Political Analysis (2014) pp. 1–21
doi:10.1093/pan/mpu009

Semi-parametric Selection Models for Potentially Non-ignorable Attrition in Panel Studies with Refreshment Samples

Yajuan Si
Department of Statistics, MC 4690, Columbia University, NY, NY 10027, USA
e-mail: ysi@stat.columbia.edu (corresponding author)

Jerome P. Reiter
Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708, USA

D. Sunshine Hillygus
Department of Political Science, Box 90204, Duke University, Durham, NC 27708, USA

Edited by R. Michael Alvarez

Panel studies typically suffer from attrition. Ignoring the attrition can result in biased inferences if the missing data are systematically related to outcomes of interest. Unfortunately, panel data alone cannot inform the extent of bias due to attrition. Many panel studies also include refreshment samples, which are data collected from a random sample of new individuals during the later waves of the panel. Refreshment samples offer information that can be utilized to correct for biases induced by non-ignorable attrition while reducing reliance on strong assumptions about the attrition process. We present a Bayesian approach to handle attrition in two-wave panels with one refreshment sample and many categorical survey variables. The approach includes (1) an additive non-ignorable selection model for the attrition process; and (2) a Dirichlet process mixture of multinomial distributions for the categorical survey variables. We present Markov chain Monte Carlo algorithms for sampling from the posterior distribution of model parameters and missing data. We apply the model to correct attrition bias in an analysis of data from the 2007–08 Associated Press/Yahoo News election panel study.

1 Introduction

The value of longitudinal or panel surveys, in which the same individuals are interviewed repeatedly at different points in time, is well recognized in political science, as are the threats from panel attrition. For example, in the most recent multi-wave panel survey of the American National Election Study (ANES), 47% of respondents who completed the first wave in January 2008 failed to complete the follow-up wave in June 2010. Such attrition can result in biased inferences when the propensity to drop out is systematically related to the substantive outcome of interest (e.g., Behr, Bellgardt, and Rendtel 2005; Olsen 2005; Bhattacharya 2008a). Too often, however, diagnostic analyses of panel attrition are conducted and reported separately from substantive research (e.g., Zabel 1998; Bartels 1999; Clinton 2001; Kruse et al. 2009). Yet, panel attrition is not just a distinct technical issue of interest only to methodologists; it can have direct implications for the substantive knowledge claims that can be made from panel surveys. For example, Bartels (1999) found that estimates of political interest were too high in the second wave of the 1992–96

Authors' note: Replication materials are available in Si et al. (2014).

ANES panel because of differential dropout among respondents. Frankel and Hillygus (2013) show that attrition in the 2008 ANES panel study biased estimates of the relationship between sex and campaign interest.

Although there is considerable recognition about the potential problems of panel attrition, the way in which it is handled in panel survey research varies widely. The most commonly used approaches for handling panel attrition rely on tenuous assumptions about the attrition process. Perhaps most often, scholars simply ignore panel attrition entirely in their reliance on listwise deletion for all missing data (e.g., Wawro 2002)—an approach that assumes data are missing completely at random (MCAR). The use of post-stratification weights (e.g., Henderson, Hillygus, and Tompson 2010) or standard approaches to multiple imputations (e.g., Pasek et al. 2009; Honaker and King 2010) assumes the data are missing at random (MAR)—dependent on observed, but not unobserved, data. Recognizing that MCAR or MAR assumptions may not be justifiable, some researchers use selection models (Hausman and Wise 1979; Brehm 1993; Kenward 1998; Scharfstein, Rotnitzky, and Robins 1999) or pattern mixture models (Little 1993; Kenward, Molenberghs, and Thijs 2003) to model attrition that is not missing at random (NMAR)—dependent on the values of unobserved data. Unfortunately, the parameters in such models cannot be identified by the panel data alone; strong and untestable assumptions about the attrition process are necessary (e.g., Schluchte 1982; Brown 1990; Diggle and Kenward 1994; Little and Wang 1996; Hogan and Daniels 2008).

The key dilemma is that it is not possible to determine if the missing-data pattern from panel attrition is ignorable or non-ignorable using only the collected data. External sources of information are necessary. When looking for external data, it is common to rely on government studies like the Current Population Survey (CPS) or the American Community Survey. For example, in the construction of post-stratification weights, the characteristics of those who answered all waves of the panel (complete cases) are compared to estimates from these "gold standard" surveys. These sources, however, severely restrict the variables available for comparison to a handful of demographic characteristics. Frankel and Hillygus (2013) show that demographic variables alone may not be sufficient to fully account for panel attrition bias. Moreover, any observed differences could reflect differences in survey design features like mode, field period, or question wording, rather than bias due to panel attrition.

Refreshment samples—cross-sectional, random samples of new respondents given the questionnaire at the same time as a second or subsequent wave of the panel—are a preferred source of external data. These samples offer information that can be leveraged to correct for non-ignorable panel attrition via statistical modeling. In particular, analysts can correct bias via an additive non-ignorable (AN) model, which comprises a joint model for the survey variables coupled with a selection model for the attrition process (Hirano et al. 1998).

To date, applications of the AN model have been limited to low dimensional settings, for example only a handful of survey variables. With low dimension, specifying the joint model for the survey variables in the AN model is relatively straightforward; see Deng et al. (2013) for an overview. However, panel analyses in political science often involve a substantial number of variables, so specifying a joint model for the survey variables can be daunting. Consider, for example, specifying a joint model for a large number of categorical variables. When using log-linear models or sequences of conditional models (e.g., specify $f(a)$, then $f(b|a)$, then $f(c|a,b)$, and so on), the number of possible models is enormous. It can be difficult to identify the interaction terms to include in the model, especially in the presence of missing data (Erosheva, Fienberg, and Junker 2002; Vermunt et al. 2008; Su et al. 2011; Si and Reiter 2013). Approaches that treat categorical variables as continuous—for example the multivariate normal assumption of `Amelia II` (King et al. 2001)—can have undesirable properties even in low dimensions (Allison 2000; Cranmer and Gill 2013; Kropko et al. 2014).

In this article, we propose a variant of the AN model for two-wave panels with many categorical survey variables. The underlying joint distribution for the survey variables is a Dirichlet process (DP) mixture of multinomial distributions, and the model for attrition is a probit AN selection model. The DP mixture model can capture complex dependencies among the survey variables automatically while being computationally efficient (Dunson and Xing 2009; Si and

Reiter 2013).[1] The Markov chain Monte Carlo (MCMC) algorithms for fitting the DP model can be easily modified to handle both attrition and item nonresponse simultaneously. The MCMC also produces completed data sets that, if desired, can be used for multiple imputation (Rubin 1987). We call the proposed approach the *semi-parametric AN model*.

The remainder of the article is structured as follows. We first offer a brief background on the use of refreshment samples in panel studies in Section 2. We then describe the semi-parametric AN model in detail in Section 3 and the MCMC algorithms in Section 4. We present results of a simulation study of the performance of the method in Section 5. We apply the semi-parametric AN model to data from the Associated Press-Yahoo 2008 Election Panel Study (APYN) in Section 6, focusing on measuring political interest in the 2008 presidential campaign.[2] We find evidence that panel attrition is non-ignorable, and correcting for it results in different estimates of political interest in the population than complete-case analysis. The panel attrition bias is especially pronounced among some subsets of the population, especially white, partisan females and white, Republican males. We also describe approaches for model diagnostics in Section 6. Finally, we summarize the article and discuss future research directions in Section 7.

2 Background

Although panel surveys offer considerable advantages over cross-sectional surveys for modeling complex relationships, they are typically more complex logistically (and more costly) because they must find and interview respondents who had participated in the first wave of the survey to maintain representativeness. This has become an increasingly challenging task in recent years because people are harder to reach and, if reached, more likely to refuse participation (Olson and Witt 2011). Thus, panel surveys tend to have more severe missing-data problems than cross-sectional surveys. Not only do participants fail to respond to individual survey questions (item nonresponse), they also fail to respond to entire follow-up survey waves (panel attrition).

Refreshment samples have become a common design feature of panel surveys as a way to check or mitigate against panel attrition. Overlapping or rotating panels, in which a new study cohort completes their first wave at the same time a previous cohort completes a second or later wave, offer an equivalent benefit. The General Social Survey, CPS, and the 2008–09 ANES Panel are just a few examples of major studies that have included refreshment samples.

Although refreshment samples are relatively common, the way they are used varies widely. Most often, refreshment samples are used for basic exploratory checks and comparisons—as the basis for discussion about potential bias in the results without statistical correction (e.g., Frick et al. 2006; Kruse et al. 2009). As initially conceived, refreshment samples were used, as the name implies, to "refresh" the panel, using the fresh respondents to directly replace those who had dropped out (Ridder 1992)—an idea that dates to some of the earliest survey methods work (Kish and Hess 1959). Others simply add the refreshment respondents to the analysis to boost the sample size, disregarding the attrition process of the original respondents (e.g., Wissen and Meurs 1989; Heeringa 1997; Thompson et al. 2006). Unfortunately, considerable research has shown that simply treating these refreshment respondents as substitutes without additional adjustments can introduce additional bias because they are more likely to resemble respondents rather than nonrespondents (Lin and Schaeffer 1995; Vehovar 1999).

The key point is simply that refreshment samples have been under-utilized in analyses of panel surveys. This is despite the fact that methods have been developed to use the information in refreshment samples to statistically correct for non-ignorable panel attrition (Hirano et al. 1998, 2001; Bartels 1999; Bhattacharya 2008b). In particular, Hirano et al. (1998, 2001) prove that an AN selection model for attrition can be identified and can correct for attrition bias in two-wave surveys with refreshment samples at the second wave. Deng et al. (2013) extend the AN model to panels with more than two waves and one refreshment sample, including scenarios where attriters in one

[1] Similar non-parametric Bayesian approaches have been used by political scientists for large-scale data analysis in other contexts, for example Grimmer (2010) and Kyung, Gill, and Casella (2011).
[2] Replication materials for Sections 5 and 6 are available in Si, Reiter, and Hillygus (2014).

wave return in a later wave. Unfortunately, however, these methods have not filtered down to substantive applications in political science, perhaps because scholars must often contend with additional data complexities not dealt with in these previous applications of the AN approach—namely, item nonresponse and high-dimensional settings. Thus, we build on the previous research by proposing a joint model that offers a more flexible imputation engine to handle a large number of categorical survey variables and is able to efficiently handle item nonresponse. The proposed semi-parametric AN model uses a DP mixture model to estimate the underlying joint distribution of the quantities of interest rather than a succession of logistic or multinomial models—an approach that is challenging (or unfeasible) in applications with a large number of variables. In addressing the real-world complexities of survey data, we hope to help make the testing and correction of non-ignorable panel attrition more practical for political science analyses of panel surveys.

