_i - \hat{Z}_i) = 0,\end{aligned}$$

where $\hat{Z}_i$ is the expected value of Z given X, which for randomized Z should be close to the proportion of people assigned to the treatment condition.

The estimation process finds the γ values that satisfy the estimating equation conditions. The third equation (h[3]) corresponds to a sample weighting condition as the weighted distribution of X will reflect the (known) population distribution of X. The fourth equation imposes the orthogonality of weighted Y and Z minus $\hat{Z}$. Standard errors are calculated based on a sandwich estimator as described in the Supplementary Material of Sun *et al.* (2018).

Digging into the estimation helps us appreciate how estimating equations produce a coefficient on Y in the response equation even though Y is missing for $R = 0$ individuals. Suppose that $\gamma_Y > 0$. This will produce more $R = 1$ observations among the $Y = 1$, $Z = 1$ group than the $Y = 0$, $Z = 1$ group. In the unweighted data, Y and Z will be correlated. The orthogonality condition in the estimating equation requires that Y and Z are conditionally uncorrelated in the full sample, which is achieved with a $\hat{\gamma}_Y > 0$ which produces a smaller weight on the $Y = 1$, $Z = 1$ observations compared to the $Y = 0$, $Z = 1$ observations. If γ_Y were zero, we would not systematically observe more $R = 1$ observations among the $Y = 1$, $Z = 1$ group than the $Y = 0$, $Z = 1$ group and Y and Z can be orthogonal in the population without down-weighting the $Y = 1$ observations. Section 4 of the Supplementary Material provides more detail on the intuition behind estimating equations in this context by providing a numerical example.

3. MNAR Imputation

MNAR imputation models require a model that predicts $Y|X$ for $R = 0$ individuals. In MAR models, $Y|X$ is the same by assumption among respondents and non-respondents. This is not the case for MNAR models. Hence, we need to formulate a model of $Y|X$ for unobserved individuals as a function of the $Y|X$ for the observed individuals. Tchetgen-Tchetgen, Robins, and Rotnitzky (2010) formulate the joint

density of (y,r) as a function of the selection bias function (see also Burger and McLaren 2017; Little and Rubin 2020, 353):

$$f(y,r) \propto exp[(r-1)\eta]f(y|R=1,x,z)Pr(R=1|Y=0,x,z).$$

The marginal distribution of y is

$$f(y|r,x,z) = \frac{P(y,r|z,x)}{\int P(y,r|z,x)d\mu(y)} = \frac{exp[(r-1)\eta]f(y|R=1,x,z)}{E[exp[(r-1)\eta]|R=1,x,z]},$$

which implies that the distribution of y when $R=0$ can be written as a function of $y|R=1$:

$$f(y|R=0,x,z) = \frac{exp[-\eta]f(y|R=1,x,z)}{E[exp[-\eta]|R=1,x,z)]}. \tag{6}$$

To estimate γ_Y for the MNAR imputation model, first estimate $f(y|R=1,x,z)$. Note that the distribution of Y given that $R=1$ is a function of X *and* Z. Even though Z is assumed not to affect Y in the full population, it can affect the expected value of Y in the observed sample via the effects of non-random selection. This equation can be, for example, a saturated model such as $Y = \lambda_0 + \lambda_Z Z + \lambda_X X + \lambda_{ZX} ZX$ for a simple model with only Z and a single X. With this model in hand, one can calculate $\hat{Y}_i^0$, the expected value of Y for $R=0$ observations given the density in Equation (6), which is a function of the selection bias function (η) and the distribution of $Y|X,Z,R=1$.

The expected value of Y_i for each i is $R_i Y_i + (1-R_i)\hat{Y}_i^0$. To satisfy the estimating equation condition that Y and Z are independent in the full population, we find the value of γ_Y that solves

$$\sum_i (Z_i - \hat{Z}_i)\{R_i Y_i + (1-R_i)\hat{Y}_i^0\} = 0. \tag{7}$$

The MNAR imputation estimate uses the estimates of η and γ_Y from the above process to estimate a population average

$$\hat{\bar{Y}} = \frac{1}{N}\sum_i [R_i Y_i + (1-R_i)\hat{Y}_i^0]. \tag{8}$$

4. Doubly Robust Estimation in MNAR Context

A doubly robust estimator is consistent if one or both of the weighting and imputation models is correct (Scharfstein, Rotnitzky, and Robins 1999; Vermeulen and Vansteelandt 2016).

A doubly robust estimator for the population average is

$$\hat{\bar{Y}} = \frac{1}{N}\sum\left[\frac{R_i Y_i}{\hat{\pi}_i} - \frac{R_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{Y}_i^0\right] \tag{9}$$

$$= \frac{1}{N}\sum\left[\frac{R_i}{\hat{\pi}_i}(Y_i - \hat{Y}_i^0) + \hat{Y}_i^0\right]. \tag{10}$$

If the propensity model is correct, Equation (9) converges to the (correctly) weighted observed data because the average of $R_i - \hat{\pi}$ converges to zero for a correctly specified propensity model. Equation (10) is mathematically identical. In a MAR context, if the imputation is correct, the observed difference of Y and $\hat{Y}_i^0$ will converge to zero and the doubly robust estimate will converge to the correctly specified (by assumption) value of Y. In an MNAR context, this convergence depends on knowing the η function (Equation (3)); see the proof of Proposition 3 in the Supplementary Material of Sun *et al.* (2018).

