Journal of Econometrics 208 (2019) 487–506

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

On the estimation of treatment effects with endogenous misreporting[☆]

Pierre Nguimkeu [a,*], Augustine Denteh [b], Rusty Tchernis [a,c]

[a] *Georgia State University, United States*
[b] *Harvard Medical School & Tulane University, United States*
[c] *NBER & IZA, United States*

ARTICLE INFO

Article history:
Received 2 August 2016
Received in revised form 8 December 2017
Accepted 14 October 2018
Available online 5 November 2018

JEL classification:
C35
C51

Keywords:
Treatment effect
Misclassification
Endogeneity
Binary regressor
Partial observability
Bias

ABSTRACT

Participation in social programs is often misreported in survey data, complicating the estimation of treatment effects. We propose a model to estimate treatment effects under endogenous participation and endogenous misreporting. We present an expression for the asymptotic bias of both OLS and IV estimators and discuss the conditions under which sign reversal may occur. We provide a method for eliminating this bias when researchers have access to information regarding participation and misreporting. We establish the consistency and asymptotic normality of our proposed estimator and assess its small sample performance through Monte Carlo simulations. An empirical example illustrates the proposed method.

1. Introduction

This paper proposes a solution to the problem of identification and estimation of treatment effects in parametric regressions when participation is endogenously misreported. In particular, we provide a two-step estimation procedure that consistently estimates the conditional average treatment effect. Participation in social programs is substantially misreported in survey data, sometimes with misreporting levels as high as 50% (Meyer et al., 2009). When a binary regressor is misreported (or misclassified), the measurement error is necessarily negatively correlated with the true underlying value of the regressor, thus making the classical measurement error assumptions implausible.[1] While earlier papers (Aigner, 1973; Lewbel, 2007) show that exogenous misreporting leads to attenuation bias, we demonstrate that the effects of endogenous misreporting are much more severe. To our knowledge, this paper is the first attempt to provide point estimates of treatment effects in the context of endogenous misreporting of a binary treatment variable.

[☆] We thank Jason Abrevaya, Victor Chernozhukov, Steven Durlauf, William Greene, Kei Hirano, Arthur Lewbel, Nikolas Mittag, Robert Moffitt, Tom Mroz, Alistair O′Malley, Dale Poirier, Seth Richards-Shubik, Gary Solon, Jeffrey Wooldridge, seminar participants at Bank of Canada, Dalhousie University, University of Paris 10, University of California Berkeley, University of Nevada-Las Vegas, Northwestern University, as well as conference participants at the Tinbergen Institute, IZA, CERGE-EI, the Southern Economic Association, and the American Society of Health Economists, for helpful comments.

[*] Correspondence to: Department of Economics, Andrew Young School of Policy Studies, Georgia State University, P.O. Box 3992, Atlanta, GA 30302-3992, United States.

E-mail address: nnguimkeu@gsu.edu (P. Nguimkeu).

[1] For empirical papers that discuss non-classical measurement errors with continuous explanatory variables, see, e.g., Stephens and Unayama (2018), Haider and Solon (2006) and the references therein.

https://doi.org/10.1016/j.jeconom.2018.10.005
0304-4076/© 2018 Elsevier B.V. All rights reserved.

Misreporting occurs when program participants report not receiving treatment when they did ("false negatives") or vice versa ("false positives"). One-sided misreporting (i.e., the occurrence of only one of the two types of misclassification errors) is pervasive in practice and many empirical studies. For example, Lynch et al. (2007) and Meyer et al. (2018) report that validation studies always find high rates of false negatives in the Supplemental Nutrition Assistance Program (SNAP) ranging from 20% to 50%, depending on the survey and period.[2] While Meyer et al. (2018) finds up to 50% rate of false negatives in SNAP participation in the CPS Annual Social and Economic Supplement, false positives are typically low with only less than 1.5% of non-recipients reporting SNAP receipt.

One-sided misreporting is not confined to government programs. For example, according to Bound (1991), there are a number of reasons to be suspicious of any survey response to questions concerning self-evaluated health, not only because respondents are being asked for subjective judgments, but also because responses may be endogenous to the outcomes we may wish to use them to explain. Brachet (2008) argues that in health-related surveys, self-reported smoking status is significantly misreported, with false negatives ranging from 3.4% in some studies to 73% in others. Other instances of one-sided misreporting can be found in the development literature where a firm's formality status is often misreported, with informal firms more likely to falsely report their status (see Gandelman and Rasteletti, 2017), or the education literature where misclassification error is more likely to arise from overreporting of qualifications (e.g. Battistin and Sianesi, 2011).

Recognizing the documented evidence of misclassification errors, a related literature considers the consequences of measurement errors in a binary regressor in Monte Carlo studies. For instance, in studying the worst-case bounds of regression coefficients under arbitrary misclassification of a binary regressor, Kreider (2010) finds that even with misclassification error rates of less than 2%, the confidence intervals from the contaminated data that the researcher observes and the true, error-free data do not overlap. Similarly, Millimet (2011) studies the performance of several estimators employed in the causal inference literature while introducing measurement error in the treatment (binary) regressor, outcome variable, and vector of covariates, and cautions researchers to be conscious of the consequences of not addressing measurement error.

The existing literature has focused on accounting for random (exogenous) misreporting when participation is exogenous. For instance, Aigner (1973) considers misclassification in exogenous binary regressors, shows that OLS estimates are biased downwards, and proposes a technique based on knowledge of the misclassification probabilities to consistently estimate the parameters of interest. More recently, Lewbel (2007) examines the identification and estimation of the treatment effect of a misclassified binary regressor in nonparametric and semiparametric regressions. Lewbel reaches the same attenuation-bias result that Aigner (1973) finds and introduces assumptions that identify the conditional average treatment effect of the misclassified binary regressor. Related works by Bollinger (1996), Black et al. (2000) and van Hasselt and Bollinger (2012) provide partial identification bounds in the linear regression model, while Chen et al. (2008a,b), provide identification in the nonparametric regression model.

Some attempts have been made to address exogenous misreporting when treatment selection (participation) is endogenous. In the education literature, Kane et al. (1999) address misreporting when estimating returns to schooling by proposing a generalized method of moments (GMM) estimator that relies on the existence of two categorical reports of educational attainment, and so may have limited applicability. In estimating the effects of maternal smoking on infant health, Brachet (2008) proposes a two-step GMM estimator, that essentially follows Hausman et al. (1998) and Kane et al. (1999). An admitted weakness of Brachet's approach is the assumption that misreporting probabilities are independent of covariates, conditional on treatment status. Frazis and Loewenstein (2003) and Mahajan (2006) study identification with the usual IV assumptions under homogenous and heterogenous treatment effects due to observables, respectively. More recently, DiTraglia and García-Jimeno (2017) derive a sharp identified set under standard first-moment assumptions and propose a Bonferroni-based procedure for identification robust inference. Also, Bollinger and van Hasselt (2017) use a Bayesian approach while Ura (2018) allows for heterogeneous treatment effects due to unobeservables in these models.

Much less is known about the case in which the regressor and the measurement error are both endogenous. Hu et al. (2015, 2016) provide identification results in a class of nonseparable index models with measurement error and endogeneity. Kreider et al. (2012) is the most closely related to our work, in the sense that they allow for both treatment endogeneity and endogenous measurement error in the case of binary treatment as we do in this paper. In estimating the effect of SNAP on health outcomes, they use auxiliary administrative data on the size of SNAP caseloads to address misreporting by bounding the average treatment effect under increasingly stronger assumptions. While this partial identification approach identifies treatment effects with their tightest bounds, it does not yield point estimates. As such its relevance for policy making may not be widespread.

This paper has three salient contributions. First, we propose a parametric model of endogenous misreporting and endogenous participation. We only analyze the case of one-sided misreporting at this stage, which is the predominant case of misreporting described in Meyer et al. (2009). Second, we show that when misreporting is endogenous, OLS and IV estimators are inconsistent and OLS estimates can be of opposite signs from the true effects (sign reversal), whether participation is endogenous or not. We provide theoretical expressions for these biases under the normality assumption as well as Monte Carlo simulation evidence. Third, we propose an estimator that is root-n consistent and asymptotically normal and show that it performs remarkably well in small samples.

[2] Misreporting has also been documented for other government programs; see, e.g., Marquis and Moore (1990) for an earlier validation study discussing measurement error in the reports of participation in eight government transfer programs in the 1984 Survey of Income and Program Participation (SIPP).

We illustrate the use of our approach in an empirical application. Identification in our framework relies on the existence of both an additional random variable that is correlated with the true (unobserved) underlying treatment, but unrelated to the outcome and the misclassification error (e.g., Frazis and Loewenstein, 2003; Mahajan, 2006), as well as a random variable that is correlated with the misclassification error and needs not be excludable from the outcome. In addition, we assume that the observed treatment probability is a joint (known) function of the treatment and misclassification probability. This allows us to pin down the marginal distribution of true participation and thus point-identify the treatment effect.

The rest of the paper is organized as follows. Section 2 presents the model of endogenous misreporting and shows the inconsistency of OLS and IV estimators. Section 3 develops the proposed estimator. Section 4 provides Monte Carlo simulations; Section 5 contains an empirical application, and Section 6 concludes. Proofs and other technical material are collected in the appendix.

2. Framework

This section describes the proposed model and associated framework, and presents our estimation strategy.

2.1. Model with endogenous misreporting

Consider the following specification of the usual treatment effects model. The outcome variable, y_i, is related to the k–vector of correctly measured exogenous covariates, x_i, and the (true) participation indicator, δ_i^*, by

$$y_i = x_i'\beta + \delta_i^*\alpha + \epsilon_i, \tag{1}$$

and we model participation as

$$\delta_i^* = \mathbf{1}\left(z_i'\theta + v_i \geq 0\right), \tag{2}$$

where α is a scalar capturing the treatment effect of interest, β and θ are parameter vectors of sizes $k \times 1$ and $q \times 1$ respectively, z_i is a q-vector of exogenous variables that includes x_i as well as additional instruments that are unrelated to ϵ_i. In this model, the (possible) endogeneity of participation is captured by the correlation between the error terms ϵ_i and v_i.

However, the researcher does not observe the true participation indicator δ_i^* but only a possibly misclassified surrogate, δ_i, contaminated by a misreporting, unobserved dummy variable, d_i, such that $\delta_i = \delta_i^* d_i$. In other words, an individual correctly reports receiving treatment only if $d_i = 1$ and reports not receiving treatment otherwise. We assume that the indicator of (mis)reporting behavior, d_i, is related to a p-vector of observable covariates w_i such that

$$d_i = \mathbf{1}\left(w_i'\gamma + u_i \geq 0\right), \tag{3}$$

where γ is a parameter vector of size $p \times 1$ and u_i is the error term.[3] Hence, the observed participation, δ_i, can be modeled by

$$\delta_i = \delta_i^* d_i = \mathbf{1}\left(z_i'\theta + v_i \geq 0,\ w_i'\gamma + u_i \geq 0\right). \tag{4}$$

Our modeling of misreported participation is generally in the spirit of a broader class of methods that have been developed for incomplete data scenarios and specifically similar to partial observability models studied in Poirier (1980). Partial observability models such as Poirier's have been widely applied in many fields of study, including Feinstein's examination of the problem of incomplete detection of violations of laws and regulations (Feinstein, 1990).

