Correcting for Survey Misreports using Auxiliary Information with an Application to Estimating Turnout

Jonathan N. Katz
California Institute of Technology

Gabriel Katz
California Institute of Technology

ABSTRACT

Misreporting is a problem that plagues researchers that use survey data. In this paper, we develop a parametric model that corrects for misclassified binary responses using information on the misreporting patterns obtained from auxiliary data sources. The model is implemented within the Bayesian framework *via* Markov Chain Monte Carlo (MCMC) methods, and can be easily extended to address other problems exhibited by survey data, such as missing response and/or covariate values. While the model is fully general, we illustrate its application in the context of estimating models of turnout using data from the American National Elections Studies.

Much of the empirical work in the social sciences is based on the analysis of survey data. However, as has been widely documented (Bound, Brown and Mathiowetz 2001), these data are often plagued by measurement errors. There are many possible sources for such errors. Interviewers may erroneously record answers to survey items, and respondents may provide inaccurate responses due to an honest mistake, misunderstanding or imperfect recall (Bound, Brown and Mathiowetz 2001; Hausman, Abrevaya and Scott-Morton 1998; Molinari 2003). Also, as underscored by the social psychology literature, survey respondents

tend to overreport socially desirable behaviors and underreport socially undesirable ones (Loftus 1975). In the case of discrete or categorical variables, mismeasurement problems have been traditionally referred to as "misclassification" errors (Molinari 2003).

In the political science literature, concerns about misclassification have been particularly prevalent in the analysis of voting behavior. Empirical studies of the determinants of voter turnout focus on how the probability of an individual voting varies according to relevant observable factors, such as citizen's level of political information, registration laws, or demographic characteristics.[1] The decision to vote, however, is typically not observed due to the use of secret ballot in the U.S. Furthermore, even if we could observe turnout from the official ballots we would not, in general, be able to observe all the characteristics — e.g., the voter's policy preferences or information about the candidates — that presumably affect the decision. Hence, political scientist rely on the use of survey instruments, such as the American National Election Study (ANES) or the Current Population Survey (CPS), that include both measures of respondents' relevant characteristics and their *self-reported* voting behavior. This almost always leads to estimation of the common logit or probit models, since the turnout decision is dichotomous, although there are alternatives such as scobit (Nagler 1994) or non-parametric models (Härdle 1990) for discrete choice models.

However, it has been long established that some survey respondents misreport voting, i.e., they report that they have voted when in fact they did not do so (Burden 2000; Katosh and Traugott 1981; Sigelman 1982). The evidence that misreporting is a problem can be found in a series of validation studies that the ANES conducted in 1964, 1976, 1978, 1980, 1984, 1988 and 1990. After administering a post-election survey to a respondent, an official from the ANES was sent to the respondent's local registrar of elections to see if in fact she was recorded as having voted in the election. This is not an easy task, since respondents

[1]The literature is far too vast to even begin to fully cite here. See Aldrich (1993) for a review of the theoretical literature and Wolfinger and Rosenstone (1980) for an influential empirical study.

often do not know where they voted, election officials differ in their ability to produce the records in a usable form, and there might be differences between the survey data and the public records due to errors in spelling or recording. This means that the validated data may also be mismeasured, but for this paper we will assume it is correct.

That said, the ANES for these years included both the respondent's *self-reported* vote and the *validated* vote. The differences between the two measures are fairly shocking. Depending on the election year, between 13.6% and 24.6% percent of the respondents claiming to have voted did in fact not according to the public records.[2] In contrast, only between 0.6% and 4.0% of the respondents in the 1964 – 1990 validated surveys who reported not having voted did vote according to the official records. Since there is no reason to believe that measurement errors should mainly be of false positives — i.e., reporting voting when the official record contradicts this claim — this lends some credence to the social pressures argument for misreporting (Bernstein, Chadha and Montjoy 2001) and should help mitigate some of our concerns about other potential sources of classification errors, such as inaccurate records. The large differences between reported and validated turnout led to a cottage industry analyzing the causes of misreporting and to a debate about how to best measure it (e.g., Abramson and Claggett (1986), Cassel (2003), Katosh and Traugott (1981), Sigelman (1982)). All of these studies find that misreporting varies systematically with some characteristics of interest, but none of them provides an estimation solution to correct for possible misreporting.

The open question then is what to do about the problem of respondents misreporting. One possibility would be to use only validated data. At some level this is an appealing option. If we are sure that the validated data is correct, then estimation and inference is straightforward. Unfortunately, collecting the validated turnout data is difficult and expen-

[2]The Current Population Survey (CPS) also exhibits considerable turnout overreporting, although the magnitude is substantially lower than for the ANES (Highton 2004). As shown by Hausman, Abrevaya and Scott-Morton (1998) and Neuhaus (1999), however, even modest amounts of misreporting can affect parameter estimates.

sive, and ANES has stopped doing validation studies for these reasons. Furthermore, even if validation studies were free, some states, such as Indiana, make it impossible to validate votes. Hence, if we are going to limit ourselves to use only fully validated data, our samples will be much smaller. Moreover, we would also be throwing away the useful information included in the already collected but non-validated studies.[3] On the other hand, simply ignoring misreporting and using self-reported turnout to estimate standard probit or logit models can result in biased and inconsistent parameter estimates and inaccurate standard errors, potentially distorting the relative impact of the characteristics of interest on the response variable and leading to erroneous conclusions (Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999).[4]

In this paper we develop a simple Bayesian approach to correct for misreporting, allowing researchers to continue to use the self-reported data while improving the accuracy of the estimates and inferences drawn in the presence of misclassified binary responses.[5] Our model draws on Hausman, Abrevaya and Scott-Morton (1998), but incorporates information on the misreporting process from auxiliary data sources, aiding in identification (Molinari 2003) and making it easier to avoid the problems that limit the use of Hausman, Abrevaya and Scott-Morton (1998)'s modified maximum likelihood estimator in small samples

[3]In the case of the ANES, turnout is one of the few survey items included since the late 1940s, covering a larger period than any other continuing survey (Burden 2000). Validation studies, on the other hand, only comprise a handful of elections.

[4]While other procedures for reducing the frequency of overreporting - e.g., altering question wording or reformulating survey questions (Bound, Brown and Mathiowetz 2001) - can also improve the quality of future datasets, we would still be wasting large amounts of data collected in previous surveys.

[5]We focus on the case of misclassified responses and error-free covariates. Several methods have been proposed to adjust for measurement error in the covariates. See (Carroll, Ruppert and Stefanski 1995) for a review.

such as those typically used in political science (Christin and Hug 2004; Gu 2006). While incorporating this information into the study of the sample of interest using frequentist methods is far from straightforward (Prescott and Garthwaite 2005), the Bayesian paradigm provides a flexible framework for summarizing and integrating historical or supplementary evidence on misreport patterns from different sources and levels of analysis (Dunson and Tindall 2000; Ibrahim and Chen 2000; Prescott and Garthwaite 2002). Using Markov chain Monte Carlo (MCMC) methods, the model presented here allows placing prior restrictions on the misclassification probabilities or on relevant regression coefficients based on estimates from previous studies or summary statistics, improving identification and convergence when adjusting for misreporting and avoiding asymptotic approximations that may not apply in small or moderate samples. Our approach also enables us to simultaneously address another important problem with survey data, namely missing outcome and/or covariate values, using Bayesian model-based imputation (Ibrahim et al. 2005). Compared to alternative imputation techniques, Bayesian methods allow easily estimating standard errors in multiparameter problems and handling "nuisance" parameters, and have been shown to be particularly efficient when data loss due to missing observations is substantial, as is the case of the Election Studies examined here (Ibrahim, Chen and Lipsitz 2002).

Although other Bayesian approaches have been proposed to adjust for misclassification using prior information to overcome fragile or poor identifiability, they either rely exclusively on elicitation of experts' opinions (e.g., Paulino, Soares and Neuhaus (2003)) or assume that information on both the true and the fallible response is available for all subjects in a random subsample of the data (Prescott and Garthwaite 2002, 2005). In contrast, the information on the misreport patterns incorporated into our model need not come from the sample of interest, and can be combined with elicitation of experts' beliefs if needed. In the empirical application presented in this paper we will use earlier and small-sample validation studies to correct for misreporting. However, matched official records, administrative registers and possibly even aggregate data might be used to gain this information. Given the potential

difficulties of eliciting probabilities from experts' opinions and the scarcity of internal validation designs relative to administrative data sets, external validation studies and other sources of ancillary information (Bound, Brown and Mathiowetz 2001), the correction developed in this paper provides a more flexible way of incorporating prior information and can be more widely applied than existing approaches.[6] In addition, these alternative approaches focus only on the case in which the misclassification rates are independent of all covariates. As mentioned above, this assumption seems to be inappropriate when examining the determinants of voter turnout, as well as in many other potential applications. The magnitude and direction of the biases when misreporting is covariate-dependent can be quite different than in the case of constant misclassification rates (Neuhaus 1999) and, in the context of analyzing voting behavior, Bernstein, Chadha and Montjoy (2001) show that ignoring the correlation between the covariates of interest and the misreport probabilities may seriously distort multivariate explanations of the turnout decision.

While our model is developed in the context of estimating the conditional probability of turning out to vote, the method is general and will be applicable whenever misclassification of a binary outcome in a survey is anticipated and there is auxiliary information on the misreporting patterns. For instance, our approach could be used to analyze survey data on participation in pension plans and social welfare programs (Molinari 2003), energy consumption (Gu 2006), employment status (Hausman, Abrevaya and Scott-Morton 1998) and many other areas where we expect to see substantial rates of misreporting and potential correlation between some of the covariates affecting the response and the misreport probabilities. The model can also be implemented when misreporting depends on covariates other than those influencing the outcome. For example, for a substantial proportion of the CPS sample, turnout is measured by proxy, rather than self-reported (Highton 2004). In this case, the

[6]In internal validation studies, the true response is available for a subset of the main study and can be compared to the imperfect or observed response. In the case of external validation designs, the misreport pattern is estimated using data outside the main study.

misclassification probabilities would be modeled using information on misreporting patterns among household members reporting other members' turnout decision, which could be obtained from validated CPS studies.[7] Extensions of our method to discrete choice models with more than two categories along the lines of Abrevaya and Hausman (1999) are possible as well.

The paper proceeds as follows. The next section formally lays out the estimation problem in the presence of misreporting and develops our proposed solution. Section 2 presents results from a Monte Carlo experiment evaluating the robustness of our approach to misspecification of the misreport model. In Section 3, we provide three applications of our methodology using data on voter turnout from the ANES. Finally, Section 4 concludes.

1. CORRECTING FOR MISREPORTING IN BINARY CHOICE MODELS

1.1. *Defining the Problem*

Let y_i be a dichotomous (dummy) variable, and denote by $\mathbf{x}_i$ a vector of individual characteristics of interest. We want to estimate the conditional distribution of y_i given $\mathbf{x}_i$, $\Pr[y_i|\mathbf{x}_i]$. However, instead of observing the "true" dependent variable y_i, assume we observe the self-reported indicator $\tilde{y}_i$. Most studies use the observed $\tilde{y}_i$ as the dependent variable, typically running either a probit or logit model to estimate $\Pr[\tilde{y}_i = 1|\mathbf{x}_i]$.

In order to know whether this substitution can lead to incorrect inferences, we need to know the relationship between $\Pr[\tilde{y}_i = 1|\mathbf{x}_i]$ and $\Pr[y_i = 1|\mathbf{x}_i]$. We can always write

$$\begin{aligned}\Pr[\tilde{y}_i = 1|\mathbf{x}_i] = \Pr[\tilde{y}_i = 1|\mathbf{x}_i, y_i = 1] \cdot \Pr[y_i = 1|\mathbf{x}_i] + \\ \Pr[\tilde{y}_i = 1|\mathbf{x}_i, y_i = 0] \cdot \Pr[y_i = 0|\mathbf{x}_i],\end{aligned} \tag{1}$$

by the law of total probability. All that we have done is to rewrite the probability $\Pr[\tilde{y}_i =$

[7] We thank an anonymous referee for pointing us to this potential application of our model.

$1|\mathbf{x}_i]$ into two components: when the self-reported or observed variable $\tilde{y}_i$ coincides with the true response y_i, and when it does not. Also, noting that $\Pr[\tilde{y}_i = 0|\mathbf{x}_i, y_i = 1] = 1 - \Pr[\tilde{y}_i = 1|\mathbf{x}_i, y_i = 1]$ we can re-write the relationship as

$$\Pr[\tilde{y}_i = 1|\mathbf{x}_i] = (1 - \pi_i^{1|0} - \pi_i^{0|1}) \Pr[y_i = 1|\mathbf{x}_i] + \pi_i^{1|0} \tag{2}$$

where $\pi_i^{1|0} = \Pr[\tilde{y}_i = 1|y_i = 0, \mathbf{x}_i]$ is the probability that the respondent falsely claims $\tilde{y}_i = 1$ when in fact $y_i = 0$, and $\pi_i^{0|1} = \Pr[\tilde{y}_i = 0|y_i = 1, \mathbf{x}_i]$ is the probability that the observed response takes the value 0 when the true response is $y_i = 1$. It is important to the probability of each type of misreporting is conditional on $\mathbf{x}_i$.