Figure 1 illustrates how doubly robust estimation works. These simulations highlight (1) the well-known result that MAR models (even doubly robust ones) are biased when nonresponse is non-ignorable and (2) that even when one is very wrong with one or the other propensity or

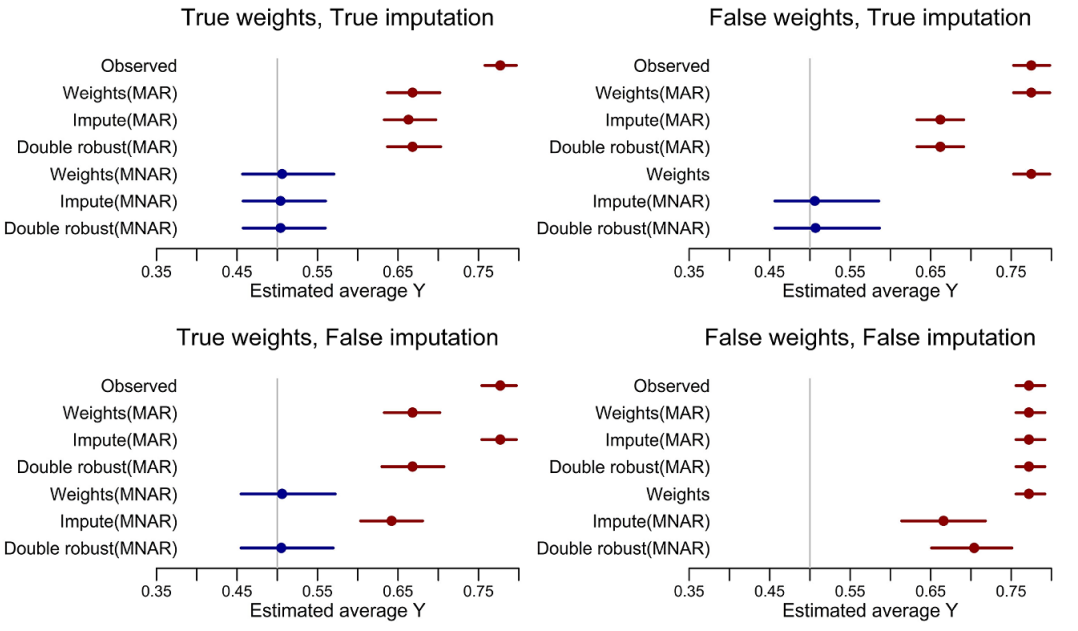

Figure 1. Simulation results for MNAR data.

imputation models, the doubly robust MNAR estimator can produce unbiased results. With better propensity or imputation models, the differences between those models and the doubly robust model would be less stark. Suppose that the true models are

$$Pr(R = 1|X, Y, Z) = logit(\gamma_0 + \gamma_Z Z + \gamma_{X1} X1 + \gamma_{X2} X2 + \gamma_Y Y),$$
$$Pr(Y = 1|X, Y, Z) = logit(\beta_0 + \beta_{X1} X1 + \beta_{X2} X2),$$

where $\gamma_Z > 0$ and $\gamma_Y > 0$. I estimate the true specifications and also two misspecified models

$$Pr(R = 1|X, Y, Z) = logit(\gamma_0),$$
$$Pr(Y = 1|X, Y, Z) = logit(\beta_0).$$

The misspecified response equation assumes everyone (whatever their X values) has the same probability of responding. The misspecified outcome equation assumes everyone (whatever their X values) has the same probability of having $Y = 1$. The point is not that these are good models; rather, the point is that the doubly robust approach can overcome weaknesses in one (but not necessarily both) of the propensity and imputation models.

The upper left of Figure 1 shows the observed mean in the data and the population estimates for the IPW, imputation and doubly robust estimators assuming MAR and MNAR, respectively, given that the correct specifications are used. The bars indicate the 5th and 95th percentile results from 200 simulations. The MAR models do poorly because they do not capture the effect of Y on R. All the MNAR models do well, which is unsurprising because the weighting and imputation models are correctly specified.

The upper right of Figure 1 shows estimates when the response equation is misspecified. The MAR models do poorly. The IPW weighted model here is not an MNAR model because it does not include a γ_Y term; it too does poorly. The MNAR imputation model is fine and—importantly—the MNAR doubly robust model overcomes the poor properties of the misspecified propensity model.

The lower left panel shows what happens when the response model is correct and the imputation model is incorrect. The MAR models do poorly. MNAR weights produce good estimates, while the

MNAR imputation model does poorly; critically, the MNAR doubly robust overcomes the weakness of the imputation model.