For the estimation, no further restrictions are imposed on x. However, we require the covariates z and w to be different but possibly overlapping and to have sufficient variation (e.g., at least one covariate in z and w is continuous) to avoid the local identification problems discussed in Poirier (1980). We also make the following basic assumptions, some of which are standard in the literature.

Assumption 1. The error term ϵ_i is independent of the exogenous variables x_i, z_i, with variance σ^2; and the error terms (u_i, v_i) are independent of all covariates x_i, z_i, w_i, and have unit variances. The correlations for the pairs (ϵ_i, u_i), (ϵ_i, v_i) and (u_i, v_i) are denoted φ_u, φ_v and ρ, respectively.

Assumption 2. The $k \times k$ matrix, $\mathbb{E}(x_i x_i')$, is nonsingular (and hence finite).

Assumption 3. The joint CDF of $(-u_i, -v_i)$ is known and defined by

$$F_{u,v}(\underline{u}, \underline{v}, \rho) = \Pr[-u_i \leq \underline{u},\ \ -v_i \leq \underline{v}], \qquad \text{for any } -\infty < \underline{u},\ \underline{v} < +\infty.$$

In particular, we assume that conditional on z_i and w_i, $(-u_i, -v_i)$ follows a bivariate normal distribution.

[3] Note that true participation, δ^*, affects misreporting to the extent that we focus on one-sided misreporting and the error terms in Eqs. (2) and (3) are correlated. This allows, for example, misreporting to be influenced by the stigma experienced by true participants in social programs.

Assumption 4. The error terms, (ϵ_i, u_i, v_i), follow a trivariate normal distribution, conditional on all covariates x_i, z_i, w_i. That is,

$$(\epsilon_i, u_i, v_i)' \,|x_i, z_i, w_i \sim N\left(0, \Sigma\right), \quad \text{with} \quad \Sigma = \begin{pmatrix} \sigma^2 & \varphi_u\sigma & \varphi_v\sigma \\ \varphi_u\sigma & 1 & \rho \\ \varphi_v\sigma & \rho & 1 \end{pmatrix}. \tag{5}$$

Assumptions 1 and 2 are quite standard. However, it is important to notice that unlike x and z, the exogeneity requirement does not apply to w, the vector of covariates associated with misreporting in Eq. (3). This could be of substantial interest in practice where exogenous covariates are often difficult to find. Assumption 3 is critical to parametrically identify the probability of true (unobserved) participation. While we assume joint normality of the disturbance terms in the observed participation equation for simplicity (as in Poirier, 1980), normality is not needed and the following discussion would hold for other absolutely continuous distributions (e.g., the bivariate logistic distributions discussed in Gumbel, 1961). Assumption 4 is only needed to derive closed-form formulas for the OLS bias (see Section 2.2) and for extensions to binary choice models and full information maximum likelihood (see Appendix B); but it is not essential for the rest of our main discussions.

Our estimation strategy relies on observing z and w. We recognize that exclusion restrictions for participation as well as relevant predictors for misreporting may be difficult to obtain in practice and our suggestion is to rely on different data sources. For instance, exclusion restrictions for participation, z, may come from qualification laws (eligibility requirements) for program participation. Relevant predictors of misreporting, w, could include peculiar features of the survey in question and its administration such as survey date, length of the survey, and interview mode.

2.2. Bias due to endogenous misreporting

We first show that a naive OLS estimator of the treatment effect is biased and may assume a sign opposite to the true effect. Since the true participation status, δ_i^*, is unobserved and only δ_i is observed, the model with reported participation status estimated by the researcher is given by

$$y_i = x_i'\beta + \delta_i\alpha + \varepsilon_i. \tag{6}$$

Given the true outcome equation defined by Eq. (1), Eq. (6) implies that

$$\varepsilon_i = \epsilon_i + \left(\delta_i^* - \delta_i\right)\alpha. \tag{7}$$

For a random sample of size n, Eq. (6) can be re-written in the matrix form as follows:

$$y = X\beta + \delta\alpha + \varepsilon, \tag{8}$$

where $y = [y_1, \ldots, y_n]'$, $X = [x_1, \ldots, x_n]'$, $\delta = [\delta_1, \ldots, \delta_n]'$, and $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]'$.[4]

Denote by $\widehat{\alpha}_{LS}$ the OLS estimator obtained by naively estimating Eq. (6) using reported participation, δ_i. Then, we have the following result.

Theorem 1. *Under Assumptions 1–4, the ordinary least squares estimator, $\widehat{\alpha}_{LS}$, is biased and inconsistent, and the asymptotic bias is given by*

$$plim(\widehat{\alpha}_{LS} - \alpha) = \frac{A - \alpha B}{C}, \tag{9}$$

with

$$A = \mathbb{E}\left[\sigma\varphi_v\phi\left(-z_i'\theta\right)\Phi\left(\frac{w_i'\gamma - \rho z_i'\theta}{\sqrt{1-\rho^2}}\right) + \sigma\varphi_u\phi\left(-w_i'\gamma\right)\Phi\left(\frac{z_i'\theta - \rho w_i'\gamma}{\sqrt{1-\rho^2}}\right)\right],$$

$$B = \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}[(\delta_i^* - \delta_i)x_i] \quad and \quad C = \mathbb{E}(\delta_i) - \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}(\delta_i x_i),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively the pdf and cdf of the standard normal.

Proof. See Appendix . □

Since the denominator in (9), C, is always positive by the Cauchy–Schwarz Inequality (see, e.g., Tripathi, 1999) the sign of the asymptotic bias only depends on the numerator of the expression. For example, if $B > 0$, then plim$(\widehat{\alpha}_{LS}) < \alpha$ for all $\alpha > A/B$ (i.e., there is an attenuation bias) and plim$(\widehat{\alpha}_{LS}) > \alpha$ for all $\alpha < A/B$ (i.e., there is an expansion bias). Also there are many instances in which plim$(\widehat{\alpha}_{LS})$ and α will have opposite signs. For example, if $B - C < 0$, then plim$(\widehat{\alpha}_{LS})$ and α have opposite signs whenever α falls between $A/(B - C)$ and 0 (Fig. 1 depicts the regions where bias and sign switching occur in this case).

[4] Re-writing the model in matrix notation is not necessary but makes the exposition (especially the proofs) less cumbersome. The matrix form also gives alternative (simpler) expressions for the various estimators considered.

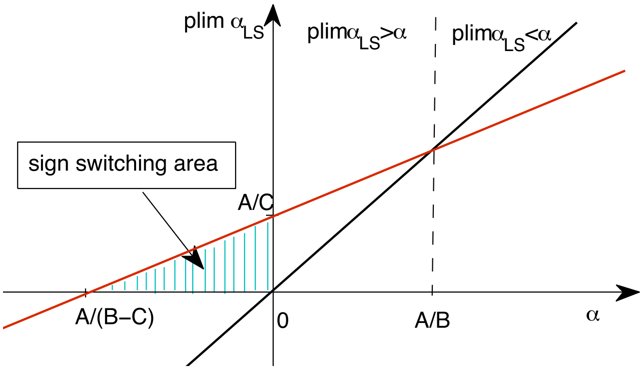

Fig. 1. Illustration of the OLS bias.

Note that sign-switching can occur even when participation is exogenous. Without loss of generality, consider the case of exogenous participation (i.e. $\varphi_v = 0$). The sign-switching result depicted in Fig. 1 follows if (i) $A > 0$, (ii) $B - C < 0$, and (iii) $A/(B - C) < \alpha < 0$. Condition (i) holds if misreporting is endogenous and the disturbance terms in Eqs. (1) and (3) are positively correlated ($\varphi_u > 0$). Thus, the size of the sign-switching region depends on how large the ratio $A/(B - C)$ is in general. In particular, in our example, the size of the sign-switching region increases with the rate of false negatives and the variance of the error term in the outcome equation and decreases with the rate of true participation, ceteris paribus. We provide evidence on the sign-switching region and these relationships in our Monte Carlo study in Section 4.

The above discussion shows that the bias related to misreporting is not merely an attenuation bias as found in many other studies (e.g., Aigner, 1973; Black et al., 2000; Lewbel, 2007). Under endogenous misreporting, the estimated treatment effect can assume an opposite sign, yielding misleading policy prescriptions. This sign reversal phenomenon would generally occur when misreporting is significant, and the direction of its correlation with the outcome is opposite to the direction of the treatment effect. For example, in the SNAP participation and obesity relationship, much empirical work has relied on self-reported participation and have found a positive or no effect on obesity. However, if people who are overweight are also more likely to correctly report participation (i.e., A positive) and since, as mentioned above, misreporting in SNAP is very severe in the data (i.e., $B - C$ is negative and small in magnitude) then we could observe a positive relationship between SNAP participation and obesity (i.e. $\text{plim}\hat{\alpha}_{LS} > 0$) even if the true effect is negative (i.e. $\alpha < 0$).

In the next section, we provide an estimation strategy that allows consistent estimation of the treatment effect, α. We first examine how well an IV estimation strategy would perform in our framework.

2.3. IV estimator under endogenous misreporting

The misreporting mechanism described above shows that in Eq. (6), the regressor, δ_i, is correlated with the error term ε_i as implied by Eq. (7). Thus, Eq. (1) can be seen as a regression with an endogenous binary regressor, even if true participation is exogenous and only misreporting is endogenous. So it may be tempting to suppose that if an instrument is present, then a standard IV estimator will address the issue raised in our framework. Here, we show that this is not the case.

Suppose we have access to a valid instrumental variable, z_i, such that $\mathbb{E}[z_i\varepsilon_i] = 0$ and $\text{Cov}(z_i, \delta_i) \neq 0$, and assume, for simplicity, that z_i is a scalar so that α is just identified. Then the (simple) instrumental variable estimator is given by

$$\widehat{\alpha}_{IV} = (z'M\delta)^{-1}z'My,$$

where $M = I - X(X'X)^{-1}X'$ is the orthogonal projection matrix onto the null space of X.

We can show using the same reasoning as above that,

$$\text{plim}(\widehat{\alpha}_{IV}) = \frac{\mathbb{E}(z_i\delta_i^*) - \mathbb{E}(z_ix_i')\mathbb{E}(x_ix_i')^{-1}\mathbb{E}(x_i\delta_i^*)}{\mathbb{E}(z_i\delta_i) - \mathbb{E}(z_ix_i')\mathbb{E}(x_ix_i')^{-1}\mathbb{E}[x_i\delta_i]}\alpha. \tag{10}$$

Thus, the IV estimator of α is inconsistent, and we cannot sign the bias in general. However, in the special case where misreporting is uncorrelated with true participation and the other covariates, it can be shown that,

$$\text{plim}(\widehat{\alpha}_{IV}) = \frac{\alpha}{\mathbb{E}[d_i]} = \frac{\alpha}{\Pr[d_i = 1]}, \quad \text{so that} \quad |\text{plim}(\widehat{\alpha}_{IV})| > |\alpha|.$$

Hence, in this specific scenario, the IV estimator is upwardly biased. This result is similar to those obtained by Black et al. (2000), (see also Frazis and Loewenstein, 2003; Brachet, 2008). The finding that the IV estimator is inconsistent is not new, given the results of the above authors and others. However, Black et al. (2000) showed that the IV estimator yields an expansion bias, which corresponds to the special case of exogenous measurement errors. By contrast, as suggested by Eq. (10), the sign of the IV bias is not obvious when misreporting is endogenous, and our simulations show that the ensuing bias can take either direction (i.e., expansion or attenuation).