Standard methods for estimating binary choice models generally assume that the conditional distribution of the dependent variable given $\mathbf{x}_i$ is known up to a parameter vector β. However, unless $\pi_i^{0|1} = \pi_i^{1|0} = 0 \quad \forall i$, estimating the conditional probability $\Pr[\tilde{y}_i = 1|\mathbf{x}_i]$ rather than $\Pr[y_i = 1|\mathbf{x}_i]$ will generally lead to biased estimates of β and inaccurate standard errors, with even small probabilities of misreporting potentially leading to significant amounts of bias (Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999). In addition, the marginal effect of covariate x on the observed response $\tilde{y}_i$ and on the true response y_i will differ by

$$\begin{aligned} \frac{\partial \Pr[\tilde{y}_i = 1|\mathbf{x}_i]}{\partial x} - \frac{\partial \Pr[y_i = 1|\mathbf{x}_i]}{\partial x} = &-\left(\frac{\partial \pi_i^{1|0}}{\partial x} + \frac{\partial \pi_i^{0|1}}{\partial x}\right) \Pr[y_i = 1|\mathbf{x}_i] \\ &- (\pi_i^{1|0} + \pi_i^{0|1}) \frac{\partial \Pr[y_i = 1|\mathbf{x}_i]}{\partial x} + \frac{\partial \pi_i^{1|0}}{\partial x}. \end{aligned} \tag{3}$$

As a result, inferences drawn on the relationship between the covariates of interest and the response variable may change substantially when estimated based on the likelihood function defined by $\Pr[\tilde{y}_i = 1|\mathbf{x}_i]$ rather than on the true model $\Pr[y_i = 1|\mathbf{x}_i]$, depending on the distribution of $\beta'\mathbf{x}_i$ and the covariate vector $\mathbf{x}_i$, on the prevalence of misclassification and on the relationship between the probabilities of misreporting and the covariates in $\mathbf{x}_i$ (Bernstein, Chadha and Montjoy 2001; Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999).

Different parametric models have been proposed to correct for misclassification of the dependent variable in binary choice models (Carroll, Ruppert and Stefanski 1995; Haus-

man, Abrevaya and Scott-Morton 1998; Paulino, Soares and Neuhaus 2003; Prescott and Garthwaite 2002, 2005).[8] In particular, Hausman, Abrevaya and Scott-Morton (1998) proposed a modified maximum likelihood estimator that requires the "monotonicity" condition $\pi_i^{1|0} + \pi_i^{0|1} < 1$ to achieve identification. Using Monte Carlo simulations, they showed that their model consistently estimates the extent of misclassification and the parameter vector β, at least in large samples. More recently, however, Christin and Hug (2004) replicated the work of Hausman, Abrevaya and Scott-Morton (1998) for different sample sizes, and found that the modified maximum likelihood estimator performed consistently better than simple probit models ignoring misclassification only in samples of 5,000 or more observations. As noted by Gu (2006), the failure of Hausman, Abrevaya and Scott-Morton (1998)'s estimator in small samples is likely due to the insufficiency of the monotonicity condition to ensure model identification. For such sample sizes typically available in political science, even moderate rates of misclassification may hinder model identification, so different assumptions may be required to put bounds on the misclassification rates and the regression coefficients. In addition, Hausman, Abrevaya and Scott-Morton (1998) and, in fact, most empirical applications of models proposed to correct for misreporting, assume constant misclassification rates, failing to account for the potential influence of the covariates of interest on $\pi^{1|0}$ and $\pi^{0|1}$.[9]

Relevant prior information on the misreport patterns is often available from auxiliary data sources, such as internal or external validation studies, small sample pilots or administrative registers, which can be used to impose restrictions on the misreport probabilities and regression coefficients to aid in identification and improve inferences on the relationship between $\mathbf{x}$ and $\mathbf{y}$ (Molinari 2003). In order to incorporate this information, we propose a simple

[8] A review of different methods developed to deal with misclassification and measurement errors in nonlinear models can be found in Carroll, Ruppert and Stefanski (1995).

[9] Abrevaya and Hausman (1999); Hausman, Abrevaya and Scott-Morton (1998) and Paulino, Soares and Neuhaus (2003), among others, discuss extensions to deal with covariate-dependent misclassification, but they do not analyze this case in practice.

Bayesian approach based on Markov Chain Monte Carlo (MCMC) methods that can be easily implemented by practitioners and applied researchers using flexible and freely available software for Bayesian analysis such as WinBUGS or JAGS (Plummer 2009; Spiegelhalter, Thomas and Best 2003).

1.2. *A Bayesian Model to Correct for Misreporting using Auxiliary Data*

We are interested in accurately estimating the effect of relevant individual characteristics on the conditional distribution of the true response. Hence, the focus of our analysis lies in the marginal posterior distribution of β, while the model for the conditional probabilities $\pi_i^{1|0}$ and $\pi_i^{0|1}$ can be regarded as "instrumental".

Since the observed response variable is dichotomous, we can start by assuming that, conditional on some set of individual characteristics, the observations are independently and identically distributed according to a Bernoulli distribution — as in Hausman, Abrevaya and Scott-Morton (1998). The probability of the sample can therefore be written as

$$\mathcal{L}(\theta|\tilde{\mathbf{y}}, \mathbf{x}) = \prod_{i=1}^{N} \Pr[\tilde{y}_i|\mathbf{x}_i, \theta]^{\tilde{y}_i}(1 - \Pr[\tilde{y}_i|\mathbf{x}_i, \theta])^{1-\tilde{y}_i}, \tag{4}$$

with $\theta = \{\pi_i^{1|0}, \pi_i^{0|1}, \beta'\}$. We will further assume that the conditional probability of the true response variable is given by $\Pr[y_i = 1|\mathbf{x}_i] = F(\beta'\mathbf{x}_i)$, where $F(\cdot)$ is some cumulative density function. For ease of exposition, we use the probit link, so that $F(\cdot)$ is the standard normal distribution denoted by $\Phi(\cdot)$.[10] We also assume that $\Pr[y_i = 1|\mathbf{x}_i]$ is *a priori* independent of $\pi_i^{1|0}$ and $\pi_i^{0|1}$.[11] Substituting for $\Pr[\tilde{y}_i|\mathbf{x}_i, \theta]$ in Equation 2 and denoting by $\mathcal{S}$ the sample

[10]This will lead to a probit model with a correction for misreporting; the use of the logit link function would result in a logit model with a correction for misreporting.

[11]This assumption simplifies the analysis considerably without entailing any obvious drawback from a practical perspective (Paulino, Soares and Neuhaus 2003).

data, we arrive at:

$$\mathcal{L}(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S}) = \prod_{i=1}^{N}\Big[\big[(1-\pi_i^{1|0}-\pi_i^{0|1})\Phi(\beta'\mathbf{x}_i)+\pi_i^{1|0}\big]^{\tilde{y}_i} \times \big[(1-\pi_i^{1|0}-\pi_i^{0|1})(1-\Phi(\beta'\mathbf{x}_i)+\pi_i^{0|1}\big]^{1-\tilde{y}_i}\Big], \tag{5}$$

which represents the probability of observing the sample under misreporting. The joint posterior density of $\theta = \{\pi_i^{1|0}, \pi_i^{0|1}, \beta\,'\}$ is therefore given by:

$$p(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S}) \propto \mathcal{L}(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S}) \times p(\beta, \pi_i^{1|0}, \pi_i^{0|1}). \tag{6}$$

Suppose that both the true and the self-reported dependent variables are recorded for all respondents in a validation study of size M. Comparing y_j to $\tilde{y}_j$ for every $j = 1, \ldots, M$, we can estimate the misreport probabilities for the validated sample. Let $\mathbf{z}_j^1$ and $\mathbf{z}_j^2$ denote sets of regressors that are useful in predicting the conditional probabilities $\pi_j^{1|0}$ and $\pi_j^{0|1}$, where the notation allows for the fact we may use different regressors to predict the two types of misreporting. $\mathbf{z}_j^1$ and $\mathbf{z}_j^2$ may include some or all of the variables in $\mathbf{x}$, as well as other variables not affecting the true response. Again, we assume probit link functions and specify the conditional probabilities of misreporting as $\pi_j^{1|0} = \Phi(\gamma_1\,'\mathbf{z}_j^1)$ and $\pi_j^{0|1} = \Phi(\gamma_2\,'\mathbf{z}_j^2)$. Letting $\mathcal{V}$ denote the data from the validation study, the likelihood from $\mathcal{V}$ is:

$$\begin{aligned}\mathcal{L}(\beta, \gamma_1, \gamma_2|\mathcal{V}) = &\prod_{j=1}^{M}\big(\Phi(\beta'\mathbf{x}_j)\big)^{y_j}\big(1-\Phi(\beta'\mathbf{x}_j)\big)^{(1-y_j)} \quad \times \\ &\prod_{y_j=1}\Phi(\gamma_1\,'\mathbf{z}_j^1)^{\tilde{y}_j} \times \big(1-\Phi(\gamma_1\,'\mathbf{z}_j^1)\big)^{1-\tilde{y}_j} \quad \times \\ &\prod_{y_j=0}\Phi(\gamma_2\,'\mathbf{z}_j^2)^{1-\tilde{y}_j} \times \big(1-\Phi(\gamma_2\,'\mathbf{z}_j^2)\big)^{\tilde{y}_j}.\end{aligned} \tag{7}$$

The posterior distributions $p(\beta, \gamma_1, \gamma_2|\mathcal{V})$ or $p(\gamma_1, \gamma_2|\mathcal{V})$ could then be used to specify the priors for β, γ_1 and γ_2 in the model fit to the sample of interest by repeated application of Bayes' theorem. However, since these posteriors cannot be expressed as tractable distributions, there is no straightforward way of transferring the relevant information from the validation study to the analysis of the main sample (Prescott and Garthwaite 2005). In

addition, unless the validation study is a random sub-sample of the main study, heterogeneity between the two samples might in some circumstances lead to misleading conclusions if inference on β is based on the pooled datasets (Dunson and Tindall 2000). Hence, we consider both samples simultaneously, combining the likelihoods in Equations 5 and 7 with vague independent priors $p(\beta), p(\gamma_1)$ and $p(\gamma_2)$ and weighting the likelihood from the validated sample by a "tunning" parameter δ that controls how much influence the validated data has relative to the main sample (Ibrahim and Chen 2000). The joint posterior density of the unknown parameters thus becomes:

$$p(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S}) \propto \mathcal{L}(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S}) \times \mathcal{L}(\beta, \gamma_1, \gamma_2|\mathcal{V})^{\delta} \times p(\beta) \times p(\gamma_1) \times p(\gamma_2) \tag{8}$$

with $0 \leq \delta \leq 1$, where $\delta = 0$ corresponds to the case in which no auxiliary information is incorporated into the analysis for the main sample, while $\delta = 1$ gives equal weights to $\mathcal{L}(\beta, \pi_i^{1|0}, \pi_i^{0|1}|\mathcal{S})$ and $\mathcal{L}(\beta, \gamma_1, \gamma_2|\mathcal{V})$. δ can be assigned either a fixed value or a prior distribution – e.g., $\delta \sim Beta(a, b)$ – (Ibrahim and Chen 2000).[12] Although Equation 8 is intractable analytically, inference can be performed using Gibbs sampling along with Metropolis steps to sample the full conditionals for β, γ_1 and γ_2 (Gelfland and Smith 1990). Under mild regularity conditions, for a sufficiently large number of iterations, samples from these conditional distributions approach samples from the joint posterior (Robert and Casella 2004). The posterior marginals obtained from these convergent samples can then summarized and used to estimate the effect of the relevant individual characteristics on the true response and the misreport probabilities.

Consequently, we only need to have validated data from a previous sample or for a subsample of the respondents in order to correct for misreporting in the model for the main study. In case several validation studies are available, they can be easily integrated into our analysis by adapting the method proposed in Ibrahim and Chen (2000) to incorporate

[12] In the latter case, the prior $p(\delta)$ would be added to Equation 8. See the discussions in Ibrahim and Chen (2000) for additional details.

historical data in binary choice models, substituting $\mathcal{L}(\beta, \gamma_1, \gamma_2|\mathcal{V})$ in Equation 8 by:

$$\prod_{d=1}^{D} \mathcal{L}(\beta, \gamma_1, \gamma_2|\mathcal{V}_d)^{\delta_d} \tag{9}$$

where $\mathcal{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_D\}$ denotes the data from D validation samples and $\delta = \{\delta_1, \ldots, \delta_D\}$, $0 \leq \delta_d \leq 1$ can be assigned I.I.D. Beta priors. Note that, while we must assume the same error structure and similar misreporting processes in the validated and non-validated samples, the covariates included in $\mathbf{x}$ and $\mathbf{z} = \{z^1, z^2\}$ do not have to be necessarily identical for both datasets. For instance, when estimating the determinants of the turnout decision, we could allow for election-specific factors affecting the turnout and the misreport probabilities, combining information from validation studies with experts' opinions, theoretical restrictions or specifying diffuse priors for some of the predictors. Covariates that were not measured in previous studies can be incorporated into the analysis of the sample of interest through the "initial" prior $p(\beta, \gamma_1, \gamma_2)$ in Equation 8 (Ibrahim et al. 2005).

Even if we did not have access to a validation sample, several other sources of information, such as administrative records or even aggregate data could be used to impose informative constraints on the misclassification rates and improve the parameter estimates. For example, in the analysis of voter turnout, we may observe turnout rates in small geographic areas, such as counties or congressional districts, that could be used to specify the misreport probabilities for all individuals in the sample belonging to a given area. Hierarchical beta priors can then be used to summarize auxiliary information available on misreporting patterns by location or relevant socio-demographic characteristics following the approach in Dunson and Tindall (2000). A Bayesian hierarchical model would also allow combining aggregate and individual data on misreport patters if available (Congdon 2002). Finally, if no auxiliary data is available to predict misreporting, constraints on the misreport probabilities could be imposed *via* elicitation of experts' opinions. Our model would then be virtually identical to Paulino, Soares and Neuhaus (2003).

Despite the advantages of our approach, it is worth mentioning that, like all parametric estimators, our model might be quite sensitive to distributional and modeling assumptions.

Although semi-parametric methods have been used to estimate discrete choice models with misclassified dependent variables (Abrevaya and Hausman 1999; Hausman, Abrevaya and Scott-Morton 1998), they are also subject to potential misspecification (Molinari 2003). A different approach would be to adapt and implement non-parametric methods based on Horowitz and Manski (1995) and Molinari (2003).[13] In particular, the "direct misclassification approach" proposed by the latter allows incorporating prior information on the misreporting pattern to obtain interval identification of parameters of interest, and can be easily applied to the case in which misclassification depends on observed covariates with relatively little computational cost. However, as is well known, non-parametric methods are subject to the curse of dimensionality, which can pose a problem in applications where the misreporting probabilities might depend on a relatively large set of covariates, and is uncertain whether point identification can be achieved in this setting (Hu 2008). To the best of our knowledge, there is very little research comparing the performance of parametric *versus* non-parametric methods to correct for covariate-dependent misclassification and evaluating the relative weaknesses and advantages of both approaches in applied work.