The lower right panel shows that the doubly robust estimator can be off if both specifications are incorrect. This is an illustrative example; it is possible to create simulations in which one or both of the MNAR IPW and MNAR imputation methods are less poorly specified and the MNAR doubly robust estimates do better than in Figure 1. In Section 3 of the Supplementary Material, I show simulations in which the MNAR estimators become less precise when the instrument has a weaker effect on response.

5. Survey Example

This section analyzes a March 2019 Ipsos Knowledge Panel survey that included a randomized treatment that discouraged response. This approach is only one of many possible instruments (see, e.g., Cohn 2022). Generally, what is needed is for a survey firm to implement two protocols: one standard protocol and one that increases or decreases response rates.

Figure 2 shows the structure of the data. To be concrete, the figure shows sample sizes for the turnout question; there are slight differences in nonresponse patterns for other dependent variables. The original pool of 3,573 people is drawn from Ipsos's online panel that was recruited using an address-based probability sample. These individuals are randomly assigned to control and treatment. For the contacted people who logged in, people in the control condition were asked political questions in a standard way. Of the 1,805 people assigned to control, 693 did not respond to Ipsos at all, 4 responded to Ipsos but did not answer the turnout question, and 1,108 responded to the turnout question. The response rate for those randomly assigned to control was 61%.

People in the treatment group who logged in were given a chance to opt out of the survey by asking them if they would like to discuss politics, sports, movies, or health. Of the 1,768 people assigned to

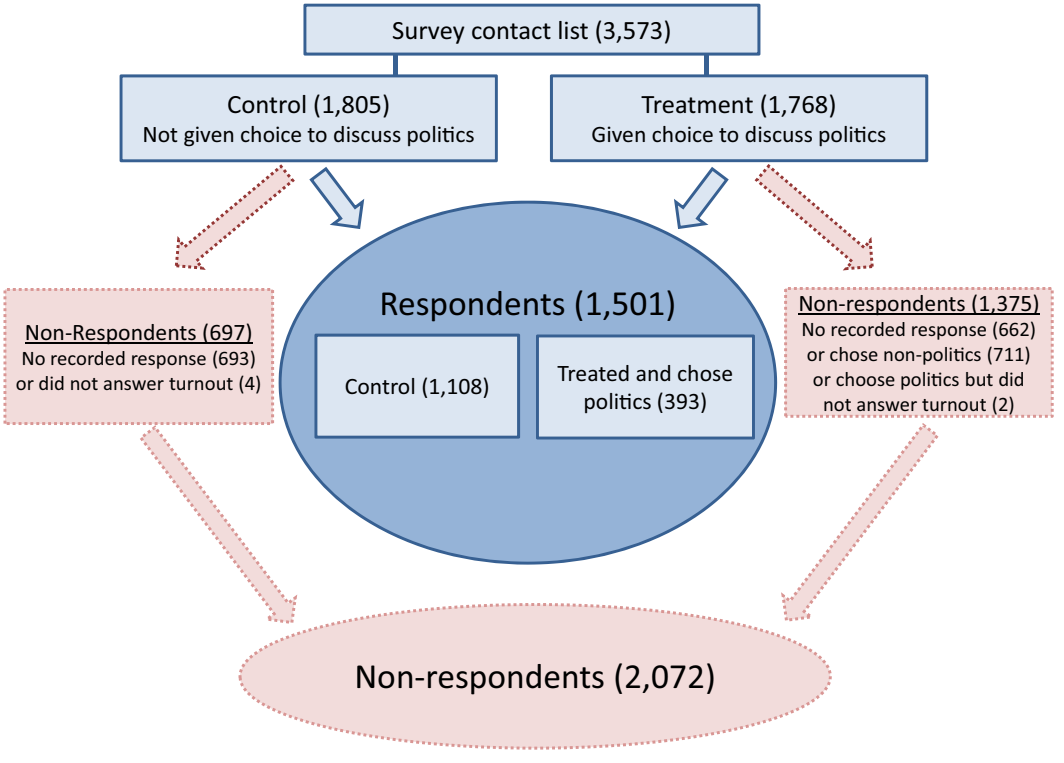

Figure 2. Survey design for turnout analysis.

treatment, 662 did not respond to Ipsos at all (and therefore presumably did not see the treatment), 711 selected a topic other than politics, 393 chose politics and answered the turnout question, and 2 people chose politics but did not answer the turnout question. Those who chose politics are respondents, and the others are non-respondents. The response rate for those assigned treatment was 22%, indicating that the treatment reduced response rates by 39 percentage points. The resulting data have 1,501 respondents and 2,072 non-respondents. Bailey (2024) analyzes these data with other MNAR protocols, all producing the same general conclusions as described below; I leave modeling differences across the types of nonresponse for future work. Table A16 in the Supplementary Material provides descriptive statistics for the survey.

5.1. Turnout

Surveys of voting typically overestimate turnout, likely in part due to non-ignorable nonresponse (Jackman and Spahn 2019). Therefore, assessing the performance of the models on turnout questions provides a good starting point for assessing how the models work.