In the sections that follow, we introduce an AN selection model for non-ignorable panel attrition with refreshment samples for large-scale categorical data. All variables used here are either binary or nominal, but similar approaches could be used to handle mixed data types of continuous and ordered categorical variables.

3 Methods

Consider a two-wave panel of N_p individuals with a refreshment sample of N_r new subjects in the second wave. For all $N = N_p + N_r$ subjects, the data include q time-invariant variables $X = (X_1, \ldots, X_q)$, such as demographic or frame variables. In the first wave, p_1 response variables are of interest, $Y_1 = (Y_{11}, \ldots, Y_{1p_1})$. The corresponding p_2 response variables in wave 2 are $Y_2 = (Y_{21}, \ldots, Y_{2p_2})$. Here, we assume $p_2 = p_1$, although this is not necessary. Among the N_p individuals, $N_{cp} < N_p$ provide at least some data in the second wave, and the remaining $N_{ip} = N_p - N_{cp}$ individuals drop out of the panel. The refreshment sample includes only (X, Y_2); by design, Y_1 are missing for all the individuals in the refreshment sample. We presume that X, Y_1, and Y_2 may be subject to item nonresponse. Thus, an individual in the panel with item nonresponse in wave 2 provides values of Y_{2k} for some non-empty set of k and does not provide values for the complementary set. Let $p = q + p_1 + p_2$ be the total number of variables, and let $Z = (Z_1, \ldots, Z_p) = (X_1, \ldots, X_q, Y_{11}, \ldots, Y_{1p_1}, Y_{21}, \ldots, Y_{2p_2})$ comprise all survey variables.

For each individual $i = 1, \ldots, N$, let $W_i = 1$ if individual i would remain in wave 2 if included in wave 1, and let $W_i = 0$ if individual i would drop out of wave 2 if included in wave 1. We note that W_i is fully observed for all individuals in the panel but is missing for the individuals in the refreshment sample, since this latter set is not provided the chance to respond in wave 1. Table 1 offers a graphical representation of the data structure of the panel and refreshment samples.

With this setup, we have mixed types of missingness due to attrition, item nonresponse, and survey design as follows: W in the refreshment sample by design; Y_1 in the refreshment sample by design; Y_2 for the N_{ip} individuals in the panel due to attrition; Y_1 in the panel due to item nonresponse; Y_2 for the N_{cp} individuals in the panel due to item nonresponse; Y_2 in the refreshment sample due to item nonresponse; X in the panel due to item nonresponse; and X in the refreshment sample due to item nonresponse.

To handle both attrition and item nonresponse in inference, we require a joint model for Z and W. We specify this in two steps, using

$$W|Z, \beta \sim g(X, Y_1, Y_2, \beta) \tag{1}$$

$$Z|\Theta \sim f(\Theta), \tag{2}$$

where β and Θ represent sets of model parameters.

For the selection model in (1), we use the probit regression

$$\Phi^{-1}(\Pr(W = 1|Y_1, Y_2, X)) = \beta_0 + \vec{\beta}_X X + \vec{\beta}_{Y_1} Y_1 + \vec{\beta}_{Y_2} Y_2. \tag{3}$$

Here, we implicitly assume that the researcher uses dummy coding to represent the levels of each variable and a separate regression coefficient for each dummy variable. Crucially, this model

Table 1 Missing data structure of panel and refreshment samples

	Wave 1	*Wave 2*
Panel sample	$X,\ Y_1$	$Y_2,\ W=1$ $Y_2=?,\ W=0$
Refreshment sample	$Y_1=?$	$X,\ Y_2,\ W=?$

assumes the probability of attrition depends on Y_1 and Y_2 through an additive function; that is, it excludes interactions between Y_1 and Y_2.[3] This quasi-separability assumption is necessary to enable identification of the model parameters, as there is insufficient information to estimate interaction effects between Y_1 and Y_2 in equation (3). We note that the model also can accommodate interaction terms among subsets of (X, Y_1) and interaction terms among subsets of (X, Y_2). Other proper link functions for binary response, such as logit, can be used in place of the probit. As a diffuse prior distribution, we use $\beta \sim N(0, \Sigma_0)$ with large variances.

As described by Hirano et al. (1998), AN models include MAR and NMAR models as special cases. When $\vec{\beta}_{Y_2} = \vec{0}$ and at least one element among $\{\vec{\beta}_X, \vec{\beta}_{Y_1}\}$ does not equal 0, the attrition is MAR. When at least one element among $\{\vec{\beta}_{Y_2}\}$ does not equal 0, the attrition is NMAR. Hence, the AN model allows the data to decide between MAR and (certain types of) NMAR attrition mechanisms.

For the joint distribution of the survey variables in (2), we propose using the Dirichlet process mixture of products of multinomial distributions (DPMPM) originally developed by Dunson and Xing (2009) and used for multiple imputation of missing data by Si and Reiter (2013). We use an approximation for computation in the DP model based on a finite number of mixture components. Without loss of generality, assume that each Z_j takes on values in $\{1, \ldots, d_j\}$, where $d_j \geq 2$ is the total number of categories for variable j. The survey variables form a contingency table of $d_1 \times d_2 \times \ldots \times d_p$ cells defined by cross-classifications of Z. Let $c_j \in \{1, \ldots, d_j\}$ be a particular value of Z_j. For any cell comprising feasible combinations of c_j, we define the cell probability as $\theta_{c_1,\ldots,c_p} = \Pr(Z_1 = c_1, \ldots, Z_p = c_p)$. Finally, for $i = 1, \ldots, N$ and $j = 1, \ldots, p$, let Z_{ij} be the value of Z_j for individual i.

Paraphrasing from Si and Reiter (2013), the truncated DPMPM assumes that each individual i belongs to exactly one of $H < \infty$ latent classes. We discuss an approach for determining H at the end of this section. For $i = 1, \ldots, N$, let $s_i \in \{1, \ldots, H\}$ indicate the class of individual i, and let $\pi_h = \Pr(s_i = h)$. We assume that $\pi = (\pi_1, \ldots, \pi_H)$ is the same for all individuals. Within any class, each of the p variables independently follows a class-specific multinomial distribution, so that individuals in the same latent class have the same cell probabilities. For any c_j, let $\psi^{(j)}_{hc_j} = \Pr(Z_{ij} = c_j | s_i = h)$ be the probability of $Z_{ij} = c_j$ given that individual i is in class h. Let $\psi = \{\psi^{(j)}_{hc_j} : c_j = 1, \ldots, d_j, j = 1, \ldots, p, h = 1, \ldots, H\}$ be the collection of all $\psi^{(j)}_{hc_j}$. The finite mixture model can be expressed as

$$Z_{ij} | s_i, \psi \overset{\text{ind}}{\sim} \text{Multinomial}(\psi^{(j)}_{s_i 1}, \ldots, \psi^{(j)}_{s_i d_j}) \text{ for all } i, j \tag{4}$$

$$s_i | \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_H) \text{ for all } i, \tag{5}$$

where each multinomial distribution has sample size equal to 1 and the number of levels is implied by the dimension of the corresponding probability vector. Integrating out the component membership indicators, this model implies that

$$\theta_{c_1,\ldots,c_p} = \sum_{h=1}^{H} \pi_h \prod_{j=1}^{p} \psi^{(j)}_{hc_j}. \tag{6}$$

[3]The impact of the separation assumption between Y_1 and Y_2 on inferences can be evaluated by sensitivity analysis (Deng et al. 2013).

For prior distributions on π and ψ, we follow Si and Reiter (2013) and use the truncated stick-breaking representation of Sethuraman (1994). We have

$$\pi_h = V_h \prod_{l<h} (1 - V_l) \quad \text{for } h = 1, \ldots, H \tag{7}$$

$$V_h \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \quad \text{for } h = 1, \ldots, H-1, \quad V_H = 1 \tag{8}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \tag{9}$$

$$\psi_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}). \tag{10}$$

We set $a_{j1} = \ldots = a_{jd_j} = 1$ for all j to correspond to uniform distributions. Following Dunson and Xing (2009) and Si and Reiter (2013), we set $(a_\alpha = 0.25, b_\alpha = 0.25)$, which represents a small prior sample size and hence vague specification for the Gamma distribution. In practice, we find these specifications allow the data to dominate the prior distribution; see Si and Reiter (2013) for further discussion.

It is desirable to make H as large as possible. To balance against required computational time, we recommend using an iterative process to determine H. Beginning with an initial proposal for H, say $H = 20$, the researcher can examine the posterior distributions of the sampled number of unique classes across MCMC iterations to diagnose if H is large enough. Significant posterior mass at a number of classes equal to H suggests that the truncation limit be increased. We note that one can use other MCMC algorithms to estimate the posterior distribution that avoid truncation, for example a slice sampler (Walker 2007; Dunson and Xing 2009) or an exact blocked sampler (Papaspiliopoulos 2008).

4 Posterior Computations

We estimate model parameters and generate completed data sets—that is, all missing values in (X, Y_1, Y_2, W) are filled in—via a blocked Gibbs sampler (Ishwaran and James 2001). The overall strategy proceeds in steps: (1) conditional on the previous iteration of completed data and parameter draws, update the latent class indicators; (2) conditional on the previous iteration of completed data set and the latest draws of latent class memberships, update the parameters (π, ψ, β); (3) conditional on the latest draws of the parameters and latent class indicators, update the missing data in (X, Y_1, Y_2, W) due to attrition and item nonresponse. We now describe each of these steps.

Step 1: Update the latent class indicators.