We now present an estimation procedure that delivers consistent and asymptotically normal estimates for the treatment effect, α.

3. The proposed estimator

Recall that our objective is to estimate α in the outcome equation (1), where true (and possibly endogenous) participation status, δ_i^*, is unobserved, but only a possibly misreported (and possibly endogenous) participation status, δ_i, is observed. The proposed estimation strategy proceeds in the following two steps.

Step 1: With the joint distribution of u_i and v_i given by $F_{u,v}(u, v, \rho)$, use the partial observability probit model given by Eq. (4) to estimate the parameter vectors θ and γ. Then, compute the predicted probability for person i's true participation status as $\hat{\delta}_i^* = \Phi(z_i'\hat{\theta})$.

Step 2: Estimate Eq. (1) by substituting $\hat{\delta}_i^*$ for δ_i^*. Assuming correct model specification and distribution of the error terms, the resulting two-step estimator of α is consistent. Moreover, with standard regularity assumptions, this estimator is asymptotically normal.

3.1. First step estimation

Following Poirier (1980), the parameters γ, θ and ρ can be jointly estimated from the joint distribution of the error terms using the binary choice model defined by

$$\Pr[\delta_i = 1|w_i, z_i] = \Pr\left[-u_i \le w_i'\gamma,\ \ -v_i \le z_i'\theta\right] = F_{u,v}\left(w_i'\gamma, z_i'\theta, \rho\right) = P_i(\gamma, \theta, \rho).$$

The log-likelihood function of this model is given by

$$L_n(\gamma, \theta, \rho) = \sum_{i=1}^{n} \delta_i \ln P_i(\gamma, \theta, \rho) + (1-\delta_i)\ln\left(1 - P_i(\gamma, \theta, \rho)\right).$$

Assuming correct distributions, the maximum likelihood estimator of the vector of parameters (γ, θ, ρ) is consistent and asymptotically normal, and the covariance matrix consistently estimated with the inverse of the information matrix. In particular, for the parameter θ, the MLE $\hat{\theta}$ is consistent and asymptotically normal, i.e.,

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{and} \quad \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta),$$

where the asymptotic variance of $\hat{\theta}$ is obtained from the information matrix equality as

$$V_\theta = \left\{\mathbb{E}\left[\frac{1}{P_i(1-P_i)} \frac{\partial P_i}{\partial \theta} \frac{\partial P_i}{\partial \theta'}\right]\right\}^{-1}. \tag{11}$$

From this expression, a consistent estimator for the variance matrix can be obtained as

$$\widehat{V}_\theta = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\widehat{P}_i(1-\widehat{P}_i)} \frac{\partial \widehat{P}_i}{\partial \theta} \frac{\partial \widehat{P}_i}{\partial \theta'}\right]^{-1}, \tag{12}$$

where $\widehat{P}_i = P_i(\hat{\gamma}, \hat{\theta}, \hat{\rho}) = F_{u,v}\left(w_i'\hat{\gamma}, z_i'\hat{\theta}, \hat{\rho}\right)$. For the normal case, the gradient takes a fairly simple form

$$\frac{\partial \widehat{P}_i}{\partial \theta} = \phi(z_i'\hat{\theta})\Phi\left(\frac{w_i'\hat{\gamma} - \hat{\rho} z_i'\hat{\theta}}{\sqrt{1-\hat{\rho}^2}}\right) z_i.$$

Since this first-step is a maximum likelihood, parametric identification of (θ, γ, ρ) can be discussed in terms of non-singularity of the corresponding information matrix (Rothenberg, 1971). This means perfect multicollinearity needs to be ruled out, implying that both w_i and z_i should satisfy the standard rank conditions as a basic requirement. Also, as we explained earlier, a single exclusion restriction between w_i and z_i (i.e., at least one covariate in z_i should not be relevant in w_i, or vice-versa) is sufficient to identify all the first step parameters locally (Poirier, 1980).[5] Also, notice that only the (correct) specification of the marginal distribution of v is necessary for the parametric identification and estimation of the model in the second step. If the distribution of u or the joint distribution of (u, v) are unknown, then one may still obtain a consistent estimator of θ in the first step by using a semiparametric approach such as the series expansion of the joint PDF of (u, v) proposed by Gallant and Nychka (1987) or the single equation multiple index model described in Ichimura and Lee (1991).

[5] Essentially, identification implies much stronger conditions than the standard rank condition for linear IV since it requires that participation and hence the (nonlinear) relationship between true treatment and instruments be fully parameterized and correctly specified.

3.2. Second step estimation

In the second step, we compute the predicted values of true unobserved participation δ_i^*, given by $\hat{\delta}_i^* = \Phi(z_i'\hat{\theta})$, which are used in lieu of δ_i^* to estimate the parameters of the new model given by

$$y_i = x_i'\beta + \hat{\delta}_i^*\alpha + \eta_i. \tag{13}$$

Using the same approach as above, the second step estimator is obtained as

$$\begin{aligned}\widehat{\alpha}_{2S} &= (\hat{\delta}^{*'}M\hat{\delta}^*)^{-1}\hat{\delta}^{*'}My \\ &= \frac{\sum_{i=1}^n \Phi(z_i'\hat{\theta})y_i - \sum_{i=1}^n \Phi(z_i'\hat{\theta})x_i'[\sum_{i=1}^n x_ix_i']^{-1}\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n \Phi(z_i'\hat{\theta})^2 - \sum_{i=1}^n \Phi(z_i'\hat{\theta})x_i'[\sum_{i=1}^n x_ix_i']^{-1}\sum_{i=1}^n x_i\Phi(z_i'\hat{\theta})}\end{aligned} \tag{14}$$

We have the following consistency result.

Theorem 2. *Under Assumptions 1–3, the two-step estimator is consistent for α, that is, $\widehat{\alpha}_{2S}\xrightarrow{p}\alpha$.*

Proof. See Appendix . □

Notice that only the component $\hat{\theta}$ of the parameter vector is used at this second stage to predict the true unobserved participation status. The other components, $\hat{\gamma}$ and $\hat{\rho}$ are only used in the computation of the asymptotic variance estimator, as described below. In this second step, exclusion restriction is not strictly needed for identification as long as nonlinearity in the marginal distribution of v_i is assumed.

We have the following asymptotic normality result.

Theorem 3. *Under the model assumptions the two-step estimator is asymptotically normal, i.e.,*

$$\sqrt{n}(\widehat{\alpha}_{2S} - \alpha)\xrightarrow{d} N(0, \sigma_\alpha^2),$$

with

$$\sigma_\alpha^2 = \frac{\alpha^2\mathbb{E}[\Lambda_i(\theta)\phi(z_i'\theta)z_i']V(\hat{\theta})\mathbb{E}[z_i\phi(z_i'\theta)\Lambda_i(\theta)]}{\mathbb{E}[\Lambda_i^2(\theta)]^2} + \frac{\alpha^2\mathbb{E}[\Lambda_i^2(\theta)\Phi(z_i'\theta)(1-\Phi(z_i'\theta))]}{\mathbb{E}[\Lambda_i^2(\theta)]^2} + \frac{\sigma^2}{\mathbb{E}[\Lambda_i^2(\theta)]}$$

where

$$\Lambda_i(\theta) = \Phi(z_i'\theta) - \mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}x_i$$

Proof. See Appendix . □

This result is an application of the central limit theorem in the context of two-step estimators and is useful for our procedure to be readily usable for parametric inference. An expression for the variance estimator $\widehat{\sigma}_\alpha^2$ of σ_α^2 is given in the Appendix . However, this variance is quite involved and can be difficult to estimate. In practice, a simpler approach to evaluate the precision of $\widehat{\alpha}_{2S}$ and make inference about the treatment effect α is to use a bootstrap.

In sum, the outcome equation requires true participation status, δ^*, which is unobserved to the econometrician. Given the observed participation, δ, the first step in our estimation procedure amounts to a partial observability probit analysis on the indicator variable δ using both z and w, which are respectively the instrumental variables driving true participation and the covariates related to misreporting. The result of this analysis is an estimator, $\hat{\theta}$, of θ, the coefficient of z, which allows constructing a proxy $\hat{\delta}^*$ for truly being a participant. By construction, this proxy is purged from both endogeneity and misreporting, and is then used in lieu of δ^* in the outcome equation of interest to derive a consistent treatment effect estimator. The estimate $\hat{\theta}$ obtained from the first step can then be used along with the other model estimates to compute a consistent variance estimator for the treatment effect estimator.

A natural alternative to our two-step procedure is to estimate our model equations jointly via maximum likelihood (ML). Under appropriate assumptions, the ML procedures yield more efficient estimators and asymptotically correct estimates of standard errors. Unfortunately, in many situations, due to sample size and other considerations, the ML estimation can be both computationally complex and costly to implement, which may limit its use. For example, the correlations between the outcome equation error and the participation and reporting equations errors, φ_u and φ_v, might not be strongly identified, resulting in a likelihood function with ridges or multiple local maxima. Also, in some applications, the researcher may be reluctant to hypothesize a specific joint distribution between the random errors of the observed participation and the outcome as is required by maximum likelihood. In the Appendix , we briefly discuss the ML estimation of this model under the assumption of joint (trivariate) normality of the errors. In the same vein, we also briefly discuss how our method can be extended to the case of binary outcomes.

While this framework focuses on one-sided misreporting (i.e., only false negatives or false positives) which may be more appealing in specific contexts (e.g., when studying scenarios of participation in risky behavior or activities associated with stigma), a more general framework should account for misreporting in both directions (i.e., both false negatives and false positives). In the following section, we provide Monte Carlo simulations results on the performance of our estimator for both the one-sided case and the case where there is a small amount of misclassification in the other direction.[6]

4. Monte Carlo simulations

This section presents the results of Monte Carlo simulations comparing the proposed two-step estimator (2S) with OLS and IV estimators. Our goal is to consistently estimate α, the (conditional) average treatment effect of participation, δ^*, on an outcome, y, given by Eq. (1). However, since true participation is unobserved, our task reduces to use the proposed method to estimate α from Eq. (6) under the assumption that observed (misclassified) participation, δ, arises according to the process described by Eq. (4), which focuses on false negatives. In the simulations, we also consider a slight departure from Eq. (4) and allow for small amounts of false positives as described below.