Figure 4 shows results for responses to the question "How likely is it you will vote in the elections in November?" Response options were on a five-point scale ranging from will not vote to absolutely certain. To keep the analysis consistent with the presentation here that is based on a dichotomous Y, I have coded "absolutely certain" as Turnout = 1 and all other answers as 0.

The figure shows estimates based on the observed sample, conventional MAR-based weighting and the MNAR models. The conventional MAR-based weights are provided by Ipsos; the results are similar if I use MAR-based IPW weights based on gender, race, education, and age.

The MNAR weighting model is estimated analogously to Equation (5). In the MNAR imputation model, I estimate $E[Y|Z,X,R=0]$ as a function of all covariates and Z in the observed data using Equation (6). This equation includes Z even though Z does not directly affect Y in the full population; the point is that conditional on response, the value of Y for R = 0 individuals can depend on Z. I include Z and all X as well as all interactions of Z with the X variables.

The MNAR models include age and dichotomous variables for the response instrument, female, Black, Hispanic, and education categories. Supplementary Table A2 provides full results.

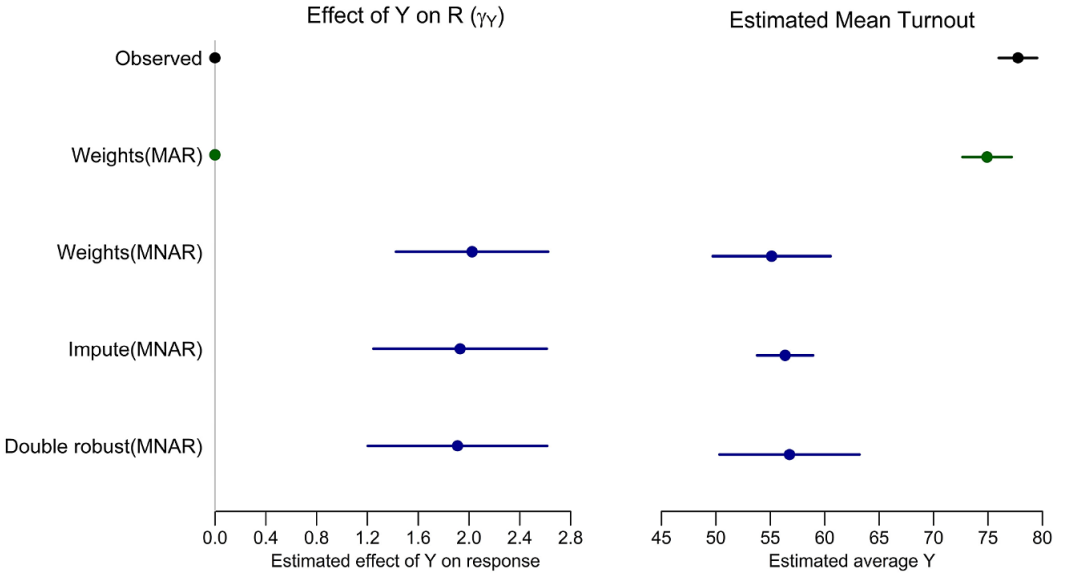

Figure 3. Analysis of turnout question.

The left panel of Figure 3 shows that the MNAR estimates of γ_Y are far from zero. The lines around the point estimates span 95% confidence intervals. These results suggest that people who are likely to turn out are also likely to respond, consistent with the selection bias suspected of existing on turnout questions. The right panel of Figure 3 shows the estimated mean of the turnout variable. Because responses to this question are on a five-point scale, I cannot directly map answers onto actual turnout (which was months away, as well). We can, however, note the pattern in turnout sentiment as measured across multiple models. In the observed sample, 77% said they were "absolutely certain" to vote. Conventional MAR-based weighting lowered this modestly to 75%. All of the MNAR models produced lower estimates of around 57%, which is closer to the typical turnout in midterms. Full results are available in Supplementary Table A2; I also discuss similar results that emerge with a different coding of the dependent variable.

The confidence intervals on the estimated mean turnout are larger with MNAR models—and this is with an observed sample of over 1,500 people and a strong first-stage response instrument. With a weak instrument and/or smaller sample size, the confidence intervals will get even larger.

5.2. Trump Approval

Polling related to Donald Trump has challenged pollsters since 2016. In 2016, the national polls were roughly correct but polls in key states, especially in the Midwest, tended to understate support for Trump. In 2020, all polls tended to understate Trump support (Clinton *et al.* 2021).

This survey asked respondents whether they approved of how Trump was handling his job as president, with response categories ranging from strongly disapprove to strongly approve plus a not sure category. I coded strongly and somewhat approve answers as approving Trump and coded not sure answers as missing.

The left panel of Figure 4 displays the estimated MNAR doubly robust γ_Y coefficient for Trump approval for two estimates. The lines indicate 95% confidence intervals. The right panel of the figure shows three different estimates of average support for Trump. The lightest shade for each population group displays the observed approval and the middle bar for each group displays the estimated approval for the group as estimated by a conventional MAR-type weighting. The darker bar for each group displays Trump approval as estimated by the doubly robust MNAR model. Full results are in Supplementary Tables A3 and A4.