Given a completed data set, for $i = 1, \ldots, N$ sample $s_i \in \{1, \ldots, H\}$ from a multinomial distribution with probabilities,

$$\Pr(s_i = h|-) = \frac{\pi_h \prod_{j=1}^{p} \psi_{hZ_{ij}}^{(j)}}{\sum_{k=1}^{H} \pi_k \prod_{j=1}^{p} \psi_{kZ_{ij}}^{(j)}}. \tag{11}$$

Step 2: Update model parameters.

We next update all model parameters using the completed data and the updated s_i values from Step 1. For the parameters in the DPMPM, we require sampling from the following distributions.

- For $h = 1, \ldots, H-1$, sample a value of V_h from the Beta distribution,

$$(V_h|-) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H} n_k), \tag{12}$$

where $n_h = \sum_{i=1}^{N} I(s_i = h)$ for all h. Here, $I(\cdot) = 1$ when the condition inside the parentheses is true and $I(\cdot) = 0$ otherwise. Set $V_H = 1$. From these H values, compute each $\pi_h = V_h \prod_{k<h} (1 - V_k)$.

- For $h = 1, \ldots, H$ and $j = 1, \ldots, p$, sample a new value of $\psi_h^{(j)} = (\psi_{h1}^{(j)}, \ldots, \psi_{hd_j}^{(j)})$ from the Dirichlet distribution,

$$(\psi_h^{(j)}|-) \sim \text{Dirichlet}(a_{j1} + \sum_{i:s_i=h} I(Z_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:s_i=h} I(Z_{ij} = d_j)). \tag{13}$$

- Sample a new value of α from the Gamma distribution,

$$(\alpha|-) \sim \text{Gamma}(a_\alpha + H - 1, b_\alpha - \log \pi_H). \tag{14}$$

For β from the AN selection model, we use latent Gaussian variables for binary probit regression models (Albert and Chib 1993) augmented with a N-dimensional latent variable t to improve mixing. Following Holmes and Held (2006), we introduce t such that

$$(\beta|t, \lambda) \sim \text{N}(B, V), \tag{15}$$

where $B = V(Z't)$ and $V = (\Sigma_0^{-1} + Z'Z)^{-1}$. Thus, given a current value of t, drawing β is straightforward. We sample a new value of t via a sequence of conditional steps.

For each $i \in \{1, \ldots, N\}$, we draw a new value of t_i from the distribution,

$$(t_i|-) \sim \begin{cases} \text{N}(m_i, v_i)I(t_i > 0) & \text{if } W_i = 1 \\ \text{N}(m_i, v_i)I(t_i \leq 0) & \text{otherwise,} \end{cases} \tag{16}$$

where $m_i = Z_iB - g_i(t_i - Z_iB)$, $v_i = 1 + g_i$, $g_i = h_i/(1 - h_i)$, where h_i are the diagonal elements of ZVZ'. After updating any particular t_i, we recompute $B = B^{\text{old}} + S_i(t_i - t_i^{\text{old}})$, where B^{old} and t_i^{old} are the values of B and t_i before the update of t_i, and S_i denotes the ith column of $S = VZ'$.

Step 3: Update missing data.

We assume that all item nonresponse is MAR; thus, we do not require modeling the item response mechanism. We use subscripts *mis* to indicate missing data, and *ref* or *pl* to indicate the refreshment sample or panel, respectively. Thus, for example, $Y_{2\text{mis}-\text{pl}}$ represents missing values of Y_2 in the panel. To update missing values in the refreshment sample, we need to sample from the full conditional distribution for $(X_{\text{mis}-\text{ref}}, Y_{1\text{mis}-\text{ref}}, Y_{2\text{mis}-\text{ref}}, W)$. To facilitate computation, we write this as

$$\begin{aligned} f(X_{\text{mis}-\text{ref}}, Y_{1\text{mis}-\text{ref}}, Y_{2\text{mis}-\text{ref}}, W|-) = {} & f(X_{\text{mis}-\text{ref}}, Y_{1\text{mis}-\text{ref}}, Y_{2\text{mis}-\text{ref}}|-) \\ & \times f(W|X_{\text{mis}-\text{ref}}, Y_{1\text{mis}-\text{ref}}, Y_{2\text{mis}-\text{ref}}, -). \end{aligned} \tag{17}$$

This allows us to take advantage of the conditional independence in the DPMPM when imputing $(X_{\text{mis}-\text{ref}}, Y_{1\text{mis}-\text{ref}}, Y_{2\text{mis}-\text{ref}})$. Thus, we have the following sampling steps:

- For $\{(i,j) : X_{ij} \in X_{\text{mis}-\text{ref}}\}$, sample a new value of X_{ij} from

$$(X_{ij}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}). \tag{18}$$

- For $\{(i,j) : Y_{i,1j} \in Y_{1\text{mis}-\text{ref}}\}$, sample a new value of $Y_{i,1j}$ from

$$(Y_{i,1j}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}). \tag{19}$$

- For $\{(i,j) : Y_{i,2j} \in Y_{2\text{mis}-\text{ref}}\}$, sample a new value of $Y_{i,2j}$ from

$$(Y_{i,2j}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{s_i1}^{(j)}, \ldots, \psi_{s_id_j}^{(j)}). \tag{20}$$

- To impute the missing W in the refreshment sample, sample from the Bernoulli distribution with success probability $\Pr(W_i = 1|-) = \Phi(Z_i\beta)$, where Z_i is the current value of the completed data for record i.

To update missing values in the panel, we sample from the full conditional distribution for $(X_{\text{mis}-\text{pl}}, Y_{1\text{mis}-\text{pl}}, Y_{2\text{mis}-\text{pl}})$. Here, we emphasize that W_i is fully observed for all units in the

panel. Hence, we must condition on the value of W_i when imputing the ith record's missing values. Once again, we leverage the conditional independence assumptions (given latent class membership) from the DPMPM. We use the following steps:

- For $\{(i,j) : X_{ij} \in X_{\text{mis-pl}}\}$, propose a new value X^*_{ij} by drawing from

$$(X_{ij}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi^{(j)}_{s_i 1}, \ldots, \psi^{(j)}_{s_i d_j}). \quad (21)$$

  The proposed new draw X^*_{ij} is accepted with probability $\Pr(W_i = 1|X^*_{ij}, -)^{(W_i=1)}\Pr(W_i = 0|X^*_{ij}, -)^{(W_i=0)}$.
- For $\{(i,j) : Y_{i,1j} \in Y_{1\text{mis-pl}}\}$, propose a new value $Y^*_{i,1j}$ by drawing from

$$(Y_{i,1j}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi^{(j)}_{s_i 1}, \ldots, \psi^{(j)}_{s_i d_j}). \quad (22)$$

  The new draw $Y^*_{i,1j}$ is accepted with probability $\Pr(W_i = 1|Y^*_{i,1j}, -)^{(W_i=1)}\Pr(W_i = 0|Y^*_{i,1j}, -)^{(W_i=0)}$.
- For $\{(i,j) : Y_{i,2j} \in Y_{2\text{mis-pl}}\}$, propose a new value $Y^*_{i,2j}$ by drawing from

$$(Y_{i,2j}|-) \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi^{(j)}_{s_i 1}, \ldots, \psi^{(j)}_{s_i d_j}). \quad (23)$$

  The new draw $Y^*_{i,2j}$ is accepted with probability $\Pr(W_i = 1|Y^*_{i,2j}, -)^{(W_i=1)}\Pr(W_i = 0|Y^*_{i,2j}, -)^{(W_i=0)}$.

5 Simulation Study

In this section, we present results of a simulation study illustrating the properties of the semi-parametric AN model. We also compare its performance with an application of multiple imputation of the missing data under an assumption of MAR attrition. To do so, we use the default imputation routines for categorical data in the popular software package, `Amelia II`.

In each replication, we generate $N_p = 800$ and $N_r = 200$ individuals in the panel and refreshment sample, respectively. Each individual has $p_1 = p_2 = 5$ binary survey variables in each wave; for simplicity we do not include X variables. We simulate values of Y_1 from a log-linear model such that

$$\begin{aligned}\log\Pr(Y_{11}, Y_{12}, Y_{13}, Y_{14}, Y_{15}) = &-2Y_{11} - 2Y_{12} - 2Y_{13} - 2Y_{14} - 2Y_{15} + Y_{11}Y_{12} + Y_{11}Y_{13} + Y_{11}Y_{14} \\ &+ Y_{11}Y_{15} + Y_{12}Y_{13} + Y_{12}Y_{14} + Y_{12}Y_{15} + Y_{13}Y_{14} - Y_{13}Y_{15} - Y_{14}Y_{15} \\ &+ Y_{11}Y_{12}Y_{13} - Y_{12}Y_{13}Y_{14} + Y_{13}Y_{14}Y_{15} + Y_{12}Y_{13}Y_{15} - 2Y_{11}Y_{14}Y_{15}.\end{aligned}$$

We simulate Y_2 from a sequence of logistic regressions such that

$$\begin{aligned}\text{logit}\ \Pr(Y_{21} = 1) &= 1.5Y_{11} + 0.2Y_{12} - 0.2Y_{13} + 0.1Y_{14} + 0.2Y_{15} - 1.2Y_{11}Y_{12} \\ \text{logit}\ \Pr(Y_{22} = 1) &= -0.1Y_{11} + Y_{12} + 0.2Y_{13} + 0.1Y_{14} + 0.1Y_{15} - Y_{21} - 0.5Y_{12}Y_{13} \\ \text{logit}\ \Pr(Y_{23} = 1) &= 1.2Y_{13} + 0.2Y_{15} + 0.1Y_{21} - 0.2Y_{22} - 0.5Y_{13}Y_{21} + 0.2Y_{21}Y_{22} + 1.1Y_{13}Y_{22} \\ &\quad - 0.4Y_{13}Y_{21}Y_{22} \\ \text{logit}\ \Pr(Y_{24} = 1) &= 0.2Y_{12} + Y_{14} + 0.1Y_{22} - 0.3Y_{23} - 1.5Y_{14}Y_{23} \\ \text{logit}\ \Pr(Y_{25} = 1) &= -0.5Y_{14} + Y_{15} + 0.2Y_{23} + 0.2Y_{24} - 1.5Y_{15}Y_{23} + Y_{23}Y_{24}.\end{aligned}$$

This simulation design ensures that pairs (Y_{1j}, Y_{2j}) are strongly correlated across waves. It also results in complex dependencies among the variables.