4.1. Simulation setup

The baseline data generating process is simulated as follows. The true treatment indicator, δ_i^*, is given by

$$\delta_i^* = \mathbf{1}\left(\theta_0 + \theta_1 z_i + v_i \geq 0\right), \quad \text{where } z_i \sim N(0,1), \quad \theta_0 = 0.1, \quad \theta_1 = 1.$$

The outcome equation y_i is given by

$$y_i = \beta_0 + x_i\beta_1 + \delta_i^*\alpha + \epsilon_i \quad \text{where } x_i \sim N(0,1) \quad \beta_0 = \beta_1 = 1, \quad \alpha = -0.2.$$

Note that $\alpha = -0.2$ is the true population treatment effect we seek to estimate.

The econometrician only observes an error-ridden treatment indicator, δ_i, defined by

$$\delta_i = \delta_i^*\mathbf{1}\left(\gamma_0 + \gamma_1 w_i + u_i \geq c\right) + (1-\delta_i^*)\mathbf{1}\left(\zeta_i < b\right),$$

where $w_i \sim N(0,1)$, $\gamma_0 = 0.01$, $\gamma_1 = 2$, and $b \in [0, 1)$.

The parameter c is the threshold that determines the proportion of false negatives in the sample.[7] The disturbance term, ζ_i, is drawn from a uniform $(0, 1)$ distribution independently from z_i and v_i so that the parameter b corresponds to the rate of false positives. For example, when $b = 0$ (baseline case), the observed treatment indicator is given by $\delta_i = \delta_i^*\mathbf{1}\left(\gamma_0 + \gamma_1 w_i + u_i \geq c\right)$ which only allows for false negatives as given by Eq. (4). However, when $b > 0$, the observed treatment indicator allows for both false negatives and a $(100 \times b)$% rate of false positives.

The disturbances ϵ_i, u_i and v_i are drawn from a trivariate distribution given by

$$(\epsilon_i, u_i, v_i) \sim IID\left(0, \Sigma\right), \quad \text{where} \quad \Sigma = \begin{pmatrix} \sigma^2 & \varphi_u\sigma & \varphi_v\sigma \\ \varphi_u\sigma & 1 & \rho \\ \varphi_v\sigma & \rho & 1 \end{pmatrix}, \quad \sigma = 1.$$

The baseline results assume joint normality although we consider non-normal distributions as well. The values of the parameters φ_v and φ_u, which are the correlations of the outcome equation error term with participation and misreporting equation disturbance terms, respectively, are varied in the simulations to examine how various degrees of the endogeneity of participation and misreporting with respect to the outcome impact the results. We also allow ρ, the correlation between participation and misreporting to vary. We estimate the treatment effect α and the associated bias using the naive OLS approach, $\hat{\alpha}_{LS}$, and the proposed two-step approach, $\hat{\alpha}_{2S}$. We also estimate the instrumental variable estimators $\hat{\alpha}_{IV1}$ and $\hat{\alpha}_{IV2}$ using z and $[z,\ w]$ as instruments, respectively.

4.2. Simulation results

We report simulation results averaged over 1000 replications each with sample size 5000 for different levels of false negatives – 0%, 5%, 10%, 20%, 40% – for $\rho \in \{0,\ 0.3\}$, $\varphi_u \in \{0,\ 0.2,\ 0.8\}$ and $\varphi_v \in \{-0.3,\ 0,\ 0.3\}$. These results are first presented for $b = 0$ (i.e., 0% false positives) and, subsequently, also for $b \in \{0.01,\ 0.05,\ 0.1\}$ (i.e., 1%, 5% and 10% false positives). Thus, no misreporting corresponds to the case of 0% false negatives and false positives. The cases of exogenous participation and exogenous misreporting correspond to $\varphi_u = \varphi_v = 0$. Table 1 presents the results of the Monte Carlo simulations for OLS, IVs, and the proposed two-step (2S) estimators when the errors are jointly normal, the false positive rate is 0%, and $\rho = 0.3$. We report both the OLS estimates using the true treatment indicator, δ_i^* (OLS-T) and the observed treatment indicator δ_i (OLS-O). Although δ_i^* is unobserved to the econometrician, the OLS-T estimates provide a theoretical benchmark for the estimates obtained using the misclassified δ_i. We also report both the IV estimates using z as an instrument (IV-1) and those using $[z,\ w]$ as instruments (IV-2). The proposed estimator is denoted (2S) in the tables.

[6] Extending this framework to the two-sided endogenous misreporting case is not straightforward. It would require at least two sets of excluded covariates, that is, w_1 and w_2, each associated with one of the misreporting directions, and possibly other additional functional form/distributional assumptions for identification.

[7] By appropriately choosing the value of c, one can simulate varying rates of misreporting.

Table 1
Monte Carlo simulations.

False Negatives	φ_u	φ_v	OLS-T	OLS-O	IV-1 [x, z]	IV-2 [x, z, w]	2S
0%	0	−0.3	−0.5523	−0.5523	−0.2006	−0.2003	−0.1998
		0	−0.1997	−0.1997	−0.2014	−0.2013	−0.2009
		0.3	0.1519	0.1519	−0.1951	−0.1954	−0.1963
	0.2	−0.3	−0.5524	−0.5524	−0.2036	−0.2033	−0.2030
		0	−0.2021	−0.2021	−0.2011	−0.2012	−0.1999
		0.3	0.1508	0.1508	−0.1983	−0.1986	−0.1972
	0.8	−0.3	−0.5507	−0.5507	−0.1991	−0.1988	−0.1980
		0	−0.1994	−0.1994	−0.2003	−0.2003	−0.1993
		0.3	0.1523	0.1523	−0.1985	−0.1987	−0.1962
5%	0	−0.3	−0.5499	−0.5066	−0.2041	−0.2152	−0.1997
		0	−0.1987	−0.1834	−0.2018	−0.2127	−0.1980
		0.3	0.1504	0.1357	−0.2069	−0.2188	−0.2012
	0.2	−0.3	−0.5513	−0.4827	−0.2061	−0.2166	−0.2001
		0	−0.2008	−0.1611	−0.2063	−0.2168	−0.2009
		0.3	0.1513	0.1615	−0.2054	−0.2167	−0.2009
	0.8	−0.3	−0.5506	−0.4072	−0.2044	−0.2158	−0.2013
		0	−0.1993	−0.0844	−0.2039	−0.2160	−0.2007
		0.3	0.1516	0.2371	−0.2036	−0.2146	−0.2003
10%	0	−0.3	−0.5515	−0.4649	−0.1949	−0.2345	−0.2000
		0	−0.2013	−0.1715	−0.1936	−0.2319	−0.1991
		0.3	0.1524	0.1250	−0.1922	−0.2323	−0.1992
	0.2	−0.3	−0.5499	−0.4203	−0.1960	−0.2350	−0.2008
		0	−0.1989	−0.1266	−0.1942	−0.2305	−0.1973
		0.3	0.1518	0.1677	−0.1952	−0.2339	−0.1995
	0.8	−0.3	−0.5513	−0.2916	−0.1953	−0.2340	−0.1998
		0	−0.1986	0.0053	−0.1936	−0.2331	−0.1978
		0.3	0.1514	0.2987	−0.1937	−0.2350	−0.2005
20%	0	−0.3	−0.5508	−0.4065	−0.1674	−0.2668	−0.1984
		0	−0.1990	−0.1496	−0.1681	−0.2666	−0.1989
		0.3	0.1501	0.1050	−0.1690	−0.2681	−0.2002
	0.2	−0.3	−0.5518	−0.3421	−0.1733	−0.2728	−0.2036
		0	−0.2004	−0.0858	−0.1690	−0.2665	−0.1992
		0.3	0.1498	0.1693	−0.1730	−0.2743	−0.2026
	0.8	−0.5	−0.5511	−0.1443	−0.1690	−0.2679	−0.1988
		0	−0.2015	0.1097	−0.1675	−0.2668	−0.1990
		0.3	0.1512	0.3680	−0.1709	−0.2728	−0.2034
40%	0	−0.3	−0.5502	−0.3161	−0.1215	−0.3740	−0.1981
		0	−0.2008	−0.1198	−0.1219	−0.3814	−0.2027
		0.3	0.1506	0.0783	−0.1210	−0.3746	−0.1987
	0.2	−0.3	−0.5501	−0.2181	−0.1160	−0.3687	−0.1963
		0	−0.2015	−0.0230	−0.1213	−0.3852	−0.2041
		0.3	0.1507	0.1743	−0.1191	−0.3788	−0.2003
	0.8	−0.3	−0.5509	0.0675	−0.1207	−0.3757	−0.1993
		0	−0.2005	0.2646	−0.1189	−0.3789	−0.2016
		0.3	0.1514	0.4607	−0.1211	−0.3783	−0.2000

Notes: The true treatment effect is $\alpha = -0.2$. Each calibration in the Monte Carlo Design involved 1000 replications each of size 5000. We report results for five false negative rates (0%, 5%, 10%, 20%, and 40%) — the proportion of true participants who misreport their status. φ_v and φ_u are correlations that indicate the extents of endogeneity of participation and misreporting, respectively. The correlation between participation and misreporting is $\rho = 0.3$. Also, the error terms are jointly normally distributed and the false positive rate is 0%.

The naive OLS estimates using δ_i (OLS-O) show that not only is the OLS estimator inconsistent as asserted in Theorem 1, but also yields the wrong (i.e., positive) sign, whether participation is exogenous or endogenous. Sign switching is observed at all nonzero false negative rates i.e. 5%, 10%, 20% and 40% and is more pronounced at higher values of φ_u. These results persist even under the special case of exogenous misreporting ($\varphi_u = 0$). The IV estimates (IV-1) and (IV-2) show that the classic IV estimator is also inconsistent and sometimes worse than the OLS, albeit keeping the correct (negative) sign.[8] Interestingly, (IV-1) shows expansion biases while (IV-2) shows attenuation biases. This confirms as we explained in Section 2.3, that we cannot generally sign the bias in the IV estimator when misreporting is endogenous.

In contrast, the proposed two-step estimator (2S), presented in the last column of Table 1, yields consistent estimates of the true treatment effect and by comparison, is superior to both the OLS and IV estimators under both endogenous and exogenous misreporting or participation. Also, the proposed estimator remains accurate and performs remarkably well, even when the rate of false negatives is substantially high in the data. Moreover, there is no cost in doing our procedure since the

[8] This is actually a better set of simulations for IV-2 because the covariate w_i can be used as an additional instrument to improve the IV. Additional simulations with w_i being endogenous yielded worse results for this IV while the proposed estimator (2S) remained consistent.

Table 2
Sensitivity of the proposed estimator to misspecification.