The left panel of Figure 4 shows that γ_Y is close to zero when we use all observations, suggesting that there was no non-ignorable nonresponse. The weighting and doubly robust population estimates in the right panel are therefore similar to each other.

As noted earlier, the nature of non-ignorability may vary by subgroups. A direct way to assess such heterogeneity is to subset the data. Figure 4 also displays results for white people in the Midwest.[4] Many assessments of the 2016 and 2020 polling errors noted that polling in the Midwest seemed particularly difficult with Trump on the ballot. In Wisconsin in 2020, for example, Biden only beat Trump by 0.6 % even though most polls suggested Biden would win easily. One late October *Washington Post* poll even showed Biden up 17% after weighting.

The estimate of γ_Y for the Midwest is −1.34 with a confidence interval that does not include zero. A negative γ_Y coefficient is consistent with non-ignorable nonresponse: people who disliked Trump were more eager to respond and therefore the observed sample was skewed against Trump, even when weighted with conventional weights.

The effect of non-ignorable nonresponse was substantial. The right panel of Figure 4 shows that for white Midwesterners, Trump support was 45% in the observed sample and 43% in the con-

[4]Parameters in these models cannot be estimated if they perfectly predict the outcome (something that is familiar for those who work with discrete choice maximum likelihood models). Hence, I omit groups who are perfectly predictable; for example, the models for the Midwest do not include African Americans because all 18 of the African American Midwesterners who responded did not approve of Trump.

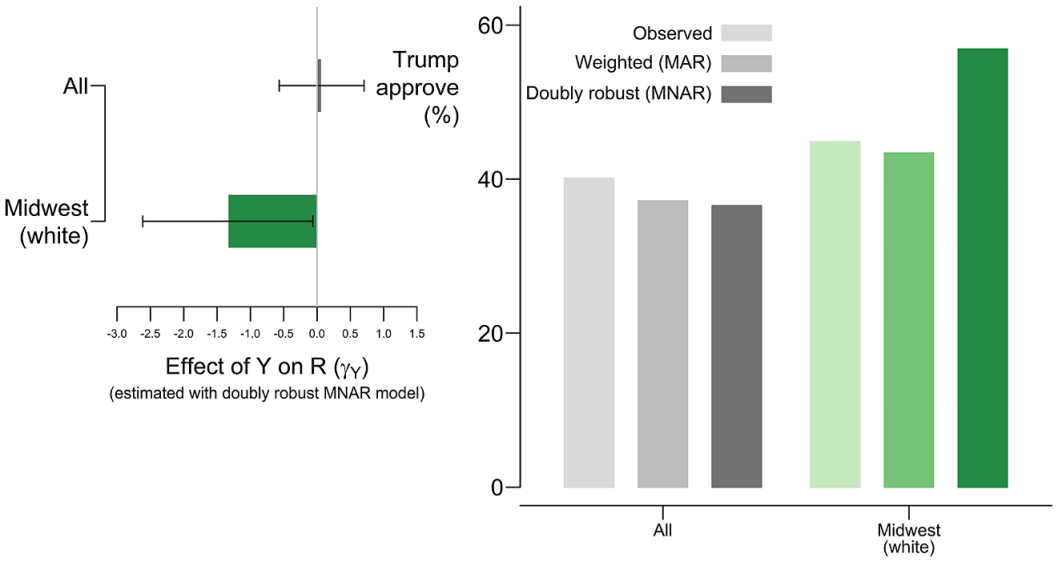

Figure 4. Analysis of Trump approval.

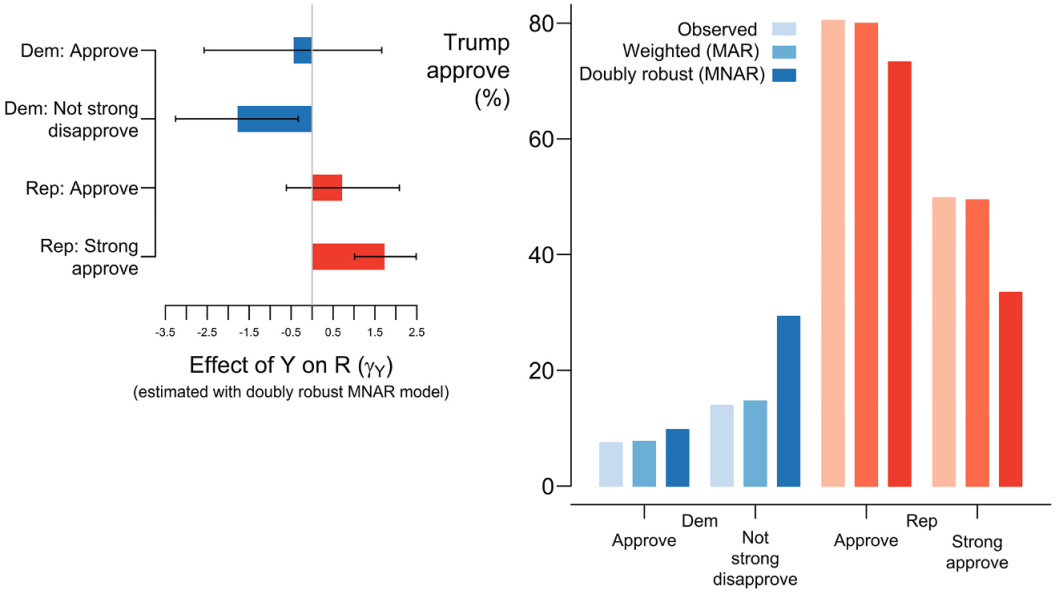

Figure 5. Analysis of Trump approval by party.