1.3. *Extending the model to account for missing data*

Besides measurement errors, survey data is often plagued with large proportions of missing outcome and covariate values due to non-response or loss of data. As is well known, unless the data are missing completely at random (MCAR), using list-wise deletion and restricting the analysis only to those respondents who are completely observed can lead to biased parameter estimates (Little and Rubin 2002).[14] Furthermore, even if the data are

[13]This was the approached taken by Jackman (1999) to handle both misclassification and non-response in surveys about political participation.

[14]It is worth mentioning, however, that there are situations in which inference based on a complete-case analysis might yield unbiased estimates and outperform imputation methods

MCAR, complete-case analyses may lead to discard a large proportion of observations and can be therefore quite inefficient (Ibrahim et al. 2005).

While several alternative procedures have been proposed to accommodate missing data, fully Bayesian methods such as the one presented in this paper are especially appealing when dealing with small sample sizes and when the fraction of missing values is considerable, and can be easily implemented withour requiring new techniques or additional steps for inference (Ibrahim, Chen and Lipsitz 2002; Ibrahim et al. 2005).[15] There is no distinction between missing data and parameters within the Bayesian framework, and thus inference in this setting essentially requires defining a prior for the missing values and sampling from the joint posterior distribution of the parameters and missing values, incorporating just an "extra-layer" in the Gibbs sampling algorithm compared to the complete-case analysis (Ibrahim et al. 2005). Hence, our model can be immediately extended to deal with missing response and covariate values, including cases with missing responses alone, with missing covariates alone, and with missing covariates and responses. This allows us to accommodate item and unit nonresponse in both the main and the validation studies.[16]

Let $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,p})'$, $i = 1, \ldots, N$, denote a $p \times 1$ vector of covariates included in $\mathbf{x}_i, \mathbf{z}_i^1$ and $\mathbf{z}_i^2$, and denote the marginal density of $\mathbf{w}_i$ by $p(\mathbf{w}_i|\alpha)$, where α parametrizes the joint distribution of the covariates. If some of hte covariates are missing, we can write $\mathbf{w}_i = (\mathbf{w}_{i,obs}, \mathbf{w}_{i,mis})$, where $\mathbf{w}_{i,mis}$ is the $q_i \times 1$ vector of missing components of $\mathbf{w}_i$, $0 \leq q \leq p$, and $\mathbf{w}_{i,obs}$ is the observed portion of $\mathbf{w}_i$. Similarly, we use $\tilde{y}_{i,mis}$ if the self-reported outcome $\tilde{y}_i$ is missing, and $\tilde{y}_{i,obs}$ otherwise. Assuming that the missing data mechanism is *ignorable*

---

even when the data are not missing completely at random (Ibrahim et al. 2005).

[15] A detailed review of different methods commonly used to handle missing data is beyond the scope of this paper. See Ibrahim et al. (2005) and Little and Rubin (2002), among others, for a detailed discussion.

[16] However, as seen in Equation 10 below, respondents with completely missing outcomes and covariates do not contribute to the likelihood function.

(Little and Rubin 2002), the *observed-data likelihood* for the main study reduces to:

$$\begin{aligned}\mathcal{L}(\beta,\gamma_1,\gamma_2,\alpha|\mathcal{S}_{obs}) = & \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=\mathbf{w}_{i,obs}} p(\tilde{y}_i|\mathbf{w}_i,\beta,\gamma_1,\gamma_2)\, p(\mathbf{w}_i|\alpha) \quad \times \\ & \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis})} \int p(\tilde{y}_i|\mathbf{w}_i,\beta,\gamma_1,\gamma_2)\, p(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}|\alpha)\, d\mathbf{w}_{i,mis} \quad \times \\ & \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\mathbf{w}_{i,obs}} p(\mathbf{w}_i|\alpha) \;\times \\ & \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis})} \int p(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}|\alpha)\, d\mathbf{w}_{i,mis}.\end{aligned} \tag{10}$$

As suggested by Ibrahim, Chen and Lipsitz (2002), it is often convenient to model the joint distribution $p(\mathbf{w}_i|\alpha)$ is as a series of one-dimensional conditional distributions:

$$\begin{aligned}p(w_{i,1},\ldots,w_{i,p}|\alpha) = & \, p(w_{i,p}|w_{i,1},\ldots,w_{i,P-1},\alpha_p) \\ & \times p(w_{i,p-1}|w_{i,1},\ldots,w_{i,p-2},\alpha_{p-1}) \times \cdots \times p(w_{i,1}|\alpha_1)\end{aligned} \tag{11}$$

where α_l, $l = 1,\ldots,p$, is a vector of parameters for the lth conditional distribution, the α_l's are distinct, and $\alpha = (\alpha_1,\ldots,\alpha_p)$. Specification 11 has the advantages of easing the prior elicitation for α and reducing the computational burden of the Gibbs algorithm required for sampling from the observed data posterior, and is particularly well-suited for cases in which $\mathbf{w}$ includes categorical and continuous covariates.[17]

Information on the misreport patterns and on all the parameters of interest can be incorporated from the validation study in essentially identical way as in the case with no missing data. A joint prior for $(\beta,\gamma_1,\gamma_2,\alpha)$ could be specified as:

$$p(\beta,\gamma_1,\gamma_2,\alpha) \propto \mathcal{L}(\beta,\gamma_1,\gamma_2,\alpha|\mathcal{V}_{obs})^{\delta} \times p(\beta) \times p(\gamma_1) \times p(\gamma_2) \times p(\alpha),$$

[17]Obviously, 11 needs to be specified only for those covariates that have missing values. If some of the covariates in $\mathbf{w}$ are completely observed for all respondents in a survey, they can be conditioned on when constructing the distribution of the missing covariates.

where $\mathcal{L}(\beta, \gamma_1, \gamma_2, \alpha|\mathcal{V}_{obs})$ is obtained from the complete-data likelihood of the validation study:

$$\mathcal{L}(\beta, \gamma_1, \gamma_2, \alpha|\mathcal{V}_{obs}) = \int \int p(\tilde{\mathbf{y}}, \mathbf{y}|\mathbf{w}, \beta, \gamma_1, \gamma_2, \alpha)\, d\tilde{y}_{mis}\, d\mathbf{w}_{mis},$$

and, as mentioned in 1.2, δ is a scalar prior parameter that weights the validated data relative to the data from the main study. Note that this specification allows for missing responses $\tilde{y}_i$ and covariate values in the validated sample as well, and can accommodate cases in which the missing self-reported variable depends on the true y_i.

In principle, it is possible to extend this approach to the case of non-ignorably missing values. However, there is usually little information on the missing data mechanism, and the parameters of the missing data model are often quite difficult to estimate (Ibrahim et al. 2005). The plausibility of the assumption that the data is missing at random (MAR) can be enhanced by including additional individual and contextual variables in the model specification (Gelman, King and Liu 1998).

2. ASSESSING ROBUSTNESS TO THE SPECIFICATION OF THE MISREPORT MODEL: A MONTE CARLO EXPERIMENT

In this section, we conduct a series of simulation analyses aimed at assessing the sensitivity of our method to misspecification of the model of misreporting. This is a particularly relevant issue, since misspecification of the misreport model may lead to inconsistent estimates of β and affect inferences on the covariate of interest (Abrevaya and Hausman 1999; Hausman, Abrevaya and Scott-Morton 1998). Drawing on research analyzing a somewhat similar problem, namely, the sensitivity of the estimated treatment effects to the specifications of the propensity score model (Zhao 2008), we examine the influence on the estimated covariate effects of misspecifying the disturbance distribution and the linear predictor of the misreport model. For reasons of space, we only present a brief overview of the results from

the Monte Carlo simulations. A detailed analysis is presented in Katz and Katz (2009).[18]

Based on the Monte Carlo design in Neuhaus (1999), we simulated 2,000 observations for two covariates: x_1 is drawn from a standard normal distribution, and x_2 is a dummy variable equal to one with probability 1/2. The true response y_i was generated as:

$$y_i = I(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i \geq 0)$$

where $I(E)$ is the indicator function equal to one if E is true and zero otherwise, $(\beta_0, \beta_1, \beta_2) = (-1, 1, 1)$ and ϵ_i drawn from a $N(0, 1)$ distribution. We also simulated a dichotomous variable d_i:

$$d_i = \begin{cases} I\big(\gamma_{1,0} + \gamma_{1,1} x_{i,1} + \gamma_{1,2} x_{i,2} + \eta_i \geq 0\big); & \text{if } y_i = 0 \\ I\big(\gamma_{2,0} + \gamma_{2,1} x_{i,1} + \gamma_{2,2} x_{i,2} + \eta_i \geq 0\big); & \text{if } y_i = 1 \end{cases}$$

where η is an error term, and $\gamma_1 = \{\gamma_{1,0}, \gamma_{1,2}, \gamma_{1,3}\}$, $\gamma_2 = \{\gamma_{2,0}, \gamma_{2,2}, \gamma_{2,3}\}$, are chosen to obtain different levels of misclassification and different degrees of correlation between the simulated covariates and the misreport probabilities $\pi_i^{1|0}$ and $\pi_i^{0|1}$. The observed response was then obtained as:

$$\tilde{y}_i = \begin{cases} I(d_i = 1); & \text{if } y_i = 0 \\ 1 - I(d_i = 1); & \text{if } y_i = 1. \end{cases}$$

In order to analyze the sensitivity of our method to misspecification of the error disturbance in the model of misreporting, we follow Horowitz (1993) and Zhao (2008) and considered 4 distributions for η: a standard normal distribution, a logistic distribution, a bimodal distribution - $\eta = 0.5N(3, 1) + 0.5N(-3, 1)$ - and heteroskedastic error terms $\eta \sim N(1, 1 + 0.1x_1^2)$. We also implemented 4 alternative specifications for the linear predictor of the

[18] In Katz and Katz (2009), we conduct a comprehensive simulation study comparing the performance of our approach *vis a vis* alternative models proposed in the literature to account for misreporting in the presence of both misclassification and missing data.

misreport model:

$$\begin{aligned}
&\text{Specification 1}: && \tau_{k,0} + \tau_{k,1} x_{i,2}; \\
&\text{Specification 2}: && \tau_{k,0} + \tau_{k,1} x_{i,1} + \tau_{k,2} x_{i,2} + \tau_{k,3} x_{i,1}^2; \\
&\text{Specification 3}: && \tau_{k,0} + \tau_{k,1} x_{i,1} + \tau_{k,2} x_{i,2} + \tau_{k,3} (x_{i,1} \times x_{i,2}); \\
&\text{Specification 4}: && \tau_{k,0} + \tau_{k,1} x_{i,1} + \tau_{k,2} x_{i,2} + \tau_{k,3} x_{i,3};
\end{aligned}$$

with $k = 1, 2$ and x_3 drawn from a log-normal distribution. We examined the effect of both forms of misspecification separately - i.e., we correctly specified the linear predictor of the misreport model when analyzing the role of misspecified error distributions and used standard normal errors when examining the influence of the functional form of the index term. In all cases, we randomly selected half of the observations in the sample and assigned them to be the validation study. We ignored the true response and the information on the misreport probabilities for the remaining 1,000 observations, using the information from the validated sub-sample to fit the model in Equation 8.[19]

Figure 1 reports the estimates of the marginal covariate effects when x_1 is omitted from the linear predictor of the misreport model (Specification 1) for different values of $\gamma_{1,1}, \gamma_{2,1}$ and average symmetric misreport rates of 5%, 10% and 20%.[20] The estimates of the marginal effect of x_1 worsen as the average misclassification rates increase and as the the correlation between the covariate and the misreport probabilities increase. However, for all values of $\gamma_{1,1}, \gamma_{2,1}$, the estimates from our model are closer to the true marginal effects than those from a model ignoring misreporting. The estimates for x_2, on the other hand, are virtually

[19]Since the validation study is a random sub-sample of the main study, a point mass prior $\delta = 1$ with probability 1 was used, equally weighting the validated and main samples. We also let the covariates in $\mathbf{z}^1$ and $\mathbf{z}^2$ differ across specifications and considered several values of γ_1, γ_2, with little change in the main substantive results presented in this section. Additional details are available in Katz and Katz (2009).

[20]In all cases, we set $\gamma_{1,2} = 1.25$, $\gamma_{2,2} = -1.25$, and adjust the value of the intercept to achieve the desired average misclassification rates.

unaffected by the omission of x_1 from the model of misreporting, and are again between 6 and 23 percentage points closer to the true effects than those from a standard probit model.

Table 1 complements the information from the figure, illustrating the influence of the other forms of misspecification considered for different values of $\bar{\pi}^{1|0}, \bar{\pi}^{0|1}, \gamma_1$ and γ_2. Adding irrelevant covariates and unnecessary nonlinear terms to the linear predictor of the misreport model has relatively little influence on the estimated marginal effects, and the same holds for the case of misspecified disturbance distributions. In all cases, the true average covariate effects lie within the central 95% credible intervals from our model, and the point estimates are betwen 4 and 18 percentage points closer to the true values than those obtained ignoring misreporting. It is worth mentioning that, as illustrated in Katz and Katz (2009), the estimates of γ_1 and γ_2 can be far away from the true coefficients when the model of misreporting is misspecified, particularly when the error terms are bimodal or heteroskedastic (Horowitz 1993; Zhao 2008). Nonetheless, the estimated covariate effects seem to be quite robust to the specification of the misreport model and much more accurate than those from standard parametric models when misclassificaton is non-negligible.

We also conducted additional simulations assuming a slightly different misreport processes for the validated and the main samples. Specifically, the values of γ_1 and γ_2 in the main sample were obtained by adding uniformly distributed errors to the corresponding parameters from the validation study. The amount of misclassification and the direction of the relationship between the covariates and the misreport probabilities was preserved, but we changed the magnitude of the effect of x_1 and x_2 on $\pi_i^{1|0}$ and $\pi_i^{0|1}$. Again, as illustrated in Table 1, the marginal effects estimated from our model are quite close to the true covariate effects. In contrast, the model ignoring misclassification systematically underestimates $\partial Pr(y = 1|\mathbf{x})/\partial x_1$ and overestimates $\partial Pr(y = 1|\mathbf{x})/\partial x_2$. We must note, though, that these results are based on limited simulation analyses and may not be true in general.