False	φ_u	Baseline	Types of misspecification of baseline				
			False positives			Distribution of errors	
Negatives			1%	5%	10%	$\Gamma(3, 1)$	$\chi^2_{(1)}$
0%	0	−0.2000	−0.2021	−0.2141	−0.2291	−0.2047	−0.2110
	0.2	−0.1970	−0.2003	−0.2126	−0.2257	−0.1971	−0.2077
	0.8	−0.1981	−0.2008	−0.2129	−0.2266	−0.1967	−0.2090
5%	0	−0.2041	−0.2057	−0.2191	−0.2317	−0.2077	−0.2059
	0.2	−0.1974	−0.2074	−0.2208	−0.2312	−0.1935	−0.2094
	0.8	−0.1989	−0.2035	−0.2169	−0.2320	−0.1957	−0.2082
10%	0	−0.1987	−0.2029	−0.2169	−0.2323	−0.1972	−0.2051
	0.2	−0.2001	−0.2060	−0.2207	−0.2331	−0.2007	−0.2042
	0.8	−0.1983	−0.2027	−0.2171	−0.2350	−0.1955	−0.2057
20%	0	−0.2018	−0.2063	−0.2234	−0.2428	−0.2013	−0.2037
	0.2	−0.1989	−0.2056	−0.2229	−0.2411	−0.1966	−0.2064
	0.8	−0.2003	−0.2036	−0.2210	−0.2377	−0.1970	−0.2073
40%	0	−0.1961	−0.2066	−0.2314	−0.2652	−0.1873	−0.2058
	0.2	−0.2000	−0.2085	−0.2333	−0.2628	−0.1977	−0.2107
	0.8	−0.2009	−0.2087	−0.2336	−0.2598	−0.2004	−0.2104
			Endogeneity of predictor			Omission of predictor	
			$corr(w, \epsilon)$ $= .5$	$corr(w, u)$ $= .5$	$corr(w, v)$ $= .5$	omitted w	omitted* w^2
0%	0	−0.2000	−0.1975	−0.1975	−0.1738	−0.1972	−0.2009
	0.2	−0.1970	−0.2008	−0.2008	−0.1765	−0.2017	−0.1986
	0.8	−0.1981	−0.1995	−0.1995	0.1754	−0.1980	−0.1967
5%	0	−0.2041	−0.1988	−0.1988	−0.1752	−0.1956	−0.2035
	0.2	−0.1974	−0.2012	−0.2013	−0.1773	−0.1979	−0.2012
	0.8	−0.1989	−0.1985	−0.1984	−0.1743	−0.1981	−0.1981
10%	0	−0.1987	−0.2017	−0.2018	−0.1776	−0.1965	−0.2001
	0.2	−0.2001	−0.1972	−0.1973	−0.1741	−0.2008	−0.1990
	0.8	−0.1983	−0.1991	−0.1992	−0.1759	−0.2013	−0.2009
20%	0	−0.2018	−0.1954	−0.1954	−0.1712	−0.2004	−0.1981
	0.2	−0.1989	−0.2008	−0.2008	−0.1763	−0.1996	−0.1997
	0.8	−0.2003	−0.1988	−0.1989	−0.1742	−0.2008	−0.1990
40%	0	−0.1961	−0.2003	−0.2004	−0.1732	−0.2022	−0.2027
	0.2	−0.2000	−0.1974	−0.1974	−0.1705	−0.1975	−0.2025
	0.8	−0.2009	−0.2004	−0.2000	−0.1732	−0.1969	−0.2037

Notes: The true treatment effect is $\alpha = -0.2$. We fix $\rho = \varphi_v = 0$. Each calibration in the Monte Carlo Design involved 1000 replications each of size 5000. We report results for five false negatives rates (0%, 5%, 10%, 20%, and 40%) and false positive rates of (0%, 1%, 5%, and 10%). The correlation φ_u indicates the extent of endogeneity of misreporting.
* Here, the true misreporting equation includes both w and w^2, but the estimation omits w^2.

proposed estimator remains as good as the OLS and the IV when there is 0% false negatives and participation is exogenous ($\varphi_v = 0$). This performance is not sensitive to the choice of parameters such as the variance of the outcome equation error or the correlation between the error terms in the participation and misreporting equations (see, e.g., the results for $\rho = 0$ in the "Baseline" column of Table 2).

To further assess the robustness of our proposed estimator, we investigate its sensitivity to misspecification in a number of directions. First, we allow the reported participation to include both false negatives (as before) and a small amount of false positives. We consider false positive rates of 1%, 5%, and 10%. Second, we allow for the error terms to be non-normal. We consider both the trivariate Gamma distribution and the trivariate Chi-squared distribution as alternatives to allow more skewness and kurtosis in the distributions of error terms.[9] Third, we allow for the misreporting equation to be misspecified by considering the case where the predictor is unavailable to the researcher (i.e., only x is included) or by introducing a quadratic term in w in the data generating process but which is omitted by the researcher in the estimation. Fourth, we introduce correlation between the predictors of misreporting w and the error terms in both participation and outcome equations.

Table 2 summarizes the results where ρ and φ_v are fixed to zero (exogenous participation), and the focus is on the sensitivity to different degrees of endogeneity of misreporting φ_u and various rates of false negatives.[10] These results show that at any false positive rate the bias increases with false negative rates. Interestingly, for small amounts of false positives,

[9] These multivariate distributions can be simulated using the Copulas method or the inverse transformation method as described in Gentle (2002)

[10] Results for other combinations of parameter values are similar and are available upon request.

Table 3
Intervals of α yielding sign-switching in the OLS.

False	φ_u	Sign-switching region	
Negatives		$\sigma = 1$	$\sigma = 4$
5%	-0.8	$[0,\ 0.2307]$	$[0,\ 0.9227]$
	-0.2	$[0,\ 0.0577]$	$[0,\ 0.2309]$
	0	$\varnothing$	$\varnothing$
	0.2	$[-0.0577,\ 0]$	$[-0.2309,\ 0]$
	0.8	$[-0.2307,\ 0]$	$[-0.9227,\ 0]$
10%	-0.8	$[0,\ 0.4054]$	$[0,\ 1.6216]$
	-0.2	$[0,\ 0.1033]$	$[0,\ 0.8368]$
	0	$\varnothing$	$\varnothing$
	0.2	$[-0.1033,\ 0]$	$[-0.8368,\ 0]$
	0.8	$[-0.4054,\ 0]$	$[-1.6216,\ 0]$
40%	-0.8	$[0,\ 1.1347]$	$[0,\ 5.3786]$
	-0.2	$[0,\ 0.3399]$	$[0,\ 1.3597]$
	0	$\varnothing$	$\varnothing$
	0.2	$[-0.3399,\ 0]$	$[-1.3597,\ 0]$
	0.8	$[-1.1347,\ 0]$	$[-5.3786,\ 0]$

Notes: Results are reported for three false negatives rates (5%, 10%, and 40%). The correlation φ_u indicates the extent of endogeneity of misreporting; φ_v and ρ are fixed to 0.

the proposed estimator still performs quite well, although it gets worse with higher rates of false positives. Specifically, when false positive rates range from 1% to 5%, the median value of the proposed estimator (2S) ranges between -0.2042 and -0.2180 for all ranges of false negatives in our setting. The proposed estimator is robust to non-normality of the error terms and remains consistent when the true error distributions are Gamma or Chi-Squared in this setting.

When the misreporting equation is misspecified by including only the covariates from the outcome equation (i.e., only x) or when this equation includes a quadratic term in w that is omitted by the researcher in the estimation, the 2S estimator still performs well. Finally, the 2S estimator is robust to introducing correlations between the predictor w and both the outcome and misreporting equation errors. When w is correlated with the outcome equation or the misreporting equation errors, the 2S estimator remains consistent. However, the 2S estimator performs poorly, exhibiting an attenuation bias just like the IV-1, when w and the participation equation error are correlated. Our recommendation is to access w and z from different data sources to minimize the chances of having w endogenous to true participation in practice.

There are a few additional facts that are worth mentioning. On the one hand, it is not surprising that the OLS estimator only works well when there are 0% false negatives and participation is exogenous ($\varphi_v = 0$). On the other hand, the IV estimator tends to work well for low levels of false negatives (0%–5%) but gets worse for higher false negative rates (10% and higher). As explained earlier, the sign-reversal regions for the OLS depends on the quantity $A/(B - C)$ (given in Theorem 1), which varies with φ_v, φ_u, σ, and the extent of misclassification. Even when participation is exogenous (i.e. $\varphi_v = 0$), various degrees of endogeneity of misreporting (e.g., $\varphi_u \in \{-0.8,\ -0.2,\ 0,\ 0.2,\ 0.8,\}$), various sizes of the error variance (e.g., $\sigma \in \{1,\ 4\}$), and various rates of false negatives (e.g., 5%, 10%, 40%) yield different sign-switching regions for the OLS, as shown in Table 3.

Table 3 shows the ranges of the true treatment effects α for which the OLS estimator $\hat{\alpha}_{LS}$ would yield the wrong (opposite) signs in our simulation design. For example, negative correlations between misreporting and outcome errors yield positive intervals of the treatment effect for which the OLS takes the wrong (negative) sign, while positive correlations between misreporting and outcome errors yield negative intervals of the true treatment effect for which the OLS takes the wrong (positive) sign. In all cases, higher levels of endogeneity of misreporting, higher rates of false positives or greater error variance in the outcome equation yield wider sign-reversal intervals. It is only when misreporting is also exogenous ($\varphi_u = 0$) that the OLS keeps the same sign as the true treatment effect (albeit still biased), so that the sign-switching set is empty (see Table 3).

Finally, Lewbel (2007)'s estimator also worked well in our setting for the special cases where both participation and misreporting where exogenous. However, Lewbel's estimator displayed large biases and sign reversals under some endogeneity cases, which is not surprising since this limitation is clearly emphasized in Lewbel (2007). These additional results are available from the authors upon request.[11]

5. Empirical example

We illustrate our proposed method with an empirical example examining the impact of SNAP on adult body mass index (BMI).[12] As mentioned in Section 1, a major difficulty in estimating SNAP's impacts is the high reporting error rates in national surveys, with false negatives being more prevalent than false positives.

[11] It is easy to slightly modify our set up to include the IV required by Lewbel's identification strategy. For that purpose, one can add a binary indicator in the true participation equation, since, as explained by Lewbel (2007), only two points of support are needed for the instrument to identify the treatment effect in the case of one-sided misreporting.

[12] BMI is defined as weight in kilograms divided by height in meters squared.

Table 4
Summary statistics by SNAP participation status.

	Non-SNAP		SNAP	
	Mean	Std. Error	Mean	Std. Error
Body Mass Index	28.11	(0.25)	29.85	(0.52)
Age	38.96	(0.09)	38.29	(0.21)
Hispanic	0.08	(0.01)	0.11	(0.02)
Black	0.21	(0.03)	0.38	(0.04)
Household Size	3.45	(0.06)	3.77	(0.09)
WIC	0.06	(0.00)	0.21	(0.02)
SSI	0.06	(0.01)	0.26	(0.02)
AFDC	0.02	(0.00)	0.45	(0.03)
Mother's Education (High school or higher)	0.52	(0.02)	0.43	(0.04)
Number of children	1.81	(0.05)	2.23	(0.08)
Household with child <5	0.14	(0.01)	0.22	(0.02)
Gross Income SNAP Eligible Household (130% FPL)	0.39	(0.02)	0.86	(0.01)
Biometric	0.16	(0.02)	0.23	(0.04)
EBT card	0.62	(0.02)	0.55	(0.03)
Phone Interview	0.36	(0.01)	0.25	(0.02)
Net Family Income (2004 Thousand Dollars)	22.04	(0.506)	12.644	(0.364)
Observations	4307		1163	

Notes: Standard errors in parentheses are adjusted for the complex survey design of the NLSY79. Based on the 1996–2004 biennial waves of the NLSY79, restricted to females with income lower than 250% of the federal poverty level.