ventionally MAR-weighted sample. In the doubly robust MNAR estimate, Trump support among white Midwesterners was 58%. In other words, the analysis suggests that Trump approval among white Midwesterners was more than 10 percentage points higher when accounting for non-ignorable nonresponse. Differences of such magnitudes were not evident in other regions.

The effects of non-ignorable nonresponse went in different directions for the two parties. Figure 5 shows results for white Democrats and Republicans. The top line in the panel on the left shows that among Democrats there was little evidence of non-ignorable nonresponse with respect to Trump approval. Only 7% of Democrats in the raw data approved of Trump, however, meaning that the statistical power was quite low (as is typical for low probability outcomes). I also present a model labeled

"not strong approve" which is 0 for Democrats who strongly disapprove of Trump and 1 for all other Democrats who answered the question (which includes those who disapproved, but not strongly). For this model, the γ_Y is negative (with a confidence interval that does not include zero) and the estimated proportions are different. In the observed data (weighted or not), 13% of Democrats did not strongly disapprove of Trump, while this number rose 35% in the doubly robust MNAR estimate.

The opposite pattern occurred for Republicans. The estimate of γ_Y was not statistically significant when distinguishing between approval and disapproval. However, the estimate of γ_Y was positive and significant in a model in which Republicans who strongly approved of Trump were coded as 1 and all other Republicans who answered the question were coded as 0 (which includes those who approved, but not strongly in addition to those who disapproved). For strong approval, the observed and weighted estimates indicated that about 49% of Republicans strongly approved of Trump, while the MNAR doubly robust estimate was that 34% of Republicans strongly approved of Trump. Full results are available in Supplementary Tables A5–A8.

This analysis prompts the question as to how best to identify subgroups. In principle, one could break the sample into cells (as in cell weighting) and estimate separate models for each cell, after which one could patch together a population estimate based on population proportions across the cells. There would likely be statistical power issues, however. My approach was to assess substantively important subgroups with a two-step process. In an earlier survey, I explored subgroups based on party and region in light of the evaluations of recent polling lapses. I found variation by region and party. The second step was to assess these subgroups in the subsequent survey reported here.

In general, the IPW, imputation and doubly robust estimates were similar. For Midwesterners, for example, the γ_Y estimates for the IPW, imputation and doubly robust estimators were −1.46, −1.41, and −1.34, respectively.

5.3. Trump Tax Cuts

The non-ignorable nonresponse tends to vary by party on policy questions. Republicans who were eager to respond tended to be more conservative on policy questions than Republicans who were less eager to respond; Democrats who were eager to respond tended to be more liberal than other Democrats. Here, I illustrate the phenomenon for a question about support for the "Republican tax cut of 2017."

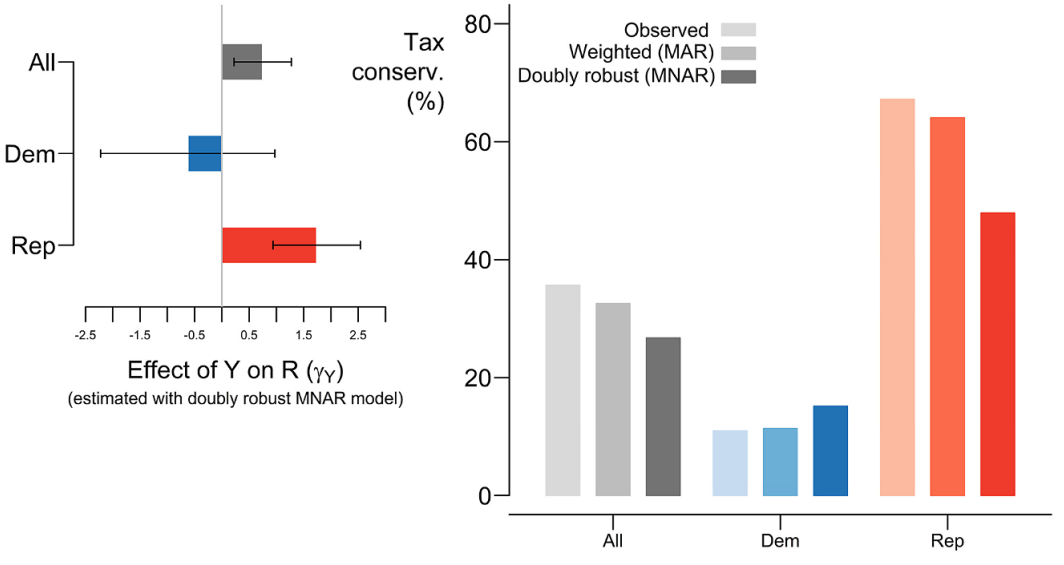

Figure 6. Analysis of tax cuts question.