We introduce attrition by sampling each W_i from a Bernoulli distribution such that

$$\Pr(W = 1) = \Phi(0.5Y_{11} - 0.5Y_{12} - Y_{22} + 2.5Y_{25}). \quad (24)$$

Any record with a simulated $W_i = 0$ is treated as a case of panel attrition. We note that this attrition mechanism is non-ignorable since it depends on components of Y_2. The attrition rate across replications is generally around 30%.

During each repetition, when fitting the semi-parametric AN selection model we set $H = 20$ and $\Sigma_0 = I_{10}$. The MCMC chains converge quickly, so that we are able to use only $T = 3000$ iterations per chain. We create $m = 20$ completed data sets by selecting every fiftieth draw from the last 1000 iterations of the chain. We also create $m = 20$ completed data sets using `Amelia II`.

We repeat the process of simulating new observed data and creating the sets of $m = 20$ imputations for $R = 100$ independent replications. In each replication, we compute point estimates and 95% confidence intervals using the multiple imputation inferences for $\Pr(Y_{2j} = 1)$ and $\Pr(Y_{1j} = 1, Y_{2j} = 1)$ for $j = 1, \ldots, 5$, and also for $\Pr(Y_{2j} = 1, Y_{25} = 1)$ for $j = 1, \ldots, 4$. We approximate the true values of these fourteen probabilities, $\{Q_j : j = 1, \ldots, 14\}$, by concatenating all $1000 \times R$ records across the R replications (before introducing missing data) and computing the resulting probabilities.

The method used for multiple imputation inference follows Rubin (1987). For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be, respectively, the estimate of some population quantity Q and the estimate of the variance of $q^{(l)}$ in completed data set $D^{(l)}$. Researchers use $\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m$ to estimate Q, and use $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate $\text{var}(\bar{q}_m)$, where $b_m = \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, with $\nu_m = (m-1)\left[1 + \bar{u}_m / \left\{(1 + 1/m)b_m\right\}\right]^2$ degrees of freedom.

For each Q_j, for each method we compute $\text{DIF}_j = |\sum_{r=1}^{R} (\bar{q}_j^r - Q_j)|/R$ and $\text{RMSE}_j = \sum_{r=1}^{R} (\bar{q}_j^r - Q_j)^2/R$, where $\bar{q}_j^r$ is the corresponding estimate $\bar{q}_m$ for Q_j during the replication r. For each j, for each method we also determine the percentage of the 95% confidence intervals that cover the corresponding Q_j.

Figures 1 and 2 summarize the performances of the two methods. As seen in Fig. 1, for nearly all probabilities, the semi-parametric AN selection model outperforms `Amelia II`, producing smaller values of DIF_j and RMSE_j. Not surprisingly, the MAR models in `Amelia II` fail to correct for the attrition bias in estimates of probabilities involving Y_2, whereas the AN model generally does. As seen in Fig. 2, the coverage rates of 95% confidence intervals when using the AN model are near the nominal 95% with one exception. This is not the case for the coverage rates that result from `Amelia II`, nearly all of which are far lower than 95%. In the bottom graph of Fig. 2, we see that the variance estimates from the AN model are not much larger than, and often actually smaller than, those from the application of `Amelia II`, showing that the better coverage rates of the AN model are not simply a result of undesirable variance inflation. In sum, this simulation demonstrates that the semi-parametric AN model is able to diagnose and correct for biases from non-ignorable panel attrition, producing inferences from panel survey data that are less biased than a standard multiple imputation approach.

6 Analysis of the AP-Yahoo News Panel

We next turn to an application of the semi-parametric AN selection model using the APYN election panel study. A number of scholars have used this study to examine opinion dynamics during the 2008 campaign (e.g., Pasek et al. 2009; Henderson and Hillygus 2011; Iyengar, Sood, and Lelkes 2012; Henderson, Hillygus, and Tompson 2010).[4] The APYN study tracked the political attitudes and behaviors over the course of the 2008 presidential campaign with eleven waves of data collection and three refreshment samples. The survey was sampled from the probability-based KnowledgePanel Internet panel.[5] Wave 1 was fielded on November 2, 2007—1 year before the

[4] In most cases, these previous works relied on post-stratification weights to correct for potential panel attrition bias, although Pasek et al. (2009) use standard multiple imputation using `Amelia II`.

[5] The panel recruits panel members via a probability-based sampling method using known published sampling frames that cover 96% of the U.S. population. Sampled non-internet households are provided a laptop computer or MSN TV unit and free internet service. The study was a collaboration between the AP and Yahoo Inc., with support from Knowledge Networks (KN) and collaboration with Norman Nie, Sunshine Hillygus, Michael Henderson, and Jon Krosnick.

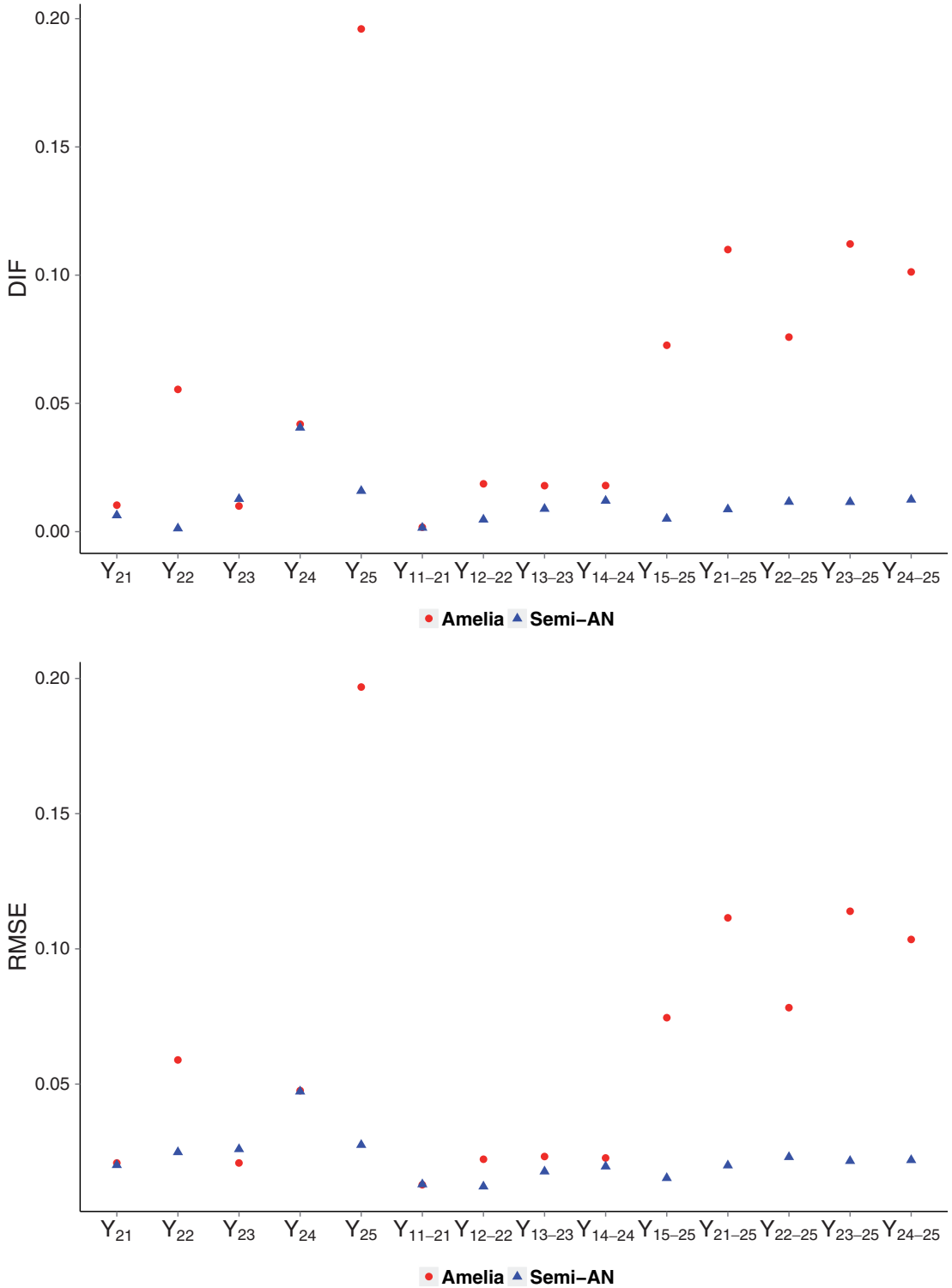

Fig. 1 Comparisons of bias and RMSE based on the semi-parametric AN model and a default application of `Amelia II` in the simulation studies. The label "Y_{2j}" denotes $\Pr(Y_{2j} = 1)$; "Y_{1j-2j}" denotes $\Pr(Y_{1j} = 1, Y_{2j} = 1)$; and "$Y_{2j-25}$" denotes $\Pr(Y_{2j} = 1, Y_{25} = 1)$.