We consider a simple treatment effects model relating BMI to a binary indicator of SNAP participation. Since SNAP participation is not randomly assigned and possibly misreported, OLS and IV estimates are biased and inconsistent for the average treatment effect.[13] We present estimates of SNAP's effect on BMI using OLS, IV, and our proposed two-step (2S) estimators using the 1996–2004 waves of female respondents of the restricted-use National Longitudinal Survey of Youth - 1979 (NLSY79).[14]

As discussed in Section 3, estimation of our 2S estimator proceeds in two steps. We first estimate a partial observability probit model described in Eq. (4) to obtain the parameters of our true participation and misreporting equations. The second step uses the predicted probabilities of true participation to estimate the effect of SNAP on BMI from Eq. (13).

To implement the 2S estimator, two sets of covariates need to be distinguished: instruments for participation (z_i in Eq. (2)) and predictors of misreporting (w_i in Eq. (3)). Regarding z_i, the excluded instruments for participation are whether the respondent's state uses a biometric identification technology ("Biometric") and the percentage of SNAP benefits issued by the state via Electronic Benefit Cards ("EBT Card") (Almada et al., 2016).

As for w_i, we take the set of regressors in the outcome equation augmented with both the biometric identification technology mentioned above ("Biometric") as well as a binary indicator for whether the interview was conducted by telephone or in person ("Phone Interview"). Intuitively, the latter captures variations in interview mode while the former could increase stigma, both of which might be correlated with misreporting, in addition to personal characteristics.[15] Notice that the vectors z_i and w_i are different albeit overlapping, as required by the theory.

The summary statistics in Table 4 by SNAP participation status suggest that participants are largely negatively selected in the program. In Table 5, we report two sets of regression estimates for the OLS, IV, and 2S. The first set of results uses five covariates (Age, Black, Hispanic, Household size, Family income) and only Biometric as the instrument. The second set uses eleven covariates (adding Number of children, Mother's education, participation in WIC, AFDC/TANF and SSI, Household with child) and Biometric EBT Card as instruments. The estimates from these methods yield different and sometimes opposite (in sign) results.

The OLS estimator suggests a positive and statistically significant effect of SNAP participation on BMI of about 0.929 BMI units for the first set and 1.006 BMI units for the second set. The IV estimator which uses Biometric as instrument suggests a 10.62 decrease in BMI units, while the one that uses Biometric and EBT Card as instruments shows only a 0.576 decrease in BMI units, although not statistically significant.[16] In contrast, our proposed estimator yields a negative and statistically significant effect of -3.093 for the First Set and -3.187 for the Second Set scenarios. These results show that while the OLS

[13] See Gundersen (2015) for a review of the literature on the SNAP-obesity relationship.

[14] We restrict our analysis sample to females who are within 250% of the federal poverty level. Determining true SNAP eligibility status is almost impossible with most national surveys due to the lack of a comprehensible set of variables needed to determine each respondent's true eligibility. Thus, we follow the existing literature and use 250% of the federal poverty level to determine our eligible sample. This type of sample restriction is common in the literature which also favors thresholds higher than the gross-income eligibility threshold of 130% (Mykerezi and Mills, 2010; Almada et al., 2016).

[15] Also, we include in w_i a dummy for being eligible for SNAP based on the Gross Income test (130% of the federal poverty level). We thank an anonymous referee for pointing these issues out.

[16] See Table C.2 in Appendix C for the first stage IV results.

Table 5
The impact of SNAP on BMI.

Variable	Set of Covariates	Dependent Variable: BMI		
		OLS	IV	2S
SNAP Participation	First Set	0.929^{**}	-10.62	-3.093^{*}
		(0.379)	(12.61)	(1.804)
	Second Set	1.006^{**}	-0.576	-3.187^{**}
		(0.432)	(7.849)	(1.538)

Notes: Standard errors in parenthesis and bootstrapped (500 replications) for the 2S estimator. Results are based on the 1996–2004 biennial waves of the NLSY79, restricted to females with income lower than 250% of the federal poverty level. For the First Set, regressors not reported include respondents' age, race, household size, household income). For the Second Set, additional regressors are number of children, mother's education, square of income, time fixed effects, and indicators for receiving WIC benefits, AFDC/TANF, SSI benefits, and having an infant living in home.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

and the 2S estimators are stable, significant and of opposite signs, the IV estimator is unstable and varies substantially with the instruments as our simulations suggested earlier.[17]

This empirical illustration is undoubtedly limited, and the results should be interpreted with caution since they are only suggestive. For instance, even if our proposed method estimates α consistently, it may not represent a causal effect of SNAP on obesity due to possible confounding by omitted variables in this application.

Subject to these caveats, the empirical example corroborates the simulation results obtained in Section 4. First, there is a possible sign reversal between the OLS and the 2S estimates. Second, the IV and the 2S estimators have the same sign, but the IV has a smaller or a larger magnitude.[18] Also, the results obtained from these methods could lead to radically different and possibly contradictory policy advice.

6. Conclusion

This paper examines the identification and estimation of the conditional average treatment effect of a binary regressor in the presence of endogenous misreporting and possibly endogenous participation. We derive and prove the consistency and asymptotic normality of our proposed two-step estimator and show that OLS and IV estimators are inconsistent and may yield wrong (opposite) signs from the true effect.

We also provide Monte Carlo simulations to this effect and illustrate our method with an empirical example examining the impact of SNAP participation on obesity. Previous studies on misclassified binary regressors are mostly concerned with exogenous or random misreporting (Aigner, 1973; Brachet, 2008; Lewbel, 2007; Mahajan, 2006; Frazis and Loewenstein, 2003), where it is commonly assumed that misclassification probabilities depend only on the true treatment status and are thus, independent of measurement errors and other regressors. Our two-step estimator relaxes this arguably strong assumption and shows that, when the researcher has access to information related to why individuals misreport, the treatment effect can be consistently estimated.

To our knowledge, this paper is the first attempt to provide point estimates of treatment effect in the context of endogenous misreporting of a binary treatment variable. This is important because of the prevalence of misreporting in public programs and survey data (Meyer et al., 2009; Bollinger, 1996; Kane and Rouse, 1995; Kane et al., 1999; Brachet, 2008). While this paper focused on one-sided endogenous misreporting when participation is possibly endogenous, future work should allow for bidirectional misreporting (i.e., false negatives and false positives). It would also be useful to show the level of dependence of our approach on distributional and functional form assumptions by considering parametric or semi-parametric estimation approaches.

Appendix A. Proofs

A.1. Proof of Theorem 1

Proof. Biasedness: By the Frisch–Waugh–Lovell Theorem, see, e.g. Davidson and MacKinnon (2004, page 68), the regression

$$My = M\delta\alpha + v$$

[17] Table C.1 in Appendix C presents the results of the first step of the 2S estimator. Panel A of Table C.1 shows both instruments have the expected signs but only "EBT Card" is statistically significantly correlated with true participation. Also, Panel B of Table C.1 suggests that being interviewed by phone is negatively correlated with the probability of truthful reporting of participation.

[18] Note that the sign reversal phenomenon obtained in this empirical illustration is neither a general result nor does its nonoccurrence invalidate the results herein. See Section 2.2 for further discussions.

yields the same least squares estimate of α as the regression equation of interest (8). It follows that,

$$\widehat{\alpha}_{LS} = (\delta' M \delta)^{-1}\delta' M y. \tag{15}$$

This implies that $\widehat{\alpha}_{LS} - \alpha = (\delta' M\delta)^{-1}\delta' M\varepsilon$.
Hence, $\mathbb{E}[\widehat{\alpha}_{LS} - \alpha | X, \delta] = (\delta' M\delta)^{-1}\delta' M \mathbb{E}[\varepsilon|X, \delta] \neq 0$, since $\mathbb{E}[\varepsilon|\delta, X] \neq 0$ by the correlation of ε and δ through u and v.
Inconsistency: We can write

$$\begin{aligned}\widehat{\alpha}_{LS} - \alpha &= (\delta' M\delta)^{-1}\delta' M\epsilon = \left(\frac{\delta' M\delta}{n}\right)^{-1}\frac{\delta' M\varepsilon}{n} \\ &= \left(\frac{\delta' M\delta}{n}\right)^{-1}\left(\frac{\delta' M\epsilon}{n} + \frac{\delta' M(\delta^* - \delta)\alpha}{n}\right) \quad \text{by Eq. (7)}\end{aligned} \tag{16}$$

Notice that,

$$\frac{\delta' M\delta}{n} = \frac{\delta'[I - X(X'X)^{-1}X']\delta}{n} = \frac{\delta'\delta}{n} - \frac{\delta' X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\delta}{n}$$

Hence, by the Weak Law of Large Numbers and the Slutsky's lemma, we have

$$\frac{\delta' M\delta}{n} \xrightarrow{p} \mathbb{E}(\delta_i^2) - \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}(\delta_i x_i)$$

By a matrix extension of the Cauchy–Schwarz inequality (see Tripathi, 1999), we know that $\mathbb{E}(\delta_i^2) - \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}(\delta_i x_i) > 0$. The Continuous Mapping Theorem then implies that

$$\left(\frac{\delta' M\delta}{n}\right)^{-1} \xrightarrow{p} \left[\mathbb{E}(\delta_i^2) - \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}(\delta_i x_i)\right]^{-1}. \tag{17}$$

Likewise, the term $\dfrac{\delta' M\epsilon}{n}$ can also be decomposed as

$$\frac{\delta' M\epsilon}{n} = \frac{\delta'[I - X(X'X)^{-1}X']\epsilon}{n} = \frac{\delta'\epsilon}{n} - \frac{\delta' X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\epsilon}{n}.$$

Then, using the same arguments as above we have

$$\frac{\delta' M\epsilon}{n} \xrightarrow{p} \mathbb{E}(\delta_i\epsilon_i) - \mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}(x_i\epsilon_i) = \mathbb{E}(\delta_i\epsilon_i),$$

where the last equality follows from Assumption 1.