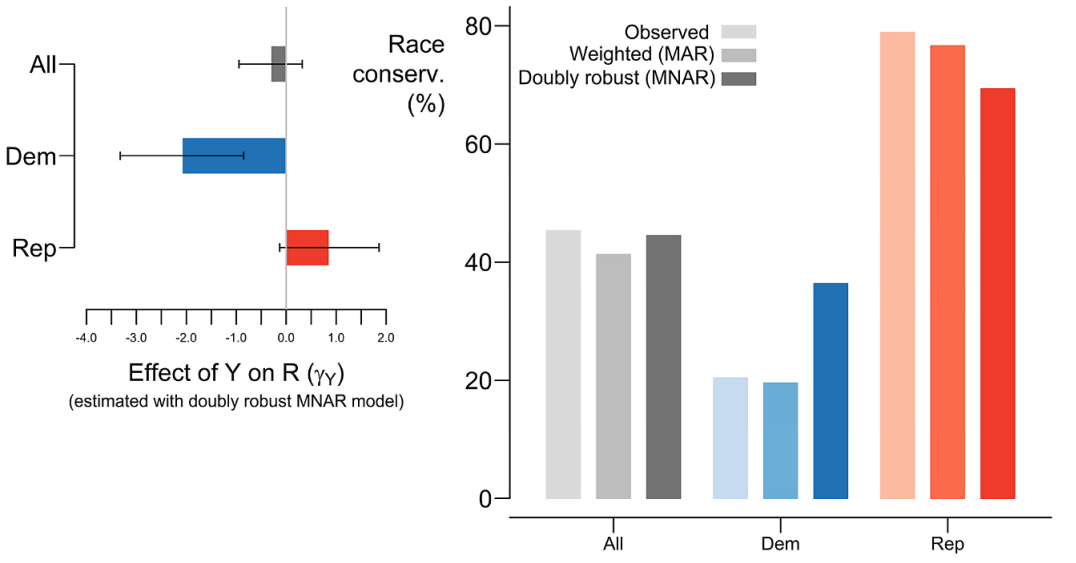

Figure 7. Analysis of race question (high values are more conservative).

Figure 6 shows the results for the whole sample and by party. Nonresponse in the whole sample modestly skews the sample toward those who favored the cuts. The party-specific results suggest that this pattern is driven by Republicans. For Democrats, γ_Y is close to zero and the estimated averages of support for the tax cuts are similar for conventionally weighted and MNAR-based doubly robust models. Only 11% of Democrats supported the tax cuts, however, meaning that the analysis was low-powered. Table A11 in the Supplementary Material shows that the γ_Y parameter was statistically significant when the dependent variable was 1 for the 38% of Democrats who did not oppose the tax (which includes those who "neither approved nor disapproved").

For Republicans, the estimate of γ_Y was positive and statistically significant. While the weighted estimate of Republican support for the tax cuts is 64%, the doubly robust estimate for support for the tax cuts among Republicans is 48%. A similar pattern occurred in a question about tariffs (which I omit for reasons of space). Full results are available in Supplementary Tables A9–A12.

5.4. Racial Conservatism

On questions related to race, some people may avoid expressing conservative views that they consider to be socially undesirable (Berinksy 1999). The survey asked a battery of race-related questions and here we illustrate how the models described perform on race using a question asking whether the respondents agreed with the statement that "Black athletes should not take a knee during the national anthem." Support for the conservative position is coded as 1 and support for the liberal position is coded as 0.

Figure 7 shows there is little sign of non-ignorable nonresponse and the full population. For Democrats, however, there is strong evidence of non-ignorable nonresponse. The estimated means for Democrats differ substantially: conventional weighting suggests 20% of Democrats opposed kneeling during the national anthem while doubly robust estimation suggests 36% of Democrats supported the racially conservative position. For Republicans, conventional weighting suggests 79% of Republicans supported the racially conservative position while doubly robust estimation suggests 71% of Republicans supported the racially conservative position. Full results are available in Supplementary Tables A13–A15.

6. Conclusion

Survey research faces huge challenges in the modern polling environment. While MAR-based approaches such as weighting are widely used and understood, less is known about models that allow for non-ignorable nonresponse.

The purpose of this paper is to introduce methods from other literatures that also grapple with non-ignorable nonresponse. In doing so, I show how the orthogonality of randomized response instruments enables the estimation of weighting and imputation models that are infeasible with the MAR-type approaches. The MNAR models do require assumptions, including an assumption that the randomized treatment does not interact with Y. Future work can extend the analysis to more general assumptions in the spirit of Levis, Kennedy, and Keele (2024).