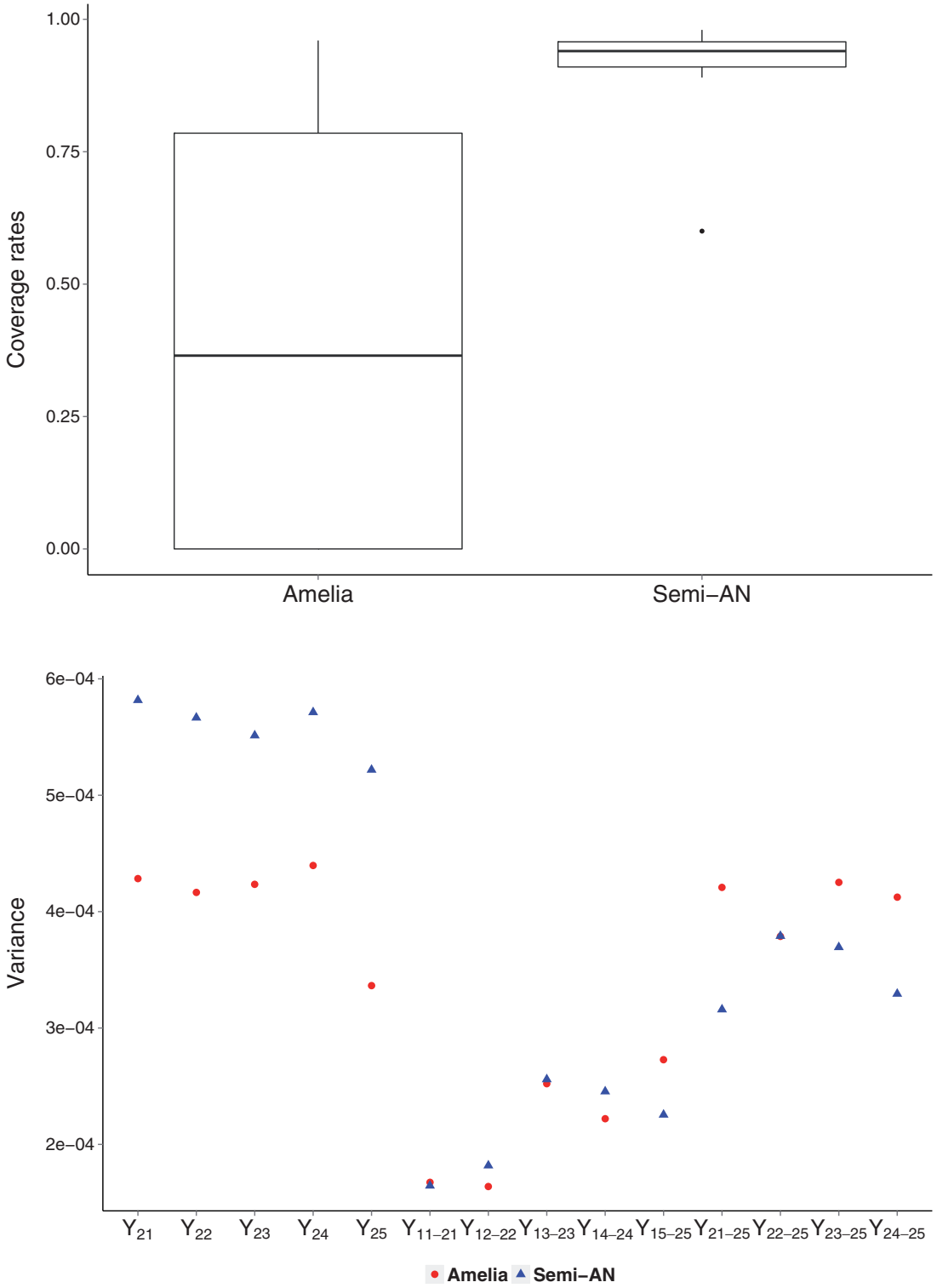

Fig. 2 Comparisons of coverage rates and variance estimates based on the semi-parametric AN model and a default application of `Amelia II` in the simulation studies. The label "Y_{2j}" denotes $\Pr(Y_{2j} = 1)$; "Y_{1j-2j}" denotes $\Pr(Y_{1j} = 1, Y_{2j} = 1)$; and "$Y_{2j-25}$" denotes $\Pr(Y_{2j} = 1, Y_{25} = 1)$.

election—and was completed by 2735 respondents.[6] We assume that the wave 1 respondents are representative of the fielded individuals.[7] After the initial wave, these wave 1 respondents were invited to participate in each follow-up wave, even if they failed to respond to the previous one. Thus, wave-to-wave attrition rates vary across time toward the end of the panel. Three external refresh cross-sections were also collected: a sample of 697 new respondents in January, 576 new respondents in September, and 464 new respondents in October. Our analysis focuses on wave 1 (November 2007) and the final wave before election with a corresponding refreshment sample (October 2008), which we label wave 2 for presentational clarity. Of those who completed wave 1, 1011 (36.97%) respondents failed to complete the October wave.

Our motivating question is a simple one: given the extent of panel attrition in the survey, can we correctly estimate population estimates of various political outcomes? We focus on several variables about voting participation and preferences that were asked in both the November 2007 and October 2008 waves, summarized in Table 2 along with demographic and political profile variables.[8] Because wave 1 of the survey was fielded before the candidates were nominated, we rely on Obama favorability as a key proxy for vote choice. We also have a number of variables related to turnout (registration, campaign interest, and likelihood to vote). These are all key measures for defining likely voters in pre-election polls (Traugott and Tucker 1984) and proxies for political interest, a substantive outcome of interest in its own right (Prior 2010). The measures are either nominal or ordinal with several fixed number of scales.[9]

Our political and demographic predictors include the following. Age (PPAGECT4) includes four categories: Age 18–29, 30–44, 45–59, and 60+. Education (PPEDUCAT) includes four education degree categories: Less than high school, High school, Some college, and Bachelor's degree or higher. Gender (PPGENDER) includes two categories: male and female. Race (PPETHM) includes two categories: Non-Hispanic White and all others. Income (PPINCIMP) includes four income categories: Less than $29,999, $30,000 to $49,999, $50,000 to $74,999, and $75,000 or more. Marital Status (PPMARIT) includes two categories: married and all others. These demographic variables are included in X in our models. We also include several political control variables. Party (PARTYID) has been recoded to have three categories: Democrat, Republican, and all others. Ideology (ID1) is the standard five-point scale, with the categories of very liberal, somewhat liberal, moderate, somewhat conservative, and very conservative. REL3 describes frequency of religious service attendance: more than once a week, once a week, a few times a month, a few times a year, and never.

For some questions, participants selected "Refused/Not Answered" and "Don't know enough to say" in both the panel and refreshment samples. We treat both of these answers as item nonresponse. Table 2 displays the counts and proportions of item nonresponse for all variables for the participants in wave 1, the remaining participants in wave 2, and the new participants in the refreshment sample. Item nonresponse missing rates vary from 0.18% up to 20.11% in wave 1, from 0.29% up to 12.76% for the non-attriters in wave 2, and from 0.65% up to 12.28% in the refreshment sample. For item nonresponse, we assume the missing values are MAR and impute them during the MCMC steps.

Following previous research (Hirano et al. 1998), we do not use survey weights in our analysis. It is, however, straightforward to adapt the semi-parametric AN model to account for a complex survey design. Analysts can impute missing values due to attrition and item nonresponse multiple times via the semi-parametric AN model, and then use survey weighted estimates in multiple

[6]This represents a 76.5% completion rate from those fielded the survey. Using the formula specified in Callegaro and DiSogra (2008), this represents a cumulative response rate of 11.2% when you take into account the panel recruitment response rate, the household profile rate, and the survey completion rate, excluding the household retention rate.
[7]That is, we assume that initial nonresponse is ignorable—a point we return to in the conclusion.
[8]Demographic and political profile variables are provided by KN at the time of the wave 1 or refreshment survey, but the variables were collected in profile surveys. These items have few missing values because they are updated as the respondent completes additional surveys or they are at times imputed based on other survey responses. Only the demographic variables were assumed to be time invariant and included in the AN model.
[9]Because of some extremely small marginal probabilities at some levels, some categories have been aggregated for FAV1-4, LV3, CND1, and LV31.

Table 2 Outcome and predictor variables

		Item nonresponse counts (%)		
Variable	*Levels*	*W1(2735)*	*W2(1724)*	*Ref(464)*
Obama favorability (FAV1-4)	2	550(20.11)	95(5.51)	20(4.31)
Voter registration (LV1)	2	9(0.33)	8(0.46)	0(0)
Campaign interest (LV3)	4	6(0.22)	5(0.29)	0(0)
Thought given to candidates (CND1)	2	5(0.18)	8(0.46)	3(0.65)
Likelihood to vote (LV31)	4	18(0.66)	220(12.76)	57(12.28)
Age (PPAGECT4)	4	0		0
Education (PPEDUCAT)	4	0		0
Race (PPETHM)	2	0		0
Gender (PPGENDER)	2	0		0
Income (PPINCIMP)	4	0		0
Marital status (PPMARIT)	2	0		0
Party (PARTYID)	3	10(0.37)		7(1.51)
Ideology (ID1)	5	57(2.08)		10(2.16)
Religiosity (REL3)	5	20(0.73)		8(1.72)

imputation inferences.[10] After imputation, analysts can disregard any adjustments to the survey weights for panel attrition, because the semi-parametric AN model is used to account for non-ignorable attrition.[11]

6.1 *Estimating the model*

We use the semi-parametric AN selection model with vague prior distributions, including (1) $\alpha \sim \text{Beta}(0.25, 0.25)$, (2) $\Psi \sim \text{Dirichlet}(1, \ldots, 1)$, and (3) $\beta \sim \text{N}(0, 1000I)$. We use $H = 30$ classes; posterior checks suggest that this is sufficiently large. We have $p_1 = p_2 = 5$ variables comprising Y_1 and Y_2, and $q = 9$ time-invariant variables comprising X.

We begin the MCMC sampler from Section 4 by initializing the parameters and filling in values for all missing data items. We set the initial value of $\alpha = 1$. We generate initial values for each V_h, where $h = 1, \ldots, H-1$, by sampling independently from Beta(1, 1); this is equivalent to draws from $H-1$ uniform distributions. We generate initial values for missing data in (X, Y_1, Y_2) due to item nonresponse by sampling from the marginal distributions of each variable based on an appropriate subset of observed cases. Specifically, we have the following initialization steps:

- For cases with missing X in the panel and the refreshment sample, sample initial values from the marginal distributions of X computed from the observed cases in the panel and the refreshment sample.
- For cases with missing Y_1 in the panel and the refreshment sample, sample initial values from the marginal distributions of Y_1 computed from the observed cases in the panel.
- For cases with missing Y_2 in the refreshment sample, sample initial values from the marginal distributions of Y_2 computed from the refreshment sample.
- For cases with missing Y_2 in the panel and $W = 1$, sample initial values from the marginal distributions of Y_2 computed from the cases with $W = 1$ in the panel.

[10] As in all multiple imputation settings, it is crucial to include variables in the imputation model that reflect salient features of the survey design, for example variables used to define primary sampling units (Reiter, Raghunathan, and Kinney 2006).