3. AN EMPIRICAL APPLICATION: CORRECTING FOR MISREPORTING IN THE ANALYSIS OF VOTER TURNOUT

Next, we illustrate the potential consequences of misreporting in the context of estimating the determinants of voter turnout and provide three different applications of our methodology using data from all the validated ANES surveys between the 1978 and 1990.[21] This dataset comprises three Midterm (1978, 1986, 1990) and three Presidential elections (1980, 1984, 1988), and has the obvious advantage of allowing us to directly compare the estimates from our model to a known benchmark, i.e., the same model estimated directly on the validated vote. We assume the validated vote to be the "gold-standard" measure of turnout, although there is considerable disagreement on this point (Burden 2000). The concern is that the validation studies are far from perfect. As stated at the outset, vote validation is expensive and difficult. The ANES is conducted in two parts, a pre- and post- election survey. In the studies from 1978, 1980, 1984, 1986, 1988 and 1990 there were in total 11,632 completed post election surveys. Unfortunately, of these completed surveys, the ANES was unable to validate 2,189 respondents, about 19.8 percent of the usable sample.[22] The majority of these failures were caused either because no registration records were found or because the local election office refused to cooperate with the ANES. If we are willing to maintain the assumption that these errors are essentially random (in the sense of being independent of the characteristics of interest), then there is no real harm done. The measurement error

[21] We use data from the 1978–1990 studies to preserve the comparability of the survey questions regarding the conditions of the interview; we will use this information to model the conditional probability of misreporting. The main substantive results reported in this Section hold for the Current Population Survey as well, and are available from the authors upon request.

[22] The rate of non-validation varies considerably across Election Studies, from around 2% of sample in 1978 to more than 31% in 1990.

will merely result in less efficient estimates of the misreporting model and a corresponding reduction in efficiency of the corrected turnout model. However, if there is systematic error, then we are just substituting one form of measurement error for another.

In Section 3.1, we estimate a simple model of the determinants of the turnout decision using both *self-reported* and *validated* turnout as the dependent variable in order to assess the consequences of ignoring misreporting. In 3.2, we re-estimate the turnout model with self-reported vote but applying our proposed solution to correct for misreporting, using a random sample of each survey as a validation sub-study. In Section 3.3, we apply our correction for misreporting under an external validation design, using information from previous ANES studies to correct for misreporting in the main sample. Both applications are based on a complete-case analysis. We deal with the problem of incomplete data in 3.4, where we account for item and unit non-response using the approach described in Section 1.3.

3.1. *Turnout misreporting in the 1978–1990 ANES*

As mentioned in the introduction, it has long been established in the political science literature that survey respondents often report to have voted when they did not actually do so (Bernstein, Chadha and Montjoy 2001; Katosh and Traugott 1981; Sigelman 1982). Figure 2 illustrates the differences between turnout rates computed from self-reported and validated vote in the six ANES studies under analysis. Validated turnout is systematically lower than reported turnout, and while both rates tend to follow similar trends, differences vary considerably across years, ranging from 7 percentage points in 1990 to more than 15 percentage points in 1980. 17.3% of the survey respondents who claimed to have voted but did not do so according to the validated data, and more than 28% of those who did not vote according to the official records responded affirmatively to the turnout question. In contrast, only 84 respondents in the 1978-1990 ANES studies reported not voting when the official record suggested they did, representing 0.7% of the sample. Additional descriptive

statistics on vote misreporting in the 1978–1990 validated ANES can be found in Table 2 in the Appendix.

In order to examine whether such high rates of overreporting affect inferences on the determinants of the turnout decision, we fit two hierarchical probit models allowing for election year and regional effects with both *self-reported* (SR) and *validated turnout* (V) as the response variable:

$$\Pr[\widetilde{y}_i = y_i^{SR}] \sim Bernoulli(p_i^{SR});$$
$$p_i^{SR} = \Phi(\lambda_t^{SR} + \eta_r^{SR} + \beta^{SR\prime}\mathbf{x}_i)$$

and

$$\Pr[y_i = y_i^{V}] \sim Bernoulli(p_i^{V});$$
$$p_i^{V} = \Phi(\lambda_t^{V} + \eta_r^{V} + \beta^{V\prime}\mathbf{x}_i);$$

where the k=1,...,K elements of β^s, s=SR, V, are assigned diffuse prior distributions:

$$\beta_k^s \sim N(\mu_{\beta_k}, \sigma_{\beta_k}^2)$$

and λ_t^s and η_r^s are election- and region- random effects, with:

$$\lambda_t^s \sim N(\mu_\lambda, \sigma_\lambda^2), \;\; s = SR, V; \; t = 1978, 1980, 1984, 1986, 1988, 1990;$$
$$\eta_r^s \sim N(\mu_\eta, \sigma_\eta^2), \;\; s = SR, V; \; r = \text{Northeast, North Central, South, West}$$

The regressors included in $\mathbf{x}_i$ are indicators for demographic and socio-economic conditions and political attitudes: *Age*, *Church Attendance*, *Education*, *Female*, *Home owner*, *Income*, *Nonwhite*, *Party Identification* and *Partisan Strength*. A description of the coding used for each of the variables may be found in the Appendix. We should note that, while this specification includes some of the variables most commonly used in models of voter turnout found in the literature (Bernstein, Chadha and Montjoy 2001; Highton 2004; Wolfinger and Rosenstone 1980), it does not examine the effect of other factors we might plausibly believe

could alter turnout, such as political information (Alvarez 1997) or differences in state-level ballot laws (Wolfinger and Rosenstone 1980). The sample used in the analysis consists of 6,411 observations for the 6 elections under study and was constructed so that they are identical for both models. Only the respondents with no missing response or covariate values are included in the analysis. The remaining observations were dropped using list-wise deletion.

Figure 3 presents the main results from both models.[23] The left panel summarizes the posterior distribution of the model's coefficients using self-reported vote as the dependent variable, and the right panel re-does the analysis with the ANES validated vote. Most of the parameter estimates are quite similar in both models, and inferences on the role of these predictors on the probability of voting agree with common expectations. For example, for both sets of estimates, older, wealthier and more educated respondents are more likely to turn out to vote. Also, strong partisans are on average 15 percentage points more likely to vote than independents, while respondents who attend church every week are on average 12 percentage points more likely to turn out to vote than those who never attend. Respondents are much more likely to turn out to vote in Presidential than in Midterm elections, and are less likely to vote if they live in the South. These results are similar using either reported or validated vote as the dependent variable. However, there are some interesting differences between the two sets of results regarding the role of some socio-demographic variables such as gender and race. In particular, the mean posterior of the coefficient for the race indicator is more than twice as large (in absolute value) using validated vote than using self-reported vote as the dependent variable.

These differences in the parameter estimates can affect inferences drawn from both models regarding the impact of the covariates on the turnout decision. In order to illustrate this

[23]Three parallel chains with dispersed initial values reached approximate convergence after 50,000 iterations, with a burn-in period of 5,000 iterations. In order to ensure that inferences are data dependent, several alternative values for the hyperparameters were tried, yielding essentially similar results.

fact, Figure 4 plots the marginal effect of race on the probability of voting using reported and validated vote for the elections under analysis. As seen in the figure, the negative effect of being *Non-white* on turnout is higher when validated vote is used as the response variable for each of the surveys considered. The average marginal effects (posterior means) are more than 6 percentage points higher than if we look only at the reported vote, with differences ranging from about 3 percentage points in the 1984 and 1986 elections to almost 11 points in the 1978 and 1988 elections. While a researcher using reported turnout would conclude that race had no significant effect on the probability of voting in the 1978 and 1988 elections at the usual confidence levels, the results obtained using validated data indicate otherwise.[24] Fitting a model of turnout using reported vote as the dependent variable will therefore tend to overpredict the probability of voting among non-white respondents and might in some cases affect substantive conclusions about the effect of race on turnout.

Finally, we examine whether over-reporting varies systematically with individuals' characteristics, fitting a probit model for $\Pr[\tilde{y}_i = 1|y_i = 0]$. As with the turnout model, the misreport model is fairly simple. The predictors include four variables that have been shown to be strongly correlated with over-reporting in previous studies: *Age*, *Church Attendance*, *Education*, *Non-white*, and *Partisan Strength* (Bernstein, Chadha and Montjoy 2001; Cassel 2003). In addition, we also include three additional covariates aimed at capturing some of the conditions of the interview. The first is an indicator of whether the interview was conducted while the respondent was alone. According to the "social pressures" argument (Loftus 1975), a respondent should be more likely to lie about voting if others will learn of the statement. The other two variables are the interviewers' assessments of the respondents' cooperation and sincerity during the interview.[25] Point and interval summaries of the

[24] In the case of the 1988 election, the marginal effect of Non-white estimated from the self-reported vote is not significant even at the 0.1 level

[25] All interviewers in the 1978 – 1990 ANES were asked to rate the level of cooperation and sincerity of the respondent after the completion of the survey.

posterior distribution of the parameters are presented in Figure 5.

In line with previous analyses, we find that overreporters tend to be more educated, older, more partisan, and are more likely to be regular church attendees. Also, consistent with the results reported in Figures 3 and 4, being *Non-white* has a positive effect on the probability of misreporting vote status: non-whites are on average 0.05 more likely to misreport than their white counterparts, and this effect is significant at the 0.1 level. Several scholars have argued that African Americans and Latinos feel pressured to appear to have voted due to the struggles and sacrifices needed to gain voting rights for their racial or ethnic group (Abramson and Claggett 1986), although recent research has suggested that the relationship between race and over-reporting is much more complex than previously thought and depends on the demographic and geographical context (Bernstein, Chadha and Montjoy 2001).[26] None of the other variables has a statistically significant effect on misreporting at the usual confidence levels. In particular, the interviewers seem unable to pick up a "feeling" that is not otherwise captured by the characteristics observable from the survey. This is probably caused by the fact that very few of the interviewers were willing to rank a respondent as uncooperative and/or insincere.[27]

Hence, the results from these simple models indicate that the probability of misreport-

[26]This relationship between race and vote over-reporting could also be associated with the socio-economic status of the non-white population. If it is the case that nonwhites, who are more concentrated in poorer areas, are more likely to be incorrectly validated or excluded from the validation studies because no records can be found (e.g., due to poorly staffed and maintained election offices), then this result could very well be an artifact. While it is difficult to rule this claim out, addressing this concern is beyond the focus of this paper. Hence, as noted above, we proceed as if the validated data provides "gold-standard" information on turnout, or is at least not subject to systematic bias.

[27]Only 1.3% of all the respondents in the sample were ranked as uncooperative by the ANES interviewers and only 0.7% were deemed to be "often insincere".

ing varies systematically with characteristics we might be interested in, and that failing to account for misreporting may affect parameter estimates and inferences about the determinants of voter turnout drawn from non-validated survey data. Unfortunately, as previously mentioned, the ANES has stopped conducting validation studies due to the cost and difficulty in collecting the data as well as to the fact that few researchers used the validated data. The next three sections allow us to evaluate the performance of our proposed method to correct for misreporting and improve estimates and inference obtained from self-reported turnout. Although our model accounts for the possibility of two types of misreporting, we saw before that virtually no one reports not voting when they did, and thus $\pi_i^{0|1}$ would be poorly estimated (Prescott and Garthwaite 2005). Therefore, in the applications below we will assume that $\pi_i^{0|1} = 0$, and we therefore only need to account for $\pi_i^{1|0}$.

3.2. *Correcting for misreporting using a validation sub-sample*

We first apply our method assuming an internal validation design. As in the simulation exercise in Section 2, we randomly assign half of the respondents in each of the 1978–1990 surveys to be the validation sub-study and ignore the validated data for the remaining respondents. We then used the information from the validated sub-sample to correct for over-reporting in the main sample, equally weighting both datasets. For illustrative purposes, we fit the same turnout and misreport models described in 3.1 for all the ANES studies considered. Nonetheless, as indicated above, the probability of voting is considerably higher in Presidential than in Midterm elections, and it is likely that different factors affect turnout in different election years. More importantly, the patterns of overreporting have also been shown to differ substantially across types of races and election years (Cassel 2003). As a result, the misreport model does not predict over-reporting very well: the mean error rate of the misreport model across election studies is 36%, while a null model that simply predicts that no respondent overreports has an error rate of 31%. The model correctly classifies

64% of the survey respondents in cases, and the mean predicted probability of misreporting averaged across simulations is 0.45; ideally this would be near zero or one for the entire sample. Therefore, while the simulation results from Section 2 suggest that our approach is quite robust to misspecification of the model of misreporting, we note that the performance of our proposed method would benefit from better modeling of the misreport process.

Figure 6 summarizes the posterior distribution of the coefficients of selected regressors estimated using validated, self-reported vote, and corrected self-reports for the two ANES studies with lowest (1978) and largest (1984) percentage of overreporters (see Table 2 in the Appendix). Assuming that the parameters estimated using validated vote are the "correct" ones, the point estimates (posterior means) from our model for the two elections are between 32% and 92% closer to the "true" values of each of the parameters than the estimates ignoring overreporting. In addition, like the "true" estimate, the estimate of $\beta_{Non-white}$ under our approach is significantly negative at the 0.05 level for the 1978 ANES. Figure 7, in turn, plots the marginal effect of race on the probability of voting estimated using our approach to correct for misreporting. A comparison of the results in the left panel of the figure with those presented in Figure 4 above shows that, after correcting for misreporting, the impact of race in the 1978 and 1988 elections is now statistically significant at the usual confidence levels. Moreover, as seen in the right panel of Figure 7, the point estimates from our model are closer to the "true" effects than those obtained from the model using self-reported vote for all the ANES studies, with differences ranging between 1 and 9 percentage points. Therefore, the evidence presented in this Section indicates that, even with the simple model of misreporting estimated here, the improvements in the accuracy of the parameter estimates obtained using our method are important, and can eventually change the substantive conclusions drawn regarding the effect of relevant covariates on the turnout decision.