I apply these tools to political examples, highlighting two important findings. First, the MNAR models produce results that accord with expectations even as they differ from MAR-based models. Second, the MNAR models produce interesting substantive findings. Based on these models, support for Donald Trump among white Midwesterners was higher than indicated by conventional weighting. In addition, the MNAR results suggested that surveys may exaggerate partisan differences, a point consistent with Cavari and Freedman (2023).

Acknowledgments. I appreciate incisive feedback from reviewers, discussants, and panel participants at the Annual Meeting of the Society for Political Methodology and the American Political Science Association.

Competing Interest. The authors have no competing interest to declare.

Data Availability Statement. The code and data to replicate the analysis described in the paper and the Supplementary Material will be available at the Harvard Dataverse.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.13.

References

Bailey, M. A. 2024. *Polling at a Crossroads: Rethinking Modern Survey Research*. Cambridge: Cambridge University Press.

Berinksy, A. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43 (4): 1209–1230.

Bradley, V. C., S. Kuriwaki, M. Isakov, D. Sejdinovic, X. Meng, and S. Flaxman. 2021. "Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake." *Nature* 600: 695–700.

Burger, R., and Z. McLaren. 2017. "An Econometric Method for Estimating Population Parameters from Non-Random Samples: An Application to Clinical Case Finding." *Health Economics* 26 (9): 1110–1122.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics*. Cambridge: Cambridge University Press.

Cavari, A., and G. Freedman. 2023. "Survey Nonresponse and Mass Polarization: The Consequences of Declining Contact and Cooperation Rates." *American Political Science Review* 117 (1): 332–339.

Chen, Y., P. Li, and C. Wu. 2020. "Doubly Robust Inference with Nonprobability Survey Samples." *Journal of the American Statistical Association* 115 (532): 2011–2021.

Clinton, J., et al. 2021. "Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls." American Association for Public Opinion Research.

Cohn, N. 2022. "Will One Small Shift Fix the Polls in 2022?" *New York Times*, November 7.

Coppock, A., A. Gerber, D. P. Green, and H. Kern. 2017. "Combining Double Sampling and Bounds to Address Nonignorable Missing Outcomes in Randomized Experiments." *Political Analysis* 25 (2): 188–206.

Hartman, E., and M. Huang. 2023. "Sensitivity Analysis for Survey Weights." *Political Analysis* 32: 1–16.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 153–162.

Isakov, M., and S. Kuriwaki. 2020. "Towards Principled Unskewing: Viewing 2020 Election Polls through a Corrective Lens from 2016." *Harvard Data Science Review* 2(4). https://doi.org/10.1162/99608f92.86a46f38

Jackman, S., and B. Spahn. 2019. "Why Does the American National Election Study Overestimate Voter Turnout?" *Political Analysis* 27: 193–207.

Kennedy, C., D. Popky, and S. Keeter. 2023. "How Public Polling Has Changed in the 21st Century." *Pew Research Center.*

Levis, A., E. Kennedy, and L. Keele. 2024. "Nonparametric Identification and Efficient Estimation of Causal Effects with Instrumental Variables." Preprint, arXiv:2402.09332.

Little, R., and D. Rubin. 2020. *Statistical Analysis with Missing Data*. New York: Wiley.

Lusinchi, D. 2012. "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" *Social Science History* 36 (1): 23–54.
Marra, G., R. Radice, T. Barnighausen, S. Wood, and M. McGovern. 2017. "A Simultaneous Equation Approach to Estimating HIV Prevalence with Nonignorable Missing Responses." *Journal of the American Statistical Association* 112 (518): 484–496.
Meng, X. 2018. "Statistical Paradises and Paradoxes in Big Data (1): Law of Large Populations, Big Data Paradox, and the 2016 Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726.
Peress, M. 2010. "Correcting for Survey Nonresponse Using Variable Response Propensity." *Journal of the American Statistical Association* 105 (492): 1418–1430.
Scharfstein, D., A. Rotnitzky, and J. Robins. 1999. "Adjusting for Nonignorable Dropout Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association* 94: 1096–1146.
Sun, B., L. Liu, W. Miao, K. Wirth, J. Robins, and E. Tchetgen-Tchetgen. 2018. "Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable." *Statistica Sinica* 28: 1965–1983.
Tchetgen, E. T., and K. Wirth. 2017. "A General Instrumental Variable Framework for Regression Analysis with Outcome Missing Not at Random." *Biometrics* 73 (4): 1123–1131.
Tchetgen-Tchetgen, E., J. Robins, and A. Rotnitzky. 2010. "On Doubly Robust Estimation in a Semiparametric Odds Ratio Model." *Biometrika* 97 (1): 171–180.
Vermeulen, K., and S. Vansteelandt. 2016. "Data-Adaptive Bias-Reduced Doubly Robust Estimation." *International Journal of Biostatistics* 12 (1): 253–282.

Cite this article: Bailey, M. A. (2025). Countering Non-Ignorable Nonresponse in Survey Models with Randomized Response Instruments and Doubly Robust Estimation. *Political Analysis*, 140–155. https://doi.org/10.1017/pan.2024.13