Using the expression of δ_i given by Eq. (4) and the trivariate normality of (ϵ_i, u_i, v_i), it can be shown by integration that

$$\begin{aligned}\mathbb{E}[\delta_i\epsilon_i] &= \mathbb{E}\left[\epsilon_i \mathbf{1}\left(z_i'\theta + v_i \geq 0, \quad w_i'\gamma + u_i \geq 0\right)\right] \\ &= \mathbb{E}\left[\Pr[u_i \geq -w_i'\gamma,\ v_i \geq -z_i'\theta, \rho]\mathbb{E}\left[\epsilon_i | u_i \geq -w_i'\gamma,\ v_i \geq -z_i'\theta\right]\right] \\ &= \mathbb{E}\left[\sigma\varphi_v\phi\left(-z_i'\theta\right)\Phi\left(\frac{w_i'\gamma - \rho z_i'\theta}{\sqrt{1-\rho^2}}\right) + \sigma\varphi_u\phi\left(-w_i'\gamma\right)\Phi\left(\frac{z_i'\theta - \rho w_i'\gamma}{\sqrt{1-\rho^2}}\right)\right],\end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of the standard normal. It follows that

$$\frac{\delta' M\epsilon}{n} \xrightarrow{p} \mathbb{E}\left[\sigma\varphi_v\phi\left(-z_i'\theta\right)\Phi\left(\frac{w_i'\gamma - \rho z_i'\theta}{\sqrt{1-\rho^2}}\right) + \sigma\varphi_u\phi\left(-w_i'\gamma\right)\Phi\left(\frac{z_i'\theta - \rho w_i'\gamma}{\sqrt{1-\rho^2}}\right)\right]. \tag{18}$$

Finally, using the same reasoning as above for the term $\dfrac{\delta' M(\delta^* - \delta)\alpha}{n}$, we have

$$\frac{\delta' M(\delta^* - \delta)\alpha}{n} \xrightarrow{p} -\alpha\mathbb{E}(\delta_i x_i')\mathbb{E}(x_i x_i')^{-1}\mathbb{E}[(\delta_i^* - \delta_i)x_i]. \tag{19}$$

The desired result follows by taking (19), (18) and (17) to Eq. (16). □

A.2. Proof of Theorem 2

Proof. We can write

$$\widehat{\alpha}_{2S} = (\hat{\delta}^{*\prime} M\hat{\delta}^*)^{-1}\hat{\delta}^{*\prime} M\delta^*\alpha + (\hat{\delta}^{*\prime} M\hat{\delta}^*)^{-1}\hat{\delta}^{*\prime} M\epsilon \tag{20}$$

By the exogeneity of X and Z given by Assumption 1, the consistency of $\hat{\theta}$, the continuity of $\Phi(\cdot)$ and the law of large numbers, we have

$$\frac{\hat{\delta}^{*\prime}M\epsilon}{n}\xrightarrow{p}\mathbb{E}[\Phi(z_i'\theta)\epsilon_i]=\mathbb{E}\left[\Phi(z_i'\theta)\mathbb{E}[\epsilon_i|z_i]\right]=0,$$

so that the second term on the RHS of Eq. (20) goes to zero. We also have, by Assumption 2, the consistency of $\hat{\theta}$, the continuity of $\Phi(\cdot)$ and the law of large numbers,

$$\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\xrightarrow{p}\mathbb{E}\left[\Phi(z_i'\theta)^2\right]-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}\mathbb{E}\left[x_i\Phi(z_i'\theta)\right]$$

and

$$\begin{aligned}\frac{\hat{\delta}^{*\prime}M\delta^{*}}{n} &\xrightarrow{p} \mathbb{E}\left[\Phi(z_i'\theta)\delta_i^*\right]-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}\mathbb{E}\left[x_i\delta_i^*\right]\\ &= \mathbb{E}\left[\Phi(z_i'\theta)\mathbb{E}[\delta_i^*|z_i]\right]-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}\mathbb{E}\left[x_i\mathbb{E}[\delta_i^*|z_i]\right]\\ &= \mathbb{E}\left[\Phi(z_i'\theta)^2\right]-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}\mathbb{E}\left[x_i\Phi(z_i'\theta)\right]\end{aligned}$$

where the last display follows from the fact that $\mathbb{E}[\delta_i^*|z_i]=\Phi(z_i'\theta)$, as implied by Eq. (2). Hence,

$$(\hat{\delta}^{*\prime}M\hat{\delta}^{*})^{-1}\hat{\delta}^{*\prime}M\delta^{*}=\left(\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\right)^{-1}\frac{\hat{\delta}^{*\prime}M\delta^{*}}{n}\xrightarrow{p}1$$

so that

$$\widehat{\alpha}_{2S}\xrightarrow{p}\alpha\quad\square$$

A.3. Proof of Theorem 3

Proof. We can write

$$\begin{aligned}\sqrt{n}(\widehat{\alpha}_{2S}-\alpha) &= \left(\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\right)^{-1}\left(\frac{\hat{\delta}^{*\prime}M(\delta^{*}-\hat{\delta}^{*})}{\sqrt{n}}\right)\alpha+\left(\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\right)^{-1}\frac{\hat{\delta}^{*\prime}M\epsilon}{\sqrt{n}}\\ &= \left(\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\right)^{-1}\left(\frac{\hat{\delta}^{*\prime}M(\Psi^{*}-\hat{\delta}^{*})}{\sqrt{n}}\right)\alpha+\left(\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n}\right)^{-1}\frac{\hat{\delta}^{*\prime}M(\alpha(\delta^{*}-\Psi^{*})+\epsilon)}{\sqrt{n}}\\ &= q_n^{-1}\left[\sqrt{n}V_{1n}\alpha+\sqrt{n}V_{2n}\right] \qquad (21)\end{aligned}$$

where

$$q_n=\frac{\hat{\delta}^{*\prime}M\hat{\delta}^{*}}{n},\quad V_{1n}=\frac{\hat{\delta}^{*\prime}M(\Psi^{*}-\hat{\delta}^{*})}{n},\quad\text{and}\quad V_{2n}=\frac{\hat{\delta}^{*\prime}M(\alpha(\delta^{*}-\Psi^{*})+\epsilon)}{n},$$

with $\Psi^*=\Psi^*(\theta)=[\Phi(z_1'\theta),\dots,\Phi(z_n'\theta)]'$.

Denote $\hat{\Lambda}_i=\hat{\delta}_i^*-\left(\frac{1}{n}\sum_{i=1}^n\hat{\delta}_i^*x_i'\right)\left(\frac{1}{n}\sum_{i=1}^n x_ix_i'\right)^{-1}x_i$ and by $\Lambda_i=\Phi(z_i'\theta)-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}x_i$ its probability limit. Notice that $q_n=\frac{1}{n}\sum_{i=1}^n\hat{\Lambda}_i^2$. We know, from the consistency results above that

$$q_n\xrightarrow{p}q=\mathbb{E}\left[\Phi(z_i'\theta)^2\right]-\mathbb{E}\left[\Phi(z_i'\theta)x_i'\right]\mathbb{E}[x_ix_i']^{-1}\mathbb{E}\left[\Phi(z_i'\theta)x_i\right]=\mathbb{E}[\Lambda_i^2]. \qquad (22)$$

Since $\hat{\delta}^*=\Psi^*(\hat{\theta})$, then expanding $\Psi^*(\theta)$ in a Taylor series about $\hat{\theta}$, we have

$$\Psi^*-\hat{\delta}^*\overset{a}{=}\psi^*(\hat{\theta})(\theta-\hat{\theta})$$

where "$\overset{a}{=}$" denotes asymptotic equivalence in probability, and $\psi^*(\theta)$ is the vector of partial derivatives given by

$$\psi^*(\theta)=\frac{\partial\Psi^*(\theta)}{\partial\theta'}=[\phi(z_1'\theta)z_1,\dots,\phi(z_n'\theta)z_n]'.$$

Therefore

$$\sqrt{n}V_{1n}\overset{a}{=}\frac{\hat{\delta}^{*\prime}M\psi^*(\hat{\theta})}{n}\sqrt{n}(\theta-\hat{\theta})=\frac{1}{n}\sum_{i=1}^n\hat{\Lambda}_i\phi(z_i'\hat{\theta})z_i'\sqrt{n}(\theta-\hat{\theta}).$$

A direct application of the central limit theorem then gives,

$$\sqrt{n}V_{1n}\alpha \xrightarrow{d} N(0, \alpha^2\nu_1^2),$$

where

$$\nu_1^2 = \mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']V(\hat{\theta})\mathbb{E}[z_i\phi(z_i'\theta)\Lambda_i]. \tag{23}$$

Likewise,

$$\sqrt{n}V_{2n} = \frac{\hat{\delta}^{*'}M(\alpha(\delta^* - \Psi^*) + \epsilon)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\Lambda}_i\zeta_i$$

where $\zeta_i = \alpha(\delta_i^* - \Phi(z_i'\theta)) + \epsilon_i$, with $\mathbb{E}[\zeta_i|z_i] = 0$ and $\mathrm{Var}[\zeta_i|z_i] = \alpha^2\Phi(z_i'\theta)(1 - \Phi(z_i'\theta)) + \sigma^2$.

Hence, by the central limit theorem,

$$\sqrt{n}V_{2n} \xrightarrow{d} N(0, \sigma_2^2),$$

where

$$\begin{aligned}\sigma_2^2 &= \mathbb{E}[\Lambda_i\left(\alpha^2\Phi(z_i'\theta)(1 - \Phi(z_i'\theta)) + \sigma^2\right)\Lambda_i] \\ &= \alpha^2\mathbb{E}[\Lambda_i^2\Phi(z_i'\theta)(1 - \Phi(z_i'\theta))] + \sigma^2\mathbb{E}[\Lambda_i^2].\end{aligned} \tag{24}$$

Finally, the asymptotic covariance term between the elements of $\sqrt{n}V_{1n}\alpha$ and $\sqrt{n}V_{2n}$ is

$$\begin{aligned}\sigma_{12} &= \mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']\mathbb{E}[(\theta - \hat{\theta})\left(\alpha(\delta_i^* - \Phi(z_i'\theta)) + \epsilon_i\right)]\mathbb{E}[\Lambda_i]\alpha \\ &= -\mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']\mathbb{E}[\hat{\theta}\epsilon_i]\mathbb{E}[\Lambda_i]\alpha.\end{aligned} \tag{25}$$

It then follows from Slutsky's Lemma, (21), (22), (23), (24) and (25) that

$$\sqrt{n}(\widehat{\alpha}_{2S} - \alpha) \xrightarrow{d} N(0, \sigma_\alpha^2), \quad \text{where}$$

$$\begin{aligned}\sigma_\alpha^2 &= \frac{\alpha^2\nu_1^2}{q^2} + 2\frac{\sigma_{12}}{q^2} + \frac{\sigma_2^2}{q^2} \\ &= \frac{\alpha^2\mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']V(\hat{\theta})\mathbb{E}[z_i\phi(z_i'\theta)\Lambda_i]}{\mathbb{E}[\Lambda_i^2]^2} - 2\frac{\alpha\mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']\mathbb{E}[\hat{\theta}\epsilon_i]\mathbb{E}[\Lambda_i]}{\mathbb{E}[\Lambda_i^2]^2} \\ &\quad + \frac{\alpha^2\mathbb{E}[\Lambda_i^2\Phi(z_i'\theta)(1 - \Phi(z_i'\theta))]}{\mathbb{E}[\Lambda_i^2]^2} + \frac{\sigma^2}{\mathbb{E}[\Lambda_i^2]}.\end{aligned}$$

With $\hat{\theta}$ and ϵ_i uncorrelated conditionally on z_i and w_i the covariance term is zero, and the variance reduces to

$$\sigma_\alpha^2 = \frac{\alpha^2\mathbb{E}[\Lambda_i\phi(z_i'\theta)z_i']V(\hat{\theta})\mathbb{E}[z_i\phi(z_i'\theta)\Lambda_i]}{\mathbb{E}[\Lambda_i^2]^2} + \frac{\alpha^2\mathbb{E}[\Lambda_i^2\Phi(z_i'\theta)(1 - \Phi(z_i'\theta))]}{\mathbb{E}[\Lambda_i^2]^2} + \frac{\sigma^2}{\mathbb{E}[\Lambda_i^2]}.$$