3.3. *Correcting for misreporting under an external validation design*

We also apply our correction for misreporting assuming an external validation design, ignoring the validated vote for the sample under analysis and incorporating information on the misreport probabilities and regression parameters from other ANES studies. Figure 8 illustrates the results of this exercise, plotting the marginal posterior distribution of selected coefficients for the 1988 and 1992 Presidential elections obtained by updating the corresponding posteriors from previous validated ANES surveys.

The upper panel compares the posterior distributions of $\beta_{Education}$, β_{Income}, $\beta_{Non-white}$ and $\beta_{Partisan\ Strength}$ for the 1988 ANES, the last Presidential election for which vote validation is available, using validated, self-reported and corrected vote. In order to implement our correction for misreporting, we used auxiliary data from the two previous Presidential elections for which validated turnout data was collected (1980 and 1984). As seen in the figure, the marginal posterior means and modes from the model accounting for overreporting are in all cases closer to "true" values than those obtained from the unadjusted self-reports. Again, as the "correct" estimate, the estimate of $\beta_{Non-white}$ under our model is significantly negative at the 0.05 level. In the case of the 1992 ANES, for which there is no validated data, we implemented our correction for misreporting using information from the previous presidential elections for which vote validation was conducted (1980, 1984 and 1988) and compared the estimates from our model with those from a model using self-reported vote. As seen in the lower panel of Figure 8, the posterior distributions of some of the parameters — $\beta_{Education}$ and $\beta_{Partisan\ Strength}$ — remain essentially unchanged when applying the correction for misreporting. However, using auxiliary information does affect the posterior distribution of the coefficients of *Income* and *Non-white*. In particular, accounting for misreporting substantially affects the marginal posterior distribution of $\beta_{Non-white}$. The mean posterior is more than twice as large (in absolute value) when using the corrected self-reports, and the effect of *Non-white* on the probability of turning out to vote is significantly negative at the

0.05 level, while it is not significant even at the 0.2 level when estimated using self-reported vote. Similar results hold when applying our model to correct for misreporting in the 1994 ANES - for which, again, vote validation was not conducted - using validated turnout data from previous Midterm elections.

We also conducted a series of sensitivity analyses aimed at assessing the robustness of the parameter estimates to changes in the composition of the auxiliary data used to correct for misreporting and in the weight assigned to the validated *vis a vis* the main sample. Figure 9 summarizes some of the results for the 1988 and 1992 ANES. The left panel plots point and interval summaries for $\beta_{Non-white}$ from our model for the 1988 ANES using two different sets of values for the weighting parameters δ_d in Equation 9: a point mass prior $\delta_d = 1$ with probability 1 $\forall d$, and uniform $Beta(1,1)$ priors $\forall d$, where $d = 1980, 1984$. In the first case, the validated and main samples are pooled together and the estimates of β for the main sample are obtained by updating the posteriors from the previous ANES surveys *via* Bayes' theorem. In the second case, we allow for different *a posteriori* weights for each of the validated samples, accounting for heterogeneity between the previous ANES studies. The right panel, in turn, compares the estimates from our model for the 1992 survey for the cases in which only validated data from the immediate previous (1988) or from all the previous (1980, 1984, 1988) Presidential elections is used to adjust for misreporting.[28] For both election years, the estimates from our model are compared to those from the unadjusted self-reports.

As illustrated in the figure, the posterior standard deviations of β tend to decrease with the amount of auxiliary data used to correct for misreporting in the main sample, but the point estimates (posterior means) and the main substantive conclusions about β seem to be quite robust to changes in the values of δ and in the size and heterogeneity of the auxiliary data. In particular, correcting for overreporting using information from previous validated studies leads to stronger negative effects of being *Non-white* on the probability of voting than

[28]For the 1992 ANES, we fixed the value of δ at 1 for this sensitivity analysis.

using self-reported vote, with differences of approximately 4 and 9 and percentage points for the 1988 and 1992 ANES, respectively.

3.4. *Accounting for item and unit non-response*

Both applications of our methodology in Sections 3.2 and 3.3 have been based on a complete-case analysis, including in the sample only those respondents for whom both the response to the turnout question and all the relevant covariates are completely observed. When respondents with missing covariates differ systematically from those with complete data with respect to the outcome of interest, this approach may lead to significantly biased estimates and inference (Little and Rubin 2002). In our sample from the 1978–1990 ANES studies, 14.5% of whites and 20.9% of non-whites have missing covariate values (other than race), and the percentage of missingness for the self-reported vote is almost 2 times larger for the latter. Since the evidence above indicates that voting patterns vary systematically with race, inferences from a complete-case analysis may be quite misleading in this setting (Ibrahim et al. 2005). In addition, list-wise deletion due to missing values in the response variable and/or the predictors leads to discard more 40% of the respondents in the 1978–1992 ANES, so that complete-case analyses are extremely wasteful and potentially inefficient. Table 3 in the Appendix reports the rates of item nonresponse for all the variables included in the turnout models estimated in 3.2 and 3.3.

In order to accommodate item and unit non-response, we implement the approach described in Section 1.3, fitting a separate model for each of the ANES studies.[29] Based on Equation 11, we specified probit regression models for all the dichotomous covariates in the model – *Female*, *Non-white*, *Own Home*, and *Alone* – while the remaining categorical covariates were assigned conditional normal distributions and discrete values were afterwards

[29]See Gelman, King and Liu (1998) for an approach to multiple imputation for multiple surveys using hierarchical modeling.

imputed for the missing responses (Gelman, King and Liu 1998).[30] In all cases, we assigned vague independent normal priors for the components of α. Figure 10 illustrates the results for the 1978 and 1992 ANES. For the former, 31% of the survey respondents have at least 1 missing covariate value, and 0.5% of the respondents failed to answer the turnout question. The corresponding rates for the latter are 47% and 9%, respectively. A complete-case analysis would keep 77% of our sample for the 1978 ANES, and only 42% for the 1992 ANES. The left panel of the figure summarizes the marginal posterior distribution of $\beta_{Non-white}$ for the 1978 ANES using reported, validated and corrected vote. As in Section 3.2, our correction for misreporting was implemented based on auxiliary information from a random sub-sample of the ANES survey. The right panel, in turn, plots the estimates for the 1992 ANES, for which we use validated turnout data from the 1988 ANES, as in 3.3. In both cases, estimates obtained using Bayesian imputation are compared to those from the complete-case analyses.

Two interesting facts emerge from the figure. First, for both election-studies, the marginal posterior distribution for $\beta_{Non-white}$ estimated using our the Bayesian imputation model is not statistically different from the that obtained using list-wise deletion, at least at the 0.05 level. However, the standard errors tend to be lower when missing values are imputed than under list-wise deletion. This result holds in fact for most of the election-years under analysis, suggesting that by omitting the cases with missing values, much information is lost on the variables that are completely or almost completely observed, thus leading to less efficient parameter estimates (Ibrahim, Chen and Lipsitz 2002; Ibrahim et al. 2005). This is likely to be an important concern in the Election Studies examined here, given that there is substantial variation in the rates of item nonresponse, with most of the variables exhibiting relatively low percentage of missing values while a few others showing very high rates of nonresponse (see the Appendix). Second, imputing missing values does not change

[30]The substantive results are essentially unchanged if, instead of the normal distributions, one-dimensional conditional gamma distributions are specified for these covariates, all of which are strictly positive.

the substantive findings reported above regarding the performance of our methodology. The results for the 1978 ANES show that the estimated effects from our model correcting for misreporting are again closer to the benchmark case – using validated vote– than the effects estimated using recalled vote. This result holds for the other ANES validated studies as well. For the 1992 election, the marginal effect of race obtained from the corrected turnout model is also higher than in the uncorrected model, as was in the complete-case analysis. For both elections, once again, the main substantive conclusions regarding the effect of being *Non-white* on the probability of voting drawn from the model correcting for misreporting differ from those obtained using recalled vote.

4. CONCLUDING REMARKS

Survey data are usually subject to measurement errors, generally referred to as classification errors when affecting discrete variables. In the political science literature, misclassification of binary dependent variables has received considerable attention in the context of estimating the determinants of voter turnout. High rates of overreporting have been documented in survey instruments commonly used to study turnout in the U.S., such as the American National Election Study (ANES) and the Current Population Survey (CPS), and most previous research has found that misreporting varies systematically with some of the relevant characteristics affecting the turnout decision.

In the presence of misreporting, standard binary choice models will generally yield biased parameter estimates and inaccurate standard errors and may lead to erroneous substantive conclusions. This paper develops a simple Bayesian method to correct for misreporting using information on the misreport mechanism from auxiliary data sources. Our model does not require full validation studies to be conducted every time a researcher is concerned about potential misreporting. As long as enough data exists to reasonably estimate the misreporting probabilities, our approach can be applied for drawing inference from the non-validated

samples, improving the accuracy of the parameter estimates and inferences regarding the effect of covariates of interest on the true response *vis a vis* standard models ignoring misclassification and methods assuming constant misreport rates. This is clearly important, since obtaining "gold-standard" data is usually quite expensive and time consuming, and thus restricting the analysis only to validated studies will generally lead to discard large amounts of useful information, as in the case of the ANES.

The proposed model is fully general and modular, can be easily implemented using freely available software, and can be readily applied in the case of missing data in the response and/or covariates. While we illustrate our technique using turnout data from the ANES, it could be applied in general to account for potential misclassification of a binary dependent variable in many other situations in which auxiliary data on the misreport structure is available. Extensions to more general discrete choice models are also straightforward. Potential avenues for future research would be to use semi- or non-parametric methods to estimate both the misreporting and turnout models, simultaneously account for response and covariate measurement errors within our model, and explore the possibility of incorporating semi-parametric approaches for inference with missing data.

While the primary focus of the paper has been on estimation techniques as opposed to substantive findings, the empirical application of our model to the analysis of the determinants of voter turnout has clear implications for researchers interested in race. Our results confirm that race does have a clear negative impact on turnout, and suggest that the null previous findings have been probably due to problems of misreporting, as had been argued by Abramson and Claggett (1986). With the correction for misreporting developed in this paper, researchers could now better estimate the effect of race over the length of the ANES datasets and not just for the few years with validated turnout data. In addition, researchers might wish to revisit Wolfinger and Rosenstone (1980) findings of the effect of registration laws to see if properly correct misreporting re-enforces or diminishes their findings.

APPENDIX: DESCRIPTION OF THE DATASET USED FOR THE ANALYSIS OF TURNOUT MISREPORTING

Variables used in the turnout model

1. Indicators for demographic, socio-economic and political characteristics

   Age: 1 if $Age < 30$; 2 if $30 \leq Age < 45$; 3 if $45 \leq Age < 60$; 4 if $Age \geq 60$.

   Church Attendance: Frequency of church attendance. Coding: 1 if never; 2 if a few times a year; 3 if once or twice a month; 4 if every week or almost every week.

   Education: Highest grade of school or year of college completed. Coding: 1 if 8 grades or less; 2 if 9–12 grades with no diploma or equivalency; 3 if 12 grades, diploma or equivalency; 4 if some college; 5 if college degree.

   Female: 1 if the respondent is female, 0 if male.

   Home owner: 1 if the respondent owns his house, 0 otherwise.

   Income: Household income. Coding: 1 if 0–16th percentile; 2 if 17h–33d percentile; 3 if 34th–67th percentile; 4 if 68th–95th percentile; 5 if 96th–100th percentile.

   Non-white: 0 if white, 1 otherwise.

   Party Identification: -1 for Democrats, 0 for Independents, 1 for Republicans.

   Partisan Strength: Coded on a four-point scale ranging from 1 for pure independents to 4 for strong partisans.

2. Additional covariates to account for misreporting

   Alone: 1 if the respondent was interviewed alone, 0 otherwise.

   Uncooperative: Respondent's level of cooperation in the interview, as evaluated by the interviewer. Coding: 1 if very good; 2 if good; 3 if fair; 4 if poor; 5 if very poor.

Sincerity: How sincere did the respondent seem to be in his/her answers, as evaluated by the interviewer. Coding: 1 if often seemed insincere; 2 if usually sincere; 3 if completely sincere.

In order to reduce the correlation between the parameters and to accelerate convergence and mixing of the Gibbs sampling algorithm, all variables where centered at their mean values (Gu 2006).

REFERENCES

Abramson, Paul R. and William Claggett. 1986. "Race-Related Differences in Self-Reported and Validated Turnout in 1984." *Journal of Politics* 48:412–422.

Abrevaya, Jason and Jerry A. Hausman. 1999. "Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells." *Annales d'Economie et de Statistique* 55–56:243–276.

Aldrich, John H. 1993. "Rational Choice and Turnout." *American Journal of Political Science* 37(1):246–278.

Alvarez, R. Michael. 1997. *Information and Elections*. Michigan: The University of Michigan Press.

Bernstein, Robert, Anita Chadha and Robert Montjoy. 2001. "Overreporting Voting. Why it Happens and Why it Matters." *Public Opinion Quarterly* 65(1):22–44.

Bound, John, Charles Brown and Nancy Mathiowetz. 2001. *Measurement Error in Survey Data*. Elsevier, North Holland: J. Heckman and E. Leamer (eds.) Handbook of Econometrics, Vol. 5, pp. 3705-3843.

Burden, Barry C. 2000. "Voter Turnout and the National Election Studies." *Political Analysis* 8(4):389–398.

Carroll, Raymond J., David Ruppert and Leonard A. Stefanski. 1995. *Measurement error in nonlinear models*. London: Chapman and Hall.

Cassel, Carol A. 2003. "Overreporting and Electoral Participation Research." *American Politics Research* 31(1):81–92.

Christin, Thomas and Simon Hug. 2004. "Methodological Issues in Studies of Conflict

Processes: Misclassifications and Endogenous Institutions." *Paper presented at the Annual Meeting of the American Political Science Association, Chicago, September 2-5.* .