[11] For the APYN data, the wave 1 weights—which are the product of the inverse sampling probabilities and post-stratification factors to adjust for nonresponse in wave 1 (but not subsequent attrition)—would be appropriate to use for making population inferences after multiple imputation with the semi-parametric AN model. Using the wave 1 weights hardly changes our estimates, and the weighted and unweighted 95% confidence intervals overlap extensively. We prefer to report the model-based estimates simply to avoid the added complexity in explaining the results and to help focus on the semi-parametric AN model and its differences with complete-case results.

- For the missing values of Y_2 due to attrition in the panel ($W = 0$), we sample initial values from the conditional probabilities obtained by $\Pr(Y_2|W{=}0) = [\Pr(Y_2) - \Pr(Y_2|W{=}1)\Pr(W{=}1)]/\Pr(W{=}0)$, where the marginal distributions $\Pr(Y_2)$ are calculated using the refreshment sample, the conditional distributions $\Pr(Y_2|W{=}1)$ are estimated on the cases with $W{=}1$ in the panel, and the probability $\Pr(W{=}1/0)$ is based on the panel.

To initialize β, we find its maximum-likelihood estimate in the probit regression of W on Z using the initially completed panel data. To initialize values for W in the refreshment sample, we independently draw from Bernoulli distributions with success probability defined by equation (3) and computed from the initialized values of β and Z.

We run the MCMC chain for 100,000 iterations. After a burn-in period of 50,000 iterations, we keep every 50th draw, resulting in 1000 samples from the posterior distribution. To evaluate convergence of the MCMC chains, we examine trace plots for $\Pr(Z_j = 1)$, where $j = 1, \ldots, p$ computed from the 1000 draws of Θ and equation (4), the posterior samples of β, and the posterior samples of α. The trace plots of the thinned quantities display good mixing behaviors and suggest MCMC convergence. The posterior mode of the number of occupied clusters is 20 with a maximum of 27, indicating that $H = 30$ is sufficient.

As a byproduct of the MCMC, we have 1000 completed versions of Z. We keep every twentieth one of these completed data sets to make $m = 50$ multiple imputations for the missing data.

6.2 *Results*

We first consider the attrition model. Figure 3 displays the posterior medians and 95% credible intervals for the thirty-one coefficients in β, based on the dummy coding for each variable in (Y_2, Y_1, X). Several coefficients of variables in Y_2 have statistically significant effects, suggesting that analyses based on the panel data alone suffer from attrition bias. Most notably, the model suggests that individuals who remain in the panel are more likely at wave 2 to be registered and to have high interest in the campaign, but not to have indicated that they thought "a lot" about the candidates. Among the demographic variables, individuals who drop out are younger, less educated, male, and more likely to be married than single. These conclusions are largely consistent with previous research examining the predictors of panel attrition (Olson and Witt 2011). The coefficient for registration switches signs from wave 1 to wave 2, a reflection of that fact that very few people change their registration status across the two waves, so that LV1 in wave 1 is highly correlated with LV1 in wave 2.

We next use the $m = 50$ multiple imputations to estimate several quantities of substantive interest. Figure 4 displays multiple imputation inferences for $\Pr(Y_2{=}1)$ and, to visualize the attrition effect, for $\Pr(Y_2{=}1|W)$, where $W{=}1$ (non-attriters) and $W{=}0$ (attriters). In other words, the figure reports the estimate of the marginal distribution for each political outcome (diamond marker) as well as the estimates for predicted attriters (circle marker) and non-attriters (triangle marker).[12] Thus, this figure allows us to compare how estimates would differ using the semi-parametric AN model (diamond marker) compared to listwise deletion (triangle marker), as is common in political science applications.[13] We see that correcting for panel attrition bias has little effect on

[12]For analysis, all variables are coded as indicator variables based on the first category level. Specifically, FAV1-4 = 1 represents favorable toward Obama; LV1 = 1 means registered; LV3 = 1 tells that the interviewers have interest on campaign; CND1 = 1 represents that interviewers have given a lot of thought on campaign; and LV31 = 1 shows the participants will very likely vote (recoded after imputation to be in the same substantive direction of the other variables).

[13]The use of post-stratification weights that adjust for attrition is another common approach used in political science to correct for potential attrition bias, but the comparison to the AN results is not entirely straightforward because of the particular weights and variables available in the data file and the treatment of item nonresponse, which is simultaneously imputed by the AN model but is often ignored (listwise deletion) in most political science applications. We computed post-stratified estimates, based only on panel non-attriters and the reported wave 2 weights on the file (approximately the wave 1 weights for panel respondents multiplied by post-stratification adjustments for attrition). We compared these estimates to estimates from imputing wave 2 values for attriters using the AN model and weighted by the wave 1 weights (which correct for design effects and unit nonresponse in initial panel, but not attrition). We find large differences in the estimates for some of the outcomes, but not others. The largest difference is found for campaign

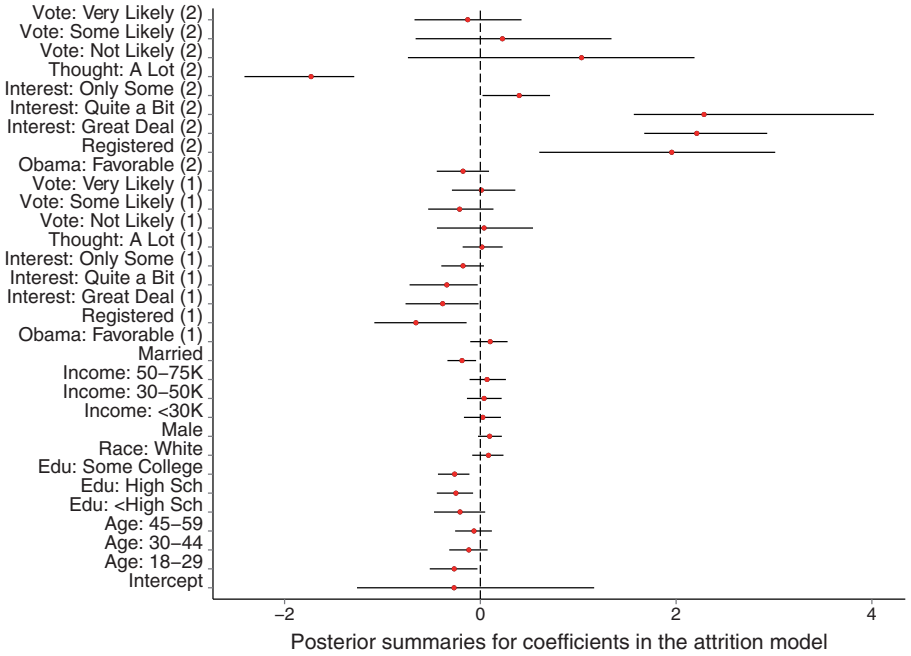

Fig. 3 Posterior medians and 95% credible intervals for regression coefficients β in the attrition model of the APYN study. In the labels, (2) indicates a measurement from wave 2, and (1) indicates a measurement from wave 1.

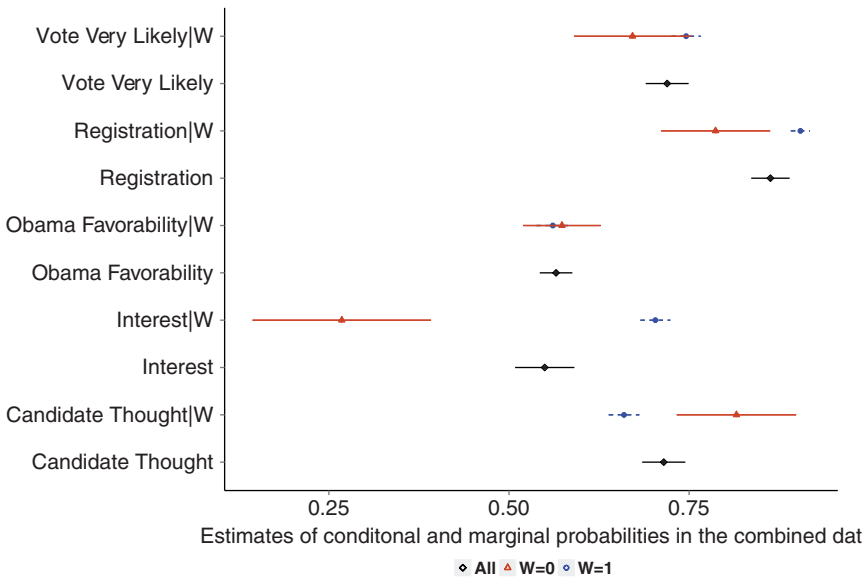

Fig. 4 Multiple imputation 95% confidence intervals for $\Pr(Y_2=1|W=1)$, $\Pr(Y_2=1|W=0)$, and $\Pr(Y_2=1)$. Inferences computed with $m=50$ completed data sets combining the panel and refreshment samples.

some estimates—most notably, Obama favorability—but large effects on others. In contrast, we would have had inaccurate estimates of campaign interest, thought about candidates, and voter

interest (AN model estimates 48% interested compared to 66% using post-stratification weights adjusted for attrition). In contrast, there is little difference in the estimate of Obama favorability using the two approaches (AN model estimates 58% favorable compared to 57% favorable using post-stratification weights for attrition).

registration using only non-attriters because these respondents are more likely to register, have more interest in the campaign, but say they have given less thought to the candidates.[14] Although we do not have external benchmarks available for most of these variables, we can compare our estimate of voter registration to that from a "gold standard" source. Using listwise deletion, the APYN survey estimates that 91% of citizens are registered to vote. The weighted estimate is 89%, although this is not directly comparable because the weights reflect design features and unit nonresponse. Correcting for panel attrition using the semi-parametric AN model reduces that estimate to 86%, which comes quite close to the 2008 ANES estimate of 87%.[15]

Although we must be cautious in drawing conclusions about the nature of panel attrition in other surveys or for other outcomes, these results are consistent with previous research and suggest that panel attrition in political surveys is more likely to bias outcomes related to political engagement than outcomes related to vote choice.