A consistent estimator for this asymptotic variance can be defined by

$$\hat{\sigma}_\alpha^2 = \frac{\hat{\alpha}_{2S}^2\hat{\nu}_1^2}{\hat{q}^2} + \frac{\hat{\alpha}_{2S}^2\hat{\nu}_2^2}{\hat{q}^2} + \frac{\hat{\sigma}^2}{\hat{q}}$$

where

$$\hat{\nu}_1^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\Lambda}_i\phi(z_i'\hat{\theta})z_i'\right)\widehat{V}(\hat{\theta})\left(\frac{1}{n}\sum_{i=1}^{n}z_i\phi(z_i'\hat{\theta})\widehat{\Lambda}_i\right),$$

$$\hat{\nu}_2^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i^2\Phi(z_i'\hat{\theta})\left(1 - \Phi(z_i'\hat{\theta})\right),$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i}\left[\left(y_i - x_i'\hat{\beta} - \hat{\alpha}_{2S}\Phi(z_i'\hat{\theta})\right)^2 - \hat{\alpha}_{2S}^2\Phi(z_i'\hat{\theta})\left(1 - \Phi(z_i'\hat{\theta})\right)\right],$$

$$\hat{q} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\Phi(z_i'\hat{\theta})^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\Phi(z_i'\hat{\theta})x_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i\Phi(z_i'\hat{\theta})\right),$$

and $\hat{\alpha}_{2S}$ is our proposed estimator of α. □

Appendix B. Extensions

B.1. Maximum likelihood estimation

Let Assumptions 1–4 hold and assume that the covariance matrix of the joint distribution of errors Σ defined in (5) is positive definite. Then our model can be estimated jointly using full information maximum likelihood. The log-likelihood function is built from the joint density of y_i, δ_i^* and d_i, which we write as the product of the conditional and the marginal densities

$$f(\delta_i^*, d_i, y_i) = f(\delta_i^*, d_i|y_i)f(y_i).$$

To derive the conditional distributions, we use results for the trivariate normal, and write

$$v_i = \rho_1\epsilon_i/\sigma + \rho_2 u_i + \eta_i, \quad \text{with} \quad \eta_i|\epsilon_i, u_i \sim N(0, \kappa^2)$$

where ρ_1, ρ_2 and κ^2 are defined in terms of the original parameters φ_v, φ_u and ρ by:[19]

$$\rho_1 = \frac{\varphi_v - \rho\varphi_u}{1-\varphi_u^2}, \quad \rho_2 = \frac{\rho - \varphi_v\varphi_u}{1-\varphi_u^2}, \quad \kappa^2 = 1 - \rho_1^2 - \rho_2^2 - 2\rho_1\rho_2\varphi_u$$

Denote $\Theta = (\theta', \gamma', \rho, \varphi_u, \varphi_v, \beta', \alpha, \sigma^2)'$ the vector of all the parameters of the model. Then,

$$\Upsilon_{11i}(\Theta) = f(\delta_i^* = 1, d_i = 1|y_i) = \Phi_2\left(\frac{z_i'\theta + \rho_1(y_i - \alpha - x_i'\beta)/\sigma}{\sqrt{\rho_2^2+\kappa^2}}, w_i'\gamma, \frac{\rho_2}{\sqrt{\rho_2^2+\kappa^2}}\right)$$

$$\Upsilon_{10i}(\Theta) = f(\delta_i^* = 1, d_i = 0|y_i) = \Phi\left(\frac{z_i'\theta + \rho_1(y_i - \alpha - x_i'\beta)/\sigma}{\sqrt{\rho_2^2+\kappa^2}}\right) - \Upsilon_{11i}(\Theta)$$

$$\Upsilon_{0i}(\Theta) = f(\delta_i^* = 0|y_i) = 1 - \Phi\left(\frac{z_i'\theta + \rho_1(y_i - x_i'\beta)/\sigma}{\sqrt{\rho_2^2+\kappa^2}}\right) = 1 - \Upsilon_{11i}(\Theta) - \Upsilon_{10i}(\Theta)$$

where $\Phi_2(\cdot,\cdot,\cdot)$ is the CDF of the standard bivariate normal distribution.

The full information log-likelihood function of the model is then defined by

$$l(\Theta) = \sum_{i=1}^n l_i(\Theta),$$

with

$$\begin{aligned} l_i(\Theta) = \delta_i \ln\left[\Upsilon_{11i}(\Theta)\frac{1}{\sigma}\phi\left(\frac{y_i - \alpha - x_i'\beta}{\sigma}\right)\right] + \\ + (1-\delta_i)\ln\left[\Upsilon_{10i}(\Theta)\frac{1}{\sigma}\phi\left(\frac{y_i - \alpha - x_i'\beta}{\sigma}\right) + \Upsilon_{0i}(\Theta)\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] \end{aligned} \tag{26}$$

Maximizing this function with respect to Θ yields a consistent and asymptotically efficient estimator of the model parameters.

B.2. Extension to binary outcomes

The method discussed in this paper can be extended to the case of binary outcomes. However, in this case we cannot just do the plug-in method described earlier, because a linear probability model would exhibit serious problems, especially the fact that it could produce a wrong sign for the treatment effect, even if the treatment status is correctly classified (see, e.g., discussion provided by Lewbel et al., 2012). A more reliable alternative in this case would be the maximum likelihood estimation. We assume the binary outcome y_i is related to the exogenous covariates x_i and to the true treatment indicator δ_i^* by

$$y_i = \mathbf{1}\left[x_i'\beta + \delta_i^*\alpha + \epsilon_i > 0\right]. \tag{27}$$

[19] The fact that the covariance matrix of the joint distribution of errors is constrained to be positive definite guarantees that these new parameters are well-defined, namely, $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, and $\kappa^2 > 0$.

True participation δ_i^* and misreporting d_i are defined, as before, by Eqs. (2) and (3), respectively. We maintain Assumptions 1–4 above, except that the conditional variance of the error ϵ_i is now normalized to 1 (as is usually the case for identification in Probit models). Given the observed participation $\delta_i = \delta_i^* d_i$, and the outcome y_i, the log-likelihood function of the binary choice model (BCM) is built up from the joint probabilities $\Pr[y_i, \delta_i^*, d_i]$ of these dichotomous variables as follows.

$$\begin{aligned} l_{BCM}(\Theta) = \sum_{i=1}^{n} \delta_i \ln \Pr[y_i, \delta_i^* = 1, d_i = 1] + \\ + (1-\delta_i) \ln \left(\Pr[y_i, \delta_i^* = 1, d_i = 0] + \Pr[y_i, \delta_i^* = 0] \right), \end{aligned} \tag{28}$$

where

$$\Pr[y_i, \delta_i^* = 1, d_i = 1] = \Pr[y_i = 1, \delta_i^* = 1, d_i = 1]^{y_i} \Pr[y_i = 0, \delta_i^* = 1, d_i = 1]^{1-y_i},$$

$$\Pr[y_i, \delta_i^* = 1, d_i = 0] = \Pr[y_i = 1, \delta_i^* = 1, d_i = 0]^{y_i} \Pr[y_i = 0, \delta_i^* = 1, d_i = 0]^{1-y_i},$$

$$\text{and} \qquad \Pr[y_i, \delta_i^* = 0] = \Pr[y_i = 1, \delta_i^* = 0]^{y_i} \Pr[y_i = 0, \delta_i^* = 0]^{1-y_i}$$

The probabilities in these equations can be obtained in terms of model parameters:

$$\Pr[y_i = 1, \delta_i^* = 1, d_i = 1] = \Phi_3 \left(x_i'\beta + \alpha, z_i'\theta, w_i'\gamma; \varphi_v, \varphi_u, \rho \right)$$

$$\Pr[y_i = 0, \delta_i^* = 1, d_i = 1] = \Phi_2 \left(z_i'\theta, w_i'\gamma; \rho \right) - \Pr[y_i = 1, \delta_i^* = 1, d_i = 1]$$

$$\Pr[y_i = 1, \delta_i^* = 1, d_i = 0] = \Phi_2 \left(x_i'\beta + \alpha, w_i'\gamma; \varphi_v \right) - \Pr[y_i = 1, \delta_i^* = 1, d_i = 1]$$

$$\Pr[y_i = 0, \delta_i^* = 1, d_i = 0] = \Phi \left(z_i'\theta \right) - \Phi_2 \left(z_i'\theta, w_i'\gamma; \rho \right) - \Pr[y_i = 1, \delta_i^* = 1, d_i = 0]$$

$$\Pr[y_i = 1, \delta_i^* = 0] = \Phi \left(x_i'\beta \right) - \Phi_2 \left(x_i'\beta, z_i'\theta; \varphi_v \right)$$

$$\Pr[y_i = 0, \delta_i^* = 0] = 1 - \Phi \left(z_i'\theta \right) - \Pr[y_i = 1, \delta_i^* = 0]$$

where $\Phi_3(\cdot, \cdot, \cdot)$ is the CDF of the standard trivariate normal distribution.

Appendix C. Additional tables

See Tables C.1 and C.2.

Table C.1
Partial observability probit estimates (first step of proposed estimator).

	Dependent Variable: SNAP Participation Dummy	
	First Set	Second Set
Panel A: Participation Equation		
Biometric	0.509*	−0.078
	(0.305)	(0.217)
EBT Card		0.242***
		(0.086)
Panel B: Misreporting Equation		
Phone Interview	−0.270***	−0.092*
	(0.066)	(0.050)
Biometric		0.040
		(0.175)
Observations	5470	5470

Notes: This table presents the first step maximum likelihood estimated coefficients of partial observability model specified in Eq. (4). Standard errors in parentheses. Results are based on the 1996–2004 biennial waves of the NLSY79, restricted to females with income lower than 250% of the federal poverty level. For the First Set column, regressors not reported include respondents's age, race, household size, household income. For the Second Set column, additional regressors are number of children, mother's education, square of income, time fixed effects, and indicators for receiving WIC benefits, AFDC/TANF, SSI benefits, and having an infant living in home.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Table C.2
First stage IV results.

	Dependent Variable: SNAP Participation Dummy	
	First Set	Second Set
Biometric	0.0456*	0.005
	(0.026)	(0.014)
EBT card		0.046***
		(0.014)
F-statistics	3.19*	5.27***
Hansen J-statistic		0.147
Observations	5470	5470

Notes: This table reports the first stage estimates of the linear IV estimator. Standard errors in parentheses. Results are based on the 1996–2004 biennial waves of the NLSY79, restricted to females with income lower than 250% of the federal poverty level. For the First Set column, regressors not reported include respondents's age, race, household size, household income. For the Second Set column, additional regressors are number of children, mother's education, square of income, time fixed effects, and indicators for receiving WIC benefits, AFDC/TANF, SSI benefits, and having an infant living in home.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

References

Aigner, D.J., 1973. Regression with a binary independent variable subject to errors of observation. J. Econometrics 1 (1), 49–59.

Almada, L., McCarthy