Congdon, Peter. 2002. *Bayesian Statistical Modelling*. New York: Wiley.

Dunson, David B. and Kenneth R. Tindall. 2000. "Bayesian Anaysis of Mutational Spectra." *Genetics* 156:1411–1418.

Gelfland, Alan E. and Adrian F. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85(410):398–409.

Gelman, Andrew, Gary King and Chuanhai Liu. 1998. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93(443):846–857.

Gu, Yuanyuan. 2006. *Misclassification of the Dependent Variable in Binary Choice Models*. Australia: Masters' Dissertation, School of Economics, University of the New South Wales.

Härdle, Wolfgang. 1990. *Applied nonparametric regression*. New York: Cambridge University Pres.

Hausman, Jerry A., Jason Abrevaya and Fiona M. Scott-Morton. 1998. "Misclassification of the dependent variable in a discrete response setting." *Journal of Econometrics* 87(2):239–269.

Highton, Benjamin. 2004. "Self-reported versus Proxy-reported Voter Turnout in the Current Population Survey." *Public Opinion Quarterly* 69(1):113–123.

Horowitz, Joel L. 1993. *Semiparametric and Nonparametric Estimation of Quantal Response Models*. Elsevier, Amsterdam: G.S. Maddala, C. R. Rao, and H. D. Vinod (eds), Handbook of Statistics, Vol. 11, pp. 45–72.

Horowitz, Joel L. and Charles F. Manski. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica* 63(2):281–302.

Hu, Yingyao. 2008. "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics* 144:27–61.

Ibrahim, Joseph G. and Min-Hui Chen. 2000. "Power Prior Distributions for Regression Models." *Statistical Science* 15(1):46–60.

Ibrahim, Joseph G., Ming-Hui Chen and Stuart R. Lipsitz. 2002. "Bayesian Methods for Generalized Linear Models With Covariates Missing at Random." *Canadian Journal of Statistics* 30:55–78.

Ibrahim, Joseph G., Ming-Hui Chen, Stuart R. Lipsitz and Amy H. Herring. 2005. "Missing-Data Methods for Generalized Linear Models: A Comparative Review." *Journal of the American Statistical Association* 100(469):332–346.

Jackman, Simon. 1999. "Correcting surveys for non-response and measurement error using auxiliary information." *Electoral Studies* 18:7–27.

Katosh, John P. and Michael W. Traugott. 1981. "The Consequences of Validated and Self-Reported Voting Measures." *Public Opinion Quarterly* 45:519–535.

Katz, Jonathan and Gabriel Katz. 2009. *Evaluating alternative models to account for misclassified dependent variables in binary choice models.* Unpublished manuscript.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical analysis with missing data.* New York: Wiley.

Loftus, Elizabeth F. 1975. "Leading Questions and the Eyewitness Report." *Cognitive Psychology* 7:145–177.

Molinari, Francesca. 2003. *Contaminated, Corrupted and Missing Data.* Evanston, IL: Doctoral Dissertation, Department of Economics, Northwestern University.

Nagler, Jonathan. 1994. "Scobit: An Alternative Estimator to Logit and Probit." *American Journal of Political Science* 38(1):230–255.

Neuhaus, John M. 1999. "Bias and efficiency loss due to misclassified responses in binary regression." *Biometrika* 86(4):843–855.

Paulino, Carlos D., Paulo Soares and John Neuhaus. 2003. "Binomial Regression with Misclassification." *Biometrics* 59:670–675.

Plummer, Martyn. 2009. *JAGS version 1.03 manual*. www-ice.iarc.fr/~/software/jags/.

Prescott, Gordon J. and Paul H. Garthwaite. 2002. "Simple Bayesian Analysis of Misclassified Binary Data with a Validation Substudy." *Biometrics* 58:454–458.

Prescott, Gordon J. and Paul H. Garthwaite. 2005. "Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study." *Statistics in Medicine* 24(3).

Robert, Christian P. and George Casella. 2004. *Monte Carlo Statistical Methods*. New York: Springer.

Sigelman, Lee. 1982. "The Nonvoting Voter in Voting Research." *American Journal of Political Science* 26:47–56.

Spiegelhalter, David J., Andrew Thomas and Nicky G. Best. 2003. *WinBUGS, Version 1.4. User Manual*. Cambridge, UK: University of Cambridge.

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

Zhao, Zong. 2008. "Sensitivity of Propensity Score Methods to the Specifications." *Economics Letters* 98(3):309–319.

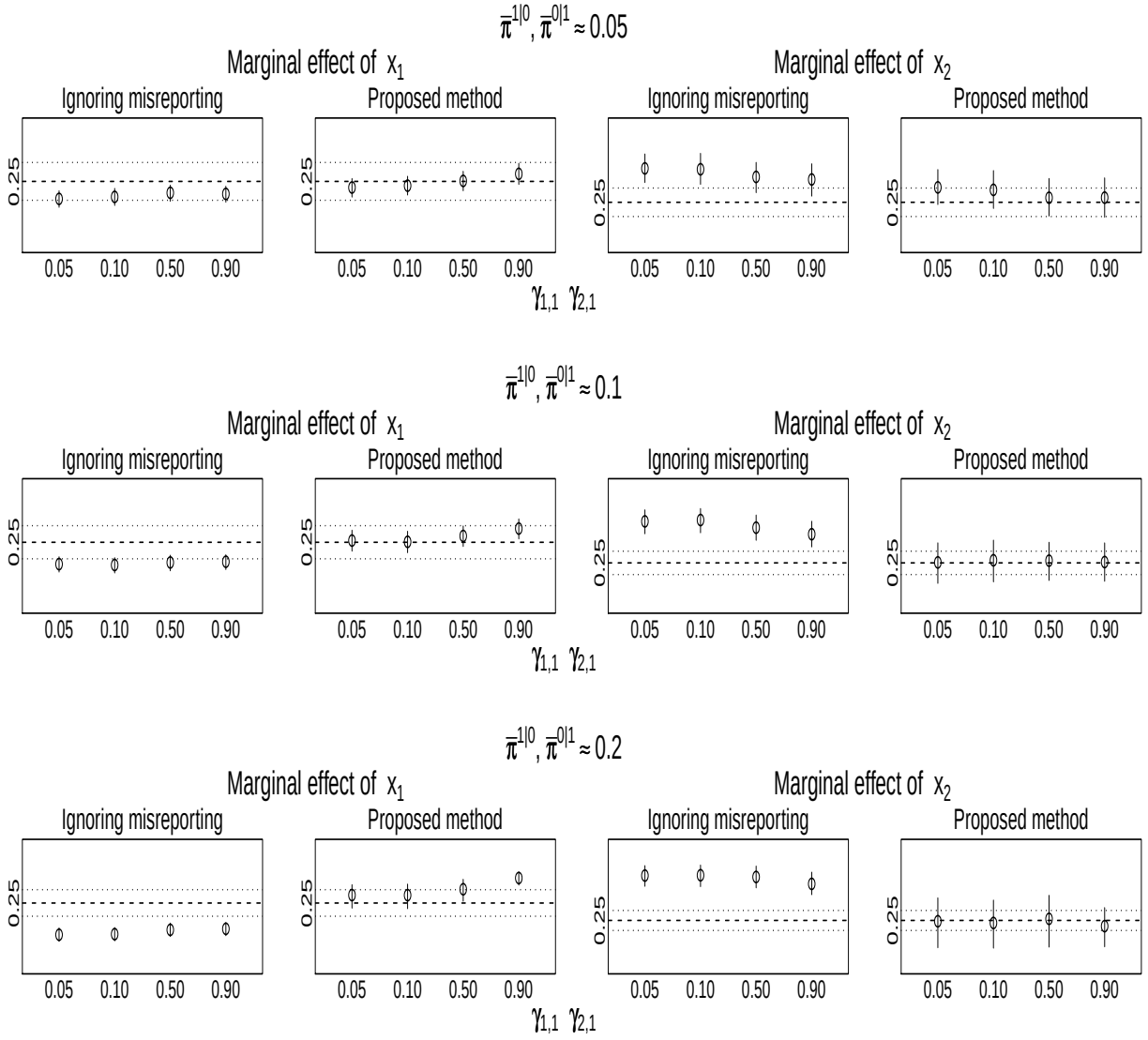

Figure 1: ***Marginal covariate effects when*** x_1 ***is omitted from the misreport model.*** *The graph plots the marginal effects* x_1 *and* x_2 *estimated under our method when* x_1 *is omitted from the linear predictor of the misreport model, for different values of* γ_1 *and* γ_2*. Results are compared to those obtained ignoring misclassification. The center dots correspond to the the posterior means, the vertical lines to the central 95% credible intervals, and the horizontal lines represent the average effects (dashed) and 95% intervals (dotted) estimated using* y_i *as the response.*

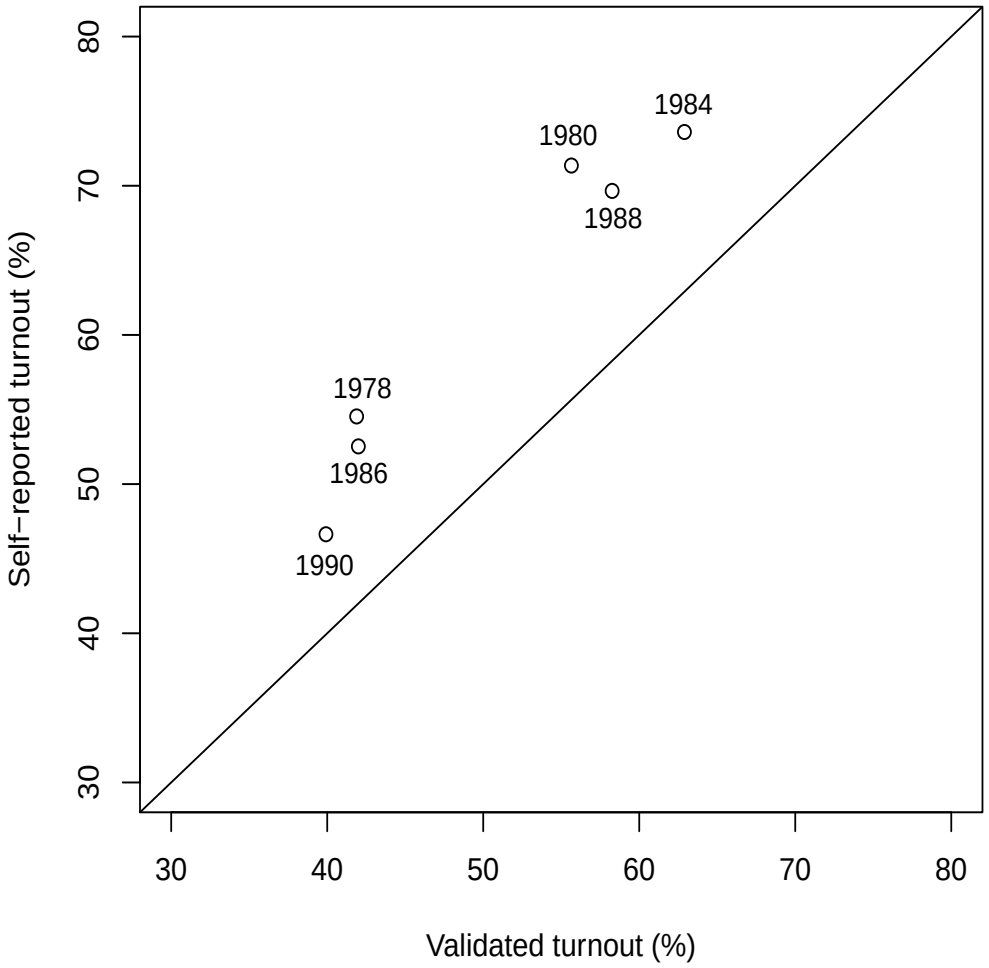

Figure 2: ***Self-reported vs. Validated Turnout, 1978–1990.*** *The graph shows the self-reported and validated turnout from the 1978–1990 ANES only in years for which there were vote validation studies. Reported turnout rates are systematically larger than the validates ones.*

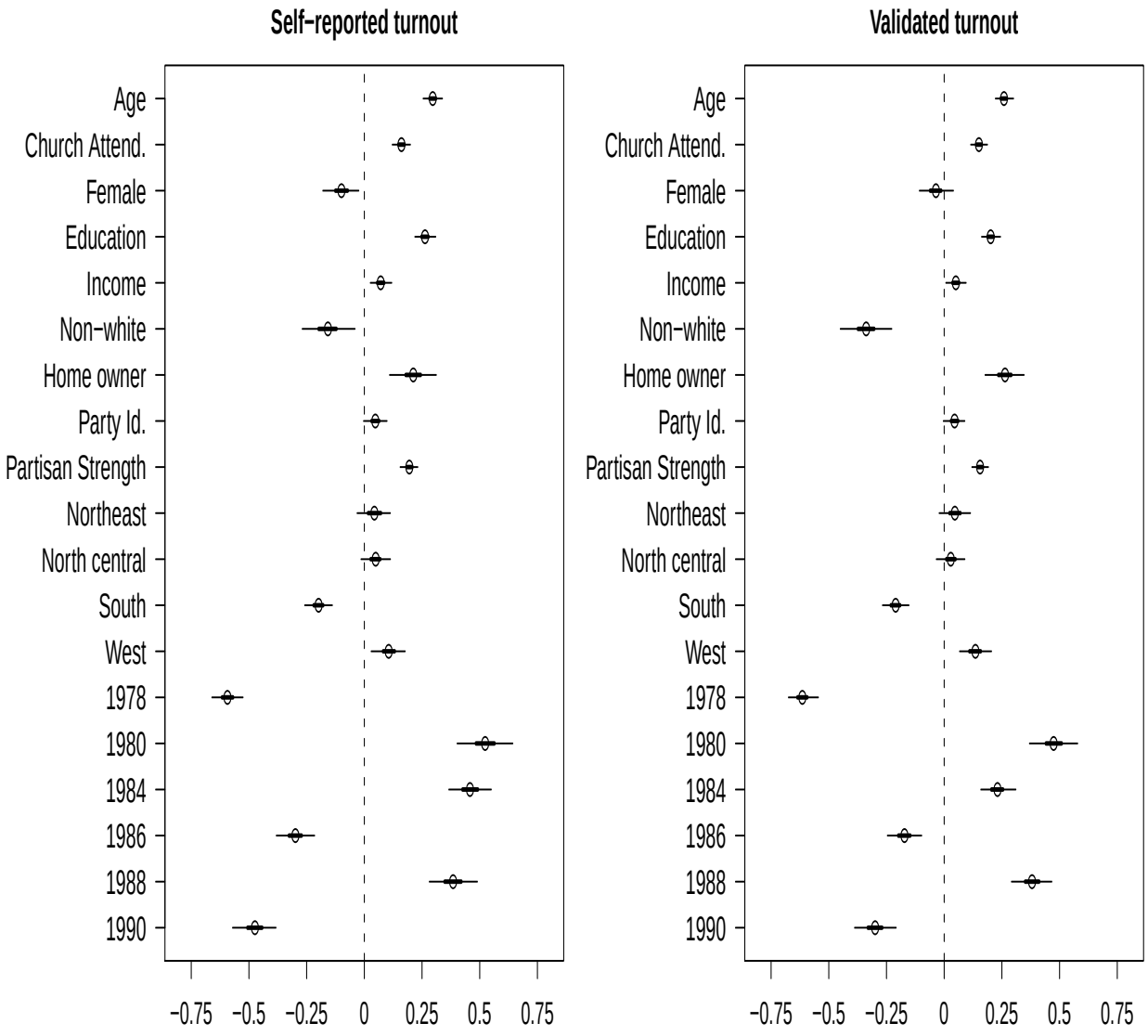

Figure 3: ***Coefficients of the probit models for Self-reported vs. Validated Turnout.*** *The graph summarizes the posterior distribution of the coefficients of the turnout model, using self-reported and validated vote as the response variable. The center dots correspond to the posterior means, the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.*