To offer a more detailed illustration of the impact of the attrition bias correction, we focus more closely on the outcome of campaign interest. We estimate the corrected and uncorrected level of campaign interest by party, gender, and race to see how correcting non-ignorable panel attrition bias can change subgroup estimates.[16] As evident in Fig. 5, the inferences based on the attrition-adjusted data set differ from those based on only the individuals who remain in the panel, even at a subgroup level. In particular, we would overestimate interest in the campaign among white Democratic and white Republican respondents if we had ignored attrition and considered only those who had stayed in the panel. For example, 77.2% of Democratic, white women are estimated to be interested in the campaign if we consider only those still left in the panel in October 2008, whereas the attrition-corrected estimate is 61.4%. Similarly, ignoring panel attrition would give us an estimate of 75.1% of Republican, white men being interested in the campaign, whereas the attrition-corrected estimate is 60.8%.

6.3 *Model Diagnostics*

To check the fit of the models, we follow the advice in Deng et al. (2013) and use posterior predictive checks (Meng 1994; Gelman et al. 2005; Burgette and Reiter 2010; He, Zaslavsky, and Landrum 2010). We use the estimated semi-parametric AN selection model to generate $T = 500$ data sets with no item missing data in (X, Y_1, Y_2). Let $\{D^{(1)}, \ldots, D^{(T)}\}$ be the collection of the T data sets with no item nonresponse. For each $D^{(t)}$, we also use the model to generate new values of Y_2 for the complete panel (cases in the panel with $W_i = 1$) and the refreshment sample, leaving any imputations for item nonresponse as fixed. This can be done after running the MCMC to convergence as follows. For given draws of parameter values and any item missing data, sample new values for the observed Y_2 using the distributions in Step 3 of Section 4. Let $\{R^{(1)}, \ldots, R^{(T)}\}$ be the collection of the T replicated data sets.

We then compare statistics of interest in $\{R^{(1)}, \ldots, R^{(T)}\}$ to those in $\{D^{(1)}, \ldots, D^{(T)}\}$. Specifically, suppose that S is some statistic of interest, such as a marginal or conditional probability in our context. For $t = 1, \ldots, T$, let $S_{R^{(t)}}$ and $S_{D^{(t)}}$ be the values of S computed from $R^{(t)}$ and $D^{(t)}$, respectively. We compute the two-sided posterior predictive probability,

$$\text{ppp} = (2/T) * \min\left(\sum_{t=1}^{T} I(S_{R^{(t)}} - S_{D^{(t)}} > 0), \sum_{t=1}^{T} I(S_{D^{(t)}} - S_{R^{(t)}} > 0)\right). \quad (25)$$

[14]Substantively, it is perhaps initially surprising that attriters were less likely to have given a lot of thought to the candidates, but recall that the model includes wave 1 measures of all outcome variables. Indeed, separate analyses similarly suggested that non-attriting panelists—while generally more politically informed, engaged, and interested than attriters—may have gotten burned out on the campaign by Election Day.

[15]The attrition-corrected numbers are closer to, but still higher than, the 2008 CPS November Supplement estimate of 71%, although those estimates are thought to be underestimated due to the ineligibility of felons and non-citizens.

[16]For simplicity, for this analysis campaign interest (LV3) is coded as a binary measure: interested (a great deal and quite a bit) or not (only some, very little, and not at all).

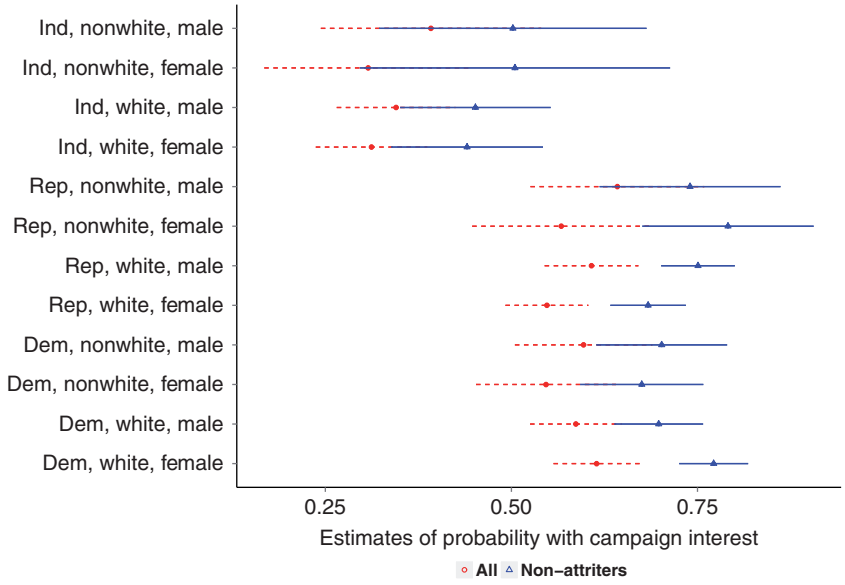

Fig. 5 Multiple imputation 95% confidence intervals based on the AN model for all data and the complete-cases estimates for the probability of being interested in the campaign. Results based on m=50 completed data sets.

When the value ppp is small, for example less than 5%, this suggests the replicated data sets are systematically different from the observed data set (after filling in the item nonresponse), with respect to that statistic. When the value of ppp is not small, the imputation model generates data that look like the completed data for that statistic.

As quantities S of interest, we use $\Pr(Y_2 = 1|W = 1)$ for individuals in the panel, and $\Pr(Y_2 = 1)$ for individuals in the refreshment samples. This results in ten statistics. Figure 6 displays a histogram of the values of ppp. The results do not suggest any serious lack of model fit.

7 Conclusions

We have presented a Bayesian approach to correct for panel attrition bias in a two-wave panel with a refreshment sample in cases with many categorical survey variables and item nonresponse. The approach includes (1) an AN selection model for the attrition process; and (2) a DP mixture of multinomial distributions for the categorical survey variables. We apply the model to the APYN panel study, illustrating how ignoring attrition leads to biased inferences of some political outcomes of interest because the tendency of participants to attrite is systematically related to their political attitudes and preferences. Most notably, we find that we significantly overestimate campaign interest in many subgroups in the electorate if we fail to correct for panel attrition bias. Given that campaign interest was one of the key measures used to define likely voters in media analysis using this data set during the 2008 campaign, the potential substantive consequences of this finding could be quite severe.

In addition to drawing attention in political science to methods for correcting non-ignorable panel attrition using refreshment samples, this article builds on the previous methodological literature by extending the AN model for use in high-dimensional settings with item nonresponse. These are contributions that address some of the real data complexities researchers encounter in panel surveys, thus helping bring attrition correction methods out of the theoretical methodological literature and into the analyses of political science research.

We do want to note potential limitations in applying the proposed method. Although the DP mixture model can capture complex dependencies among the substantive variables, one still has to specify the selection model. In the APYN application we include only main effects. We do not have any specific *a priori* expectations that this is problematic in this application, but it is possible that

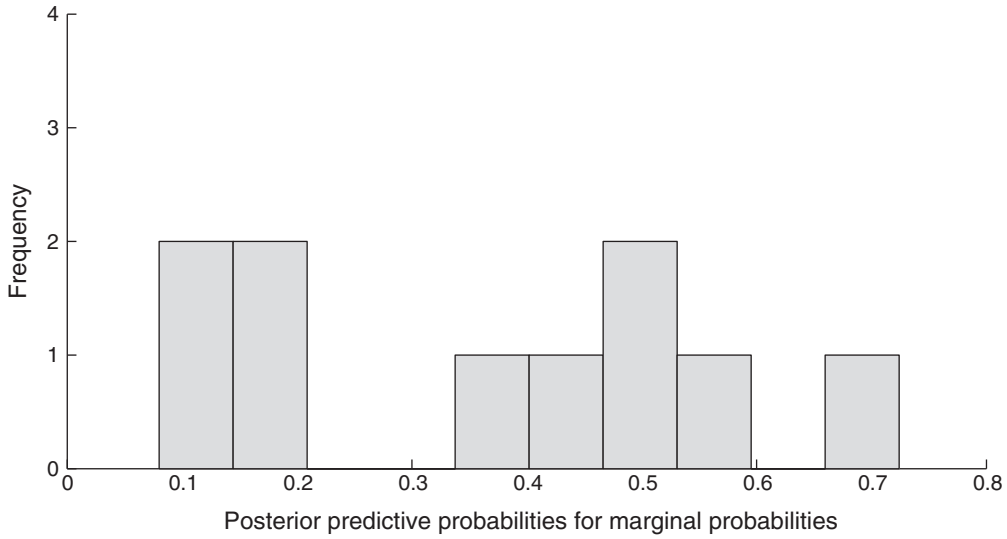

Fig. 6 Posterior predictive probabilities for $\Pr(Y_2 = 1|W = 1)$ for individuals in the panel and $\Pr(Y_2 = 1)$ for individuals in the refreshment samples.

there are interactions between variables (of course, like all AN models, for identification we require no interaction between Y_1 and Y_2) that are predictors of attrition. This approach might not be appropriate for applications in which there are expected complex dependencies in the attrition process that are not readily observable from the data. Second, and relatedly, the parameters of the selection model can be difficult to estimate, particularly when the dimension of the predictors is large. Binary regression models with many predictors are subject to separation problems, which can make estimation unstable. Bayesian MCMC for these models can be very slow to converge, particularly with large numbers of regression coefficients. In such cases, analysts may need to reduce the predictor space, for example by using various model selection techniques to predict W from X in the panel data.

Finally by ignoring unit nonresponse in wave 1 and the refreshment samples, we effectively rely on a (partially) ignorable missingness mechanism. Although this is common in the literature on panel attrition and AN models (e.g., Hirano et al. 1998; Bhattacharya 2008b), it may not be desirable given recent trends in nonresponse to surveys.[17] This would be especially problematic if there were differential patterns of initial unit nonresponse to wave 1 compared to the refreshment sample. For example, suppose politically disinterested individuals refused to respond in the refreshment sample at higher rates than politically interested individuals (given covariates) because they were especially sick of politics by mid-October. The resulting over-representation of $Y_{2i} = 1$ in the complete cases in the refreshment sample would show up as apparent non-ignorable attrition, even if in fact attrition was a MAR mechanism. Similarly, suppose individuals who are politically interested at both waves failed to respond to the initial wave—and hence are not available for wave 2—at higher rates than individuals who are politically interested in wave 1 and disinterested in wave 2. The under-representation of $Y_{2i} = 1$ (given $Y_{1i} = 1$ and X) in the complete cases in the panel again would present as apparent and perhaps specious non-ignorable attrition.