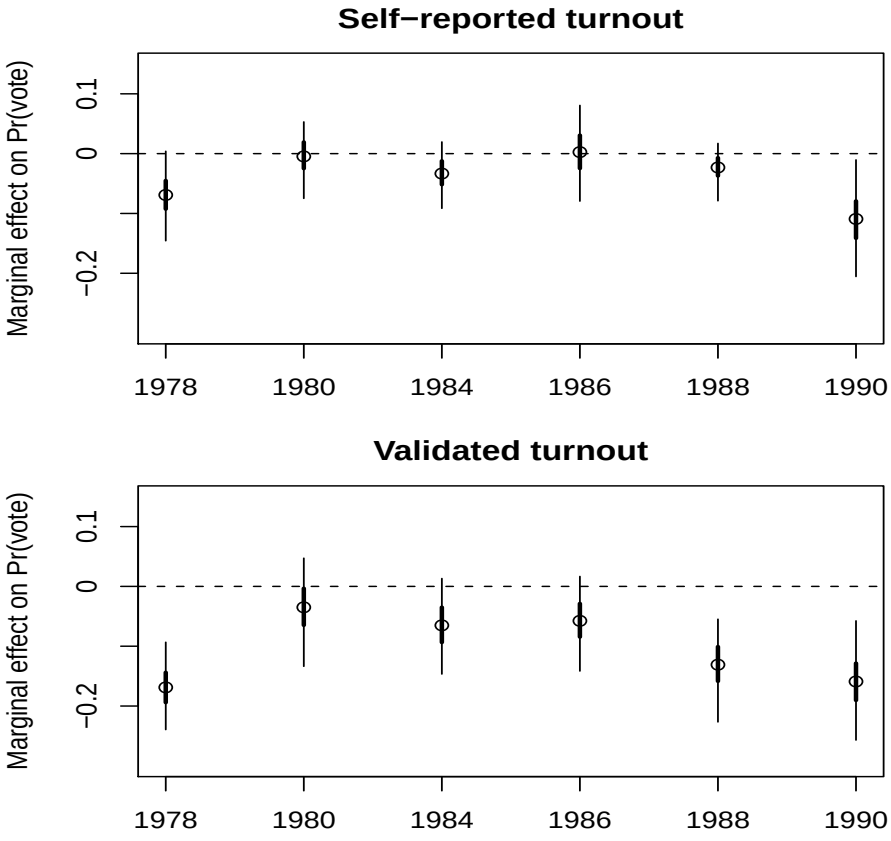

Figure 4: ***Marginal effect of race on turnout.*** *The graph shows the marginal effect of the race indicator on the likelihood of voting for each election year under study, using both reported and validated vote. The center dots correspond to the point estimates (posterior means), the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.*

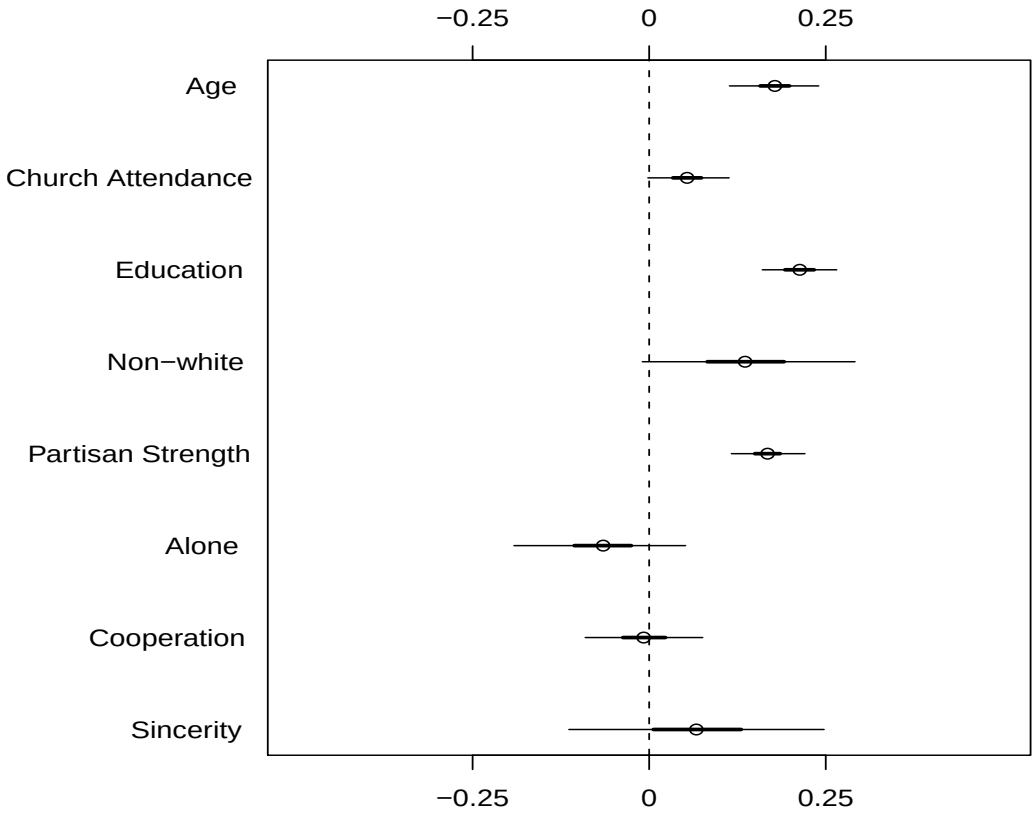

Figure 5: ***Determinants of misreporting.*** *The graph shows the parameter estimates for the model of over-reporting. The center dots correspond to the point estimates (posterior means), the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.*

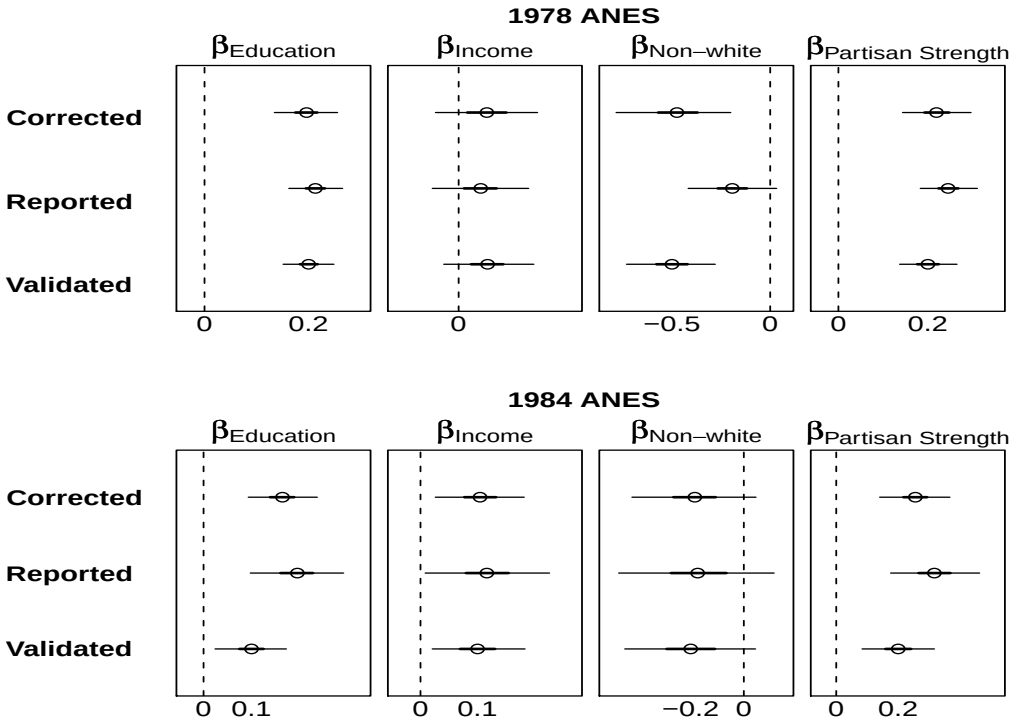

Figure 6: ***Posterior summaries for selected parameters under an internal validation design.*** *The figure plots point and interval summaries of the posterior distributions of selected coefficients for the 1978 and 1984 ANES Presidential elections, using corrected, self-reported, and validated vote. The center dots correspond to the posterior means, the thick horizontal lines to the central 50% credible intervals, and the thin lines to the central 95% credible intervals from the three different models.*

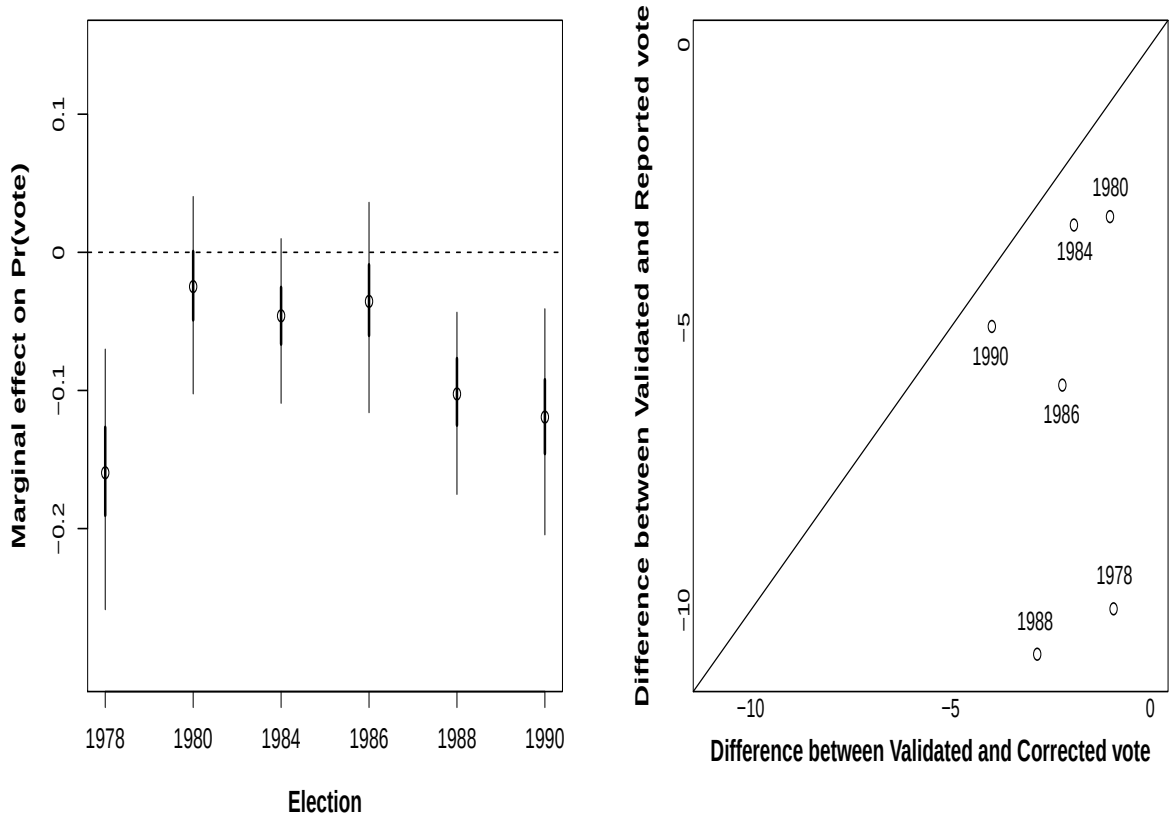

Figure 7: ***Marginal effect of race on turnout estimated under our proposed method.*** *The left panel of the graph plots the point and interval (50% and 95%) estimates of the marginal effect of race on the probability of voting obtained using our method to correct for misreporting. The right panel compares the point estimates from our model and the model ignoring misreporting with the estimates obtained using the validated data.*

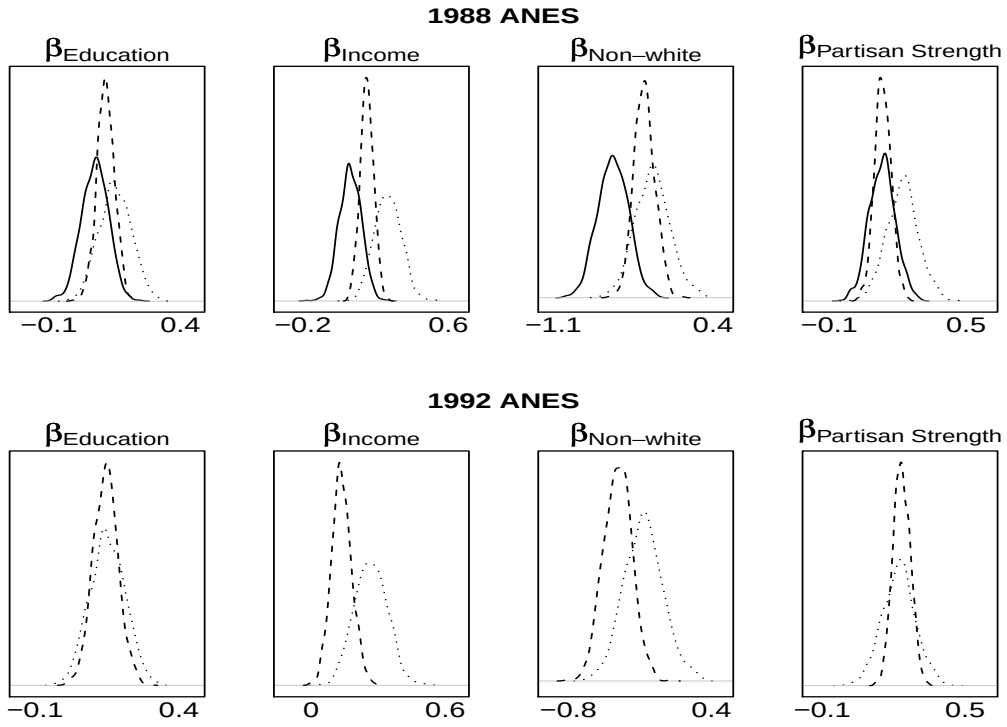

Figure 8: ***Posterior densities of*** β ***under an external validation design.*** *The figure compares the posterior densities of selected coefficients for the 1988 and 1992 Presidential elections. The solid lines plot the posterior distributions of the parameters estimated from the validated vote, the dotted lines represent the estimates obtained using self-reported vote, and the dashed lines the ones obtained adjusting for misreporting.*

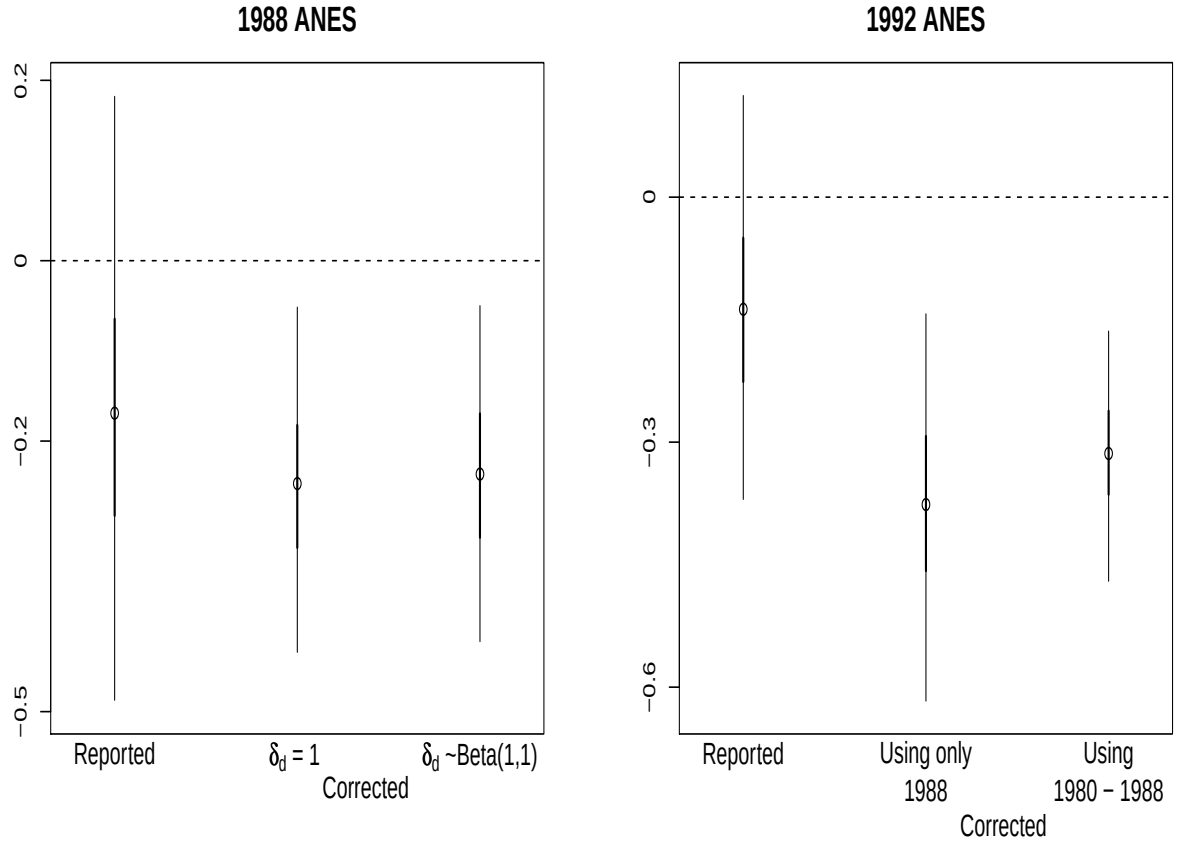

Figure 9: ***Sensitivity analysis for the external validation design.*** *The graph summarizes the posterior distribution of* $\beta_{Non-white}$ *from our model for the 1988 and 1992 elections, using alternative strategies to incorporate information from previous validated ANES studies. The estimates are compared to those obtained using self-reported vote. The center dots correspond to the posterior means, the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.*

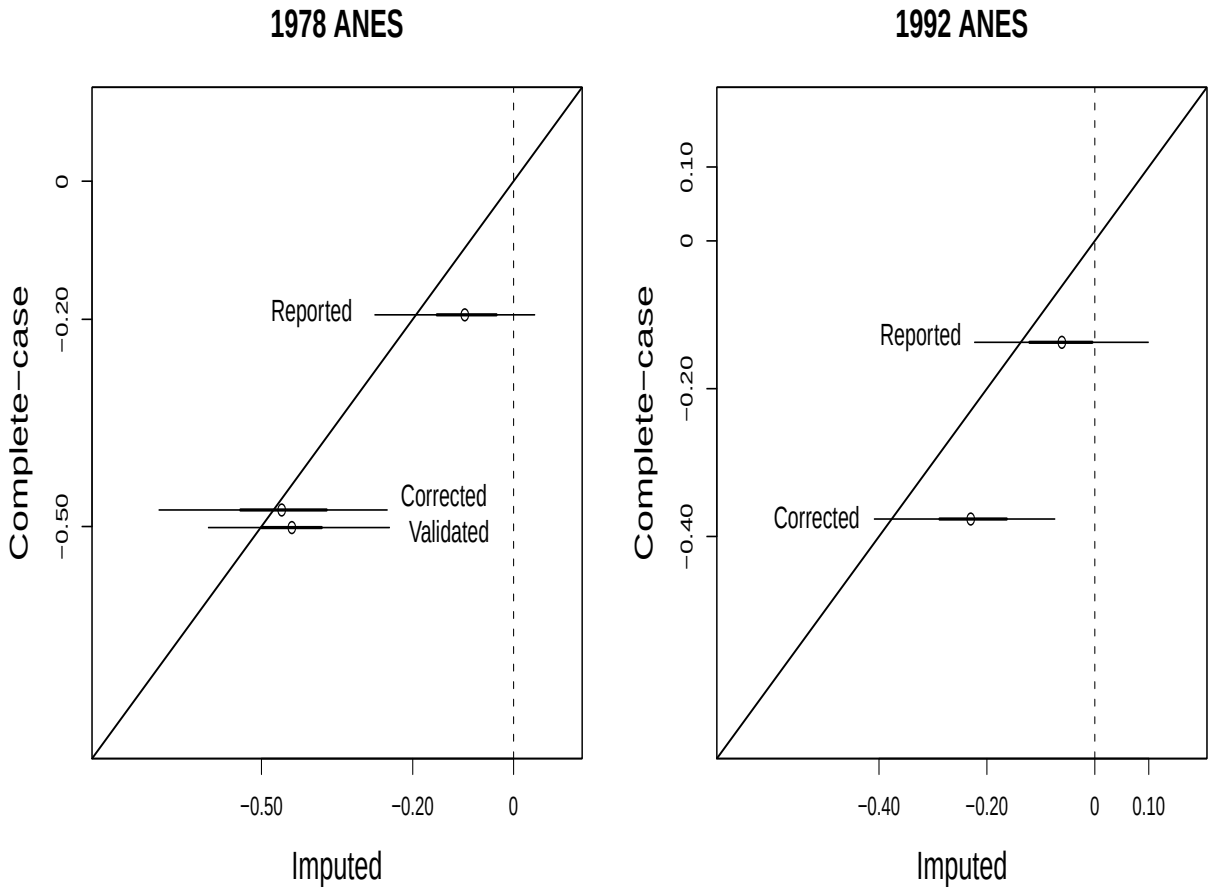

Figure 10: ***Posterior summaries for*** $\beta_{Non-white}$ ***with list-wise deletion versus Bayesian imputation.*** *The graph plots point and interval summaries for* $\beta_{Non-white}$ *for the 1978 and 1992 ANES, using list-wise deletion and fully Bayesian imputation. The center dots correspond to the point estimates (posterior means), and the horizontal bars indicate the 90% and 50% credible intervals for the models with imputed missing values.*

Table 1: Marginal covariate effects under alternative specifications of the misreport model

	Estimator	$\partial Pr(y=1\|\mathbf{x})/\partial x_1$	$\partial Pr(y=1\|\mathbf{x})/\partial x_2$
	True Model	0.25 (0.24, 0.27)	0.25 (0.21, 0.30)
Linearpredictor[a]	Specification 2	0.29 (0.24, 0.33)	0.26 (0.13, 0.38)
	Specification 3	0.29 (0.24, 0.32)	0.27 (0.15, 0.38)
	Specification 4	0.29 (0.24, 0.33)	0.26 (0.15, 0.38)
Error disturbance	Logistic distribution[b]	0.27 (0.21, 0.32)	0.23 (0.09, 0.38)
	Bimodal distribution[c]	0.24 (0.20, 0.28)	0.21 (0.10, 0.29)
	Heteroskedastic[d]	0.24 (0.17, 0.29)	0.28 (0.17, 0.39)
Different misreport models in both sub-samples[e]		0.24 (0.21, 0.28)	0.30 (0.21, 0.38)

[a] $\gamma_{1,0} = -1.5, \gamma_{1,1} = 0.05, \gamma_{1,2} = 1.25, \gamma_{2,0} = -0.2, \gamma_{2,1} = 0.05, \gamma_{2,2} = -1.25, \bar{\pi}^{1|0}, \bar{\pi}^{0|1} \approx 0.2.$

[b] $\gamma_{1,0} = -1.75, \gamma_{1,1} = 0.65, \gamma_{1,2} = 1.3, \gamma_{2,0} = -0.75, \gamma_{2,1} = 0.20, \gamma_{2,2} = -1.3, \bar{\pi}^{1|0}, \bar{\pi}^{0|1} \approx 0.2.$

[c] $\gamma_{1,0} = -1.6, \gamma_{1,1} = 0.5, \gamma_{1,2} = 1.3, \gamma_{2,0} = -1, \gamma_{2,1} = 0.5, \gamma_{2,2} = -1.30, \bar{\pi}^{1|0}, \bar{\pi}^{0|1} \approx 0.1.$

[d] $\gamma_{1,0} = -2.05, \gamma_{1,1} = 0.95, \gamma_{1,2} = 0.1, \gamma_{2,0} = -1.5, \gamma_{2,1} = -2.5, \gamma_{2,2} = -0.70, \bar{\pi}^{1|0} \approx 0.1, \bar{\pi}^{0|1} \approx 0.2.$

[e] $\bar{\pi}^{1|0}, \bar{\pi}^{0|1} \approx 0.1$

Validation sample: $\gamma_{1,0} = -1.8, \gamma_{1,1} = 0.52, \gamma_{1,2} = 1.3, \gamma_{2,0} = -1.1, \gamma_{2,1} = 0.5, \gamma_{2,2} = -1.3.$

Main sample: $\gamma_{1,0} = -2.14, \gamma_{1,1} = 0.89, \gamma_{1,2} = 1.74, \gamma_{2,0} = -1.22, \gamma_{2,1} = 0.76, \gamma_{2,2} = -1.32.$

Table 2: Vote misreporting in 1978–1990 ANES[a]

Election	$P(\tilde{y}_i = 1 \mid y_i = 0)$	$P(y_i = 0 \mid \tilde{y}_i = 1)$	$P(\tilde{y}_i = 0 \mid y_i = 1)$	$P(y_i = 1 \mid \tilde{y}_i = 0)$
1978	23.27	24.55	3.02	2.84
1980	24.48	16.52	0.58	1.37
1984	38.83	13.63	0.22	1.70
1986	31.55	17.70	0.66	1.40
1988	36.30	14.63	1.06	7.10
1990	26.83	16.83	3.67	6.46

[a] In percentage points.

Table 3: Rates of nonresponse for the variables included in the voter turnout models

Variable	**1978-1990 validated ANES**	**1992 ANES**
Age	2.07	0.00
Church Attendance	13.20	33.72
Education	0.80	2.61
Female	4.28	0.00
Income	13.58	10.66
Non-white	4.41	1.41
Home owner	0.70	6.44
Partisan Strength	4.44	0.56
Party Identification	2.60	0.36
Alone	4.55	1.57
Cooperation	4.49	0.16
Sincerity	0.47	0.24
Reported turnout	6.12	9.30
Total sample	11,632	2,485
Complete-case sample	6,411	1,206