We are not aware of any published work in which non-ignorable nonresponse in the initial panel or refreshment samples is accounted for in inference. One potential path forward is to incorporate in the AN model non-ignorable mechanisms for unit nonresponse via selection or pattern mixture models developed for cross-sectional data (Little and Rubin 2002). The analyst could then investigate the sensitivity of inferences to multiple assumptions about the non-ignorable missingness mechanisms in the initial wave and refreshment samples. Although the semi-parametric AN model

[17]Reassuringly, some research finds that lower response rates often do not bias survey estimates (Keeter et al. 2006).

does not address all potential sources of bias in panel surveys, the correction of non-ignorable panel attrition simultaneously with item nonresponse in a computationally efficient approach that can be applied to high-dimensional settings offers researchers an important tool for being able to draw inferences from panel surveys.

Funding

This work was supported by the National Science Foundation [SES1131897, SES1061241].

References

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422):669–79.

Allison, P. 2000. Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research* 28:301–9.

Bartels, L. M. 1999. Panel effects in the American National Election Studies. *Political Analysis* 8(1):1–20.

Behr, A., E. Bellgardt, and U. Rendtel. 2005. Extent and determinants of panel attrition in the European community household panel. *European Sociological Review* 21(5):489–512.

Bhattacharya, D. 2008a. Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics* 144(2):430–46.

———. 2008b. Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics* 144:430–46.

Brehm, J. 1993. *The phantom respondents: Opinion surveys and political representation*. Ann Arbor: University of Michigan Press.

Brown, C. H. 1990. Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 46(1):143–55.

Burgette, L. F., and J. P. Reiter. 2010. Multiple imputation via sequential regression trees. *American Journal of Epidemiology* 172:1070–6.

Callegaro, M., and C. DiSogra. 2008. Computing response metrics for online panels. *Public Opinion Quarterly* 72(5):1008–32.

Clinton, J. 2001. Panel bias from attrition and conditioning: A case study of the Knowledge Networks panel. In *AAPOR 55th Annual Conference*.

Cranmer, S. J., and J. Gill. 2013. We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science* 43(02):425–49.

Deng, Y., D. S. Hillygus, J. P. Reiter, Y. Si, and S. Zheng. 2013. Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science* 22:238–56.

Diggle, P., and M. G. Kenward. 1994. Informative dropout in longitudinal data analysis. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 43(1):49–93.

Dunson, D. B., and C. Xing. 2009. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* 104:1042–51.

Erosheva, E. A., S. E. Fienberg, and B. W. Junker. 2002. Alternative statistical models and representations for large sparse multi-dimensional contingency tables. *Annales de la Faculté des Sciences de Toulouse* 11(4):485–505.

Frankel, L., and S. Hillygus. 2013. Looking beyond demographics: Panel attrition in the ANES and GSS. *Political Analysis* 22(1):1–18.

Frick, J. R., J. Goebel, E. Schechtman, G. G. Wagner, and S. Yitzhaki. 2006. Using analysis of Gini (ANOGI) for detecting whether two subsamples represent the same universe: The German Socio-Economic Panel Study (SOEP) experience. *Sociological Methods Research* 34:427–68.

Gelman, A., I. Van Mechelen, G. Verbeke, D. F. Heitjan, and M. Meulders. 2005. Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* 61:74–85.

Grimmer, J. 2010. A Bayesian hierarchical topic model for political texts: Measuring expresses agendas in Senate press releases. *Political Analysis* 18(1):1–35.

Hausman, J. A., and D. A. Wise. 1979. Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica* 47(2):455–73.

He, Y., A. M. Zaslavsky, and M. B. Landrum. 2010. Multiple imputation in a large-scale complex survey: A guide. *Statistical Methods in Medical Research* 19:653–70.

Heeringa, S. 1997. *Russia longitudinal monitoring survey sample attrition, replenishment, and weighting: Rounds V–VII*. University of Michigan Institute for Social Research.

Henderson, M., and D. S. Hillygus. 2011. The dynamics of health care opinion, 2008–2010: Partisanship, self-interest, and racial resentment. *Journal of Health Politics, Policy, and Law* 36(6):945–60.

Henderson, M., D. Hillygus, and T. Tompson. 2010. "Sour grapes" or rational voting? Voter decision making among thwarted primary voters in 2008. *Public Opinion Quarterly* 74(3):499–529.

Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin. 1998. *Combining panel data sets with attrition and refreshment samples*. Technical Report 230, National Bureau of Economic Research.

———. 2001. Combining panel data sets with attrition and refreshment samples. *Econometrica* 69:1645–59.

Hogan, J. W., and M. J. Daniels. 2008. *Missing data in longitudinal studies*. Boca Raton, FL: Chapman and Hall.
Holmes, C. C., and L. Held. 2006. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1(1):145–68.
Honaker, J., and G. King. 2010. What to do about missing values in time-series cross-section data. *American Journal of Political Science* 54(2):561–81.
Ishwaran, H., and L. F. James. 2001. Gibbs sampling for stick-breaking priors. *Journal of the American Statistical Association* 96:161–73.
Iyengar, S., G. Sood, and Y. Lelkes. 2012. Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly* 76(3):405–31.
Keeter, S., C. Kennedy, M. Dimock, J. Best, and P. Craighill. 2006. Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly* 70(5):759–79.
Kenward, M. G. 1998. Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine* 17:2723–32.
Kenward, M. G., G. Molenberghs, and H. Thijs. 2003. Pattern-mixture models with proper time dependence. *Biometrika* 90:52–71.
King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95:49–69.
Kish, L., and I. Hess. 1959. A "replacement" procedure for reducing the bias of nonresponse. *American Statistician* 13:17–19.
Kropko, J., B. Goodrich, A. Gelman, and J. Hill. 2014. Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*. Published online doi:10.1093/pan/mpu007.
Kruse, Y., M. Callegaro, J. Dennis, S. Subias, M. Lawrence, C. DiSogra, and T. Tompson. 2009. Panel conditioning and attrition in the AP-Yahoo! News Election Panel Study. In *64th Conference of the American Association for Public Opinion Research (AAPOR)*. Hollywood, FL.
Kyung, M., J. Gill, and G. Casella. 2011. New findings from terrorism data: Dirichlet process random effects models for latent groups. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 60:701–21.
Lin, I., and N. C. Schaeffer. 1995. Using survey participants to estimate the impact of nonparticipation. *Public Opinion Quarterly* 59:236–58.
Little, R. J. A. 1993. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 88:125–34.
Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley & Sons.
Little, R. J. A., and Y. Wang. 1996. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 52(1):98–111.
Meng, X. 1994. Posterior predictive *p*-values. *Annals of Statistics* 22:1142–60.
Olsen, R. J. 2005. The problem of respondent attrition: Survey methodology is key. *Monthly Labor Review* 128:63–71.
Olson, K., and L. Witt. 2011. Are we keeping the people who used to stay? Changes in correlates of panel survey attrition over time. *Social Science Research* 40(4):1037–50.
Papaspiliopoulos, O. 2008. *A note on posterior sampling from Dirichlet mixture models*. Technical report, Centre for Research in Statistical Methodology, University of Warwick.
Pasek, J., A. Tahk, Y. Lelkes, J. A. Krosnick, B. K. Payne, O. Akhtar, and T. Tompson. 2009. Determinants of turnout and candidate choice in the 2008 US presidential election illuminating the impact of racial prejudice and other considerations. *Public Opinion Quarterly* 73(5):943–94.
Prior, M. 2010. You've either got it or you don't? The stability of political interest over the life cycle. *Journal of Politics* 72:747–66.
Reiter, J. P., T. E. Raghunathan, and S. Kinney. 2006. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology* 32(2):143–50.
Ridder, G. 1992. An empirical evaluation of some models for non-random attrition in panel data. *Structural Change and Economic Dynamics* 3:337–55.
Rubin, D. B. 1987. *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.
Scharfstein, D. O., A. Rotnitzky, and J. M. Robins. 1999. Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association* 94(448):1096–120.
Schluchte, M. D. 1982. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 11(14):1861–70.
Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica* 4:639–50.
Si, Y., and J. P. Reiter. 2013. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* 38(5):499–521.
Si, Y., J. P. Reiter, and D. S. Hillygus. 2014. *Replication data for: Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples*. http://dx.doi.org/10.7910/DVN/25367 (accessed April 19, 2014). IQSS Dataverse Network, V1.
Su, Y.-S., M. Yajima, A. E. Gelman, and J. Hill. 2011. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* 45(2):1–31.

Thompson, M., G. Fong, D. Hammond, C. Boudreau, P. Driezen, A. Hyland, R. Borland, K. Cummings, G. Hastings, M. Siahpush. 2006. Methods of the International Tobacco Control (ITC) four-country survey. *Tobacco Control* 15(Suppl. 3) iii12–iii18.

Traugott, M. W., and C. Tucker. 1984. Strategies for predicting whether a citizen will vote and estimation of electoral outcomes. *Public Opinion Quarterly* 48(1):330–43.

Vehovar, V. 1999. Field substitution and unit nonresponse. *Journal of Official Statistics* 15:335–50.

Vermunt, J. K., J. R. Van Ginkel, V. Der Ark, L. Andries, and K. Sijtsma. 2008. Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* 38(1):369–97.

Walker, S. G. 2007. Sampling the Dirichlet mixture models with slices. *Computations in Statistics-Simulation and Computation* 36:45–54.

Wawro, G. 2002. Estimating dynamic panel data models in political science. *Political Analysis* 10:25–48.

Wissen, L., and H. Meurs. 1989. The Dutch mobility panel: Experiences and evaluation. *Transportation* 16:99–119.

Zabel, J. 1998. An analysis of attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an application to a model of labor market behavior. *Journal of Human Resources* 33:479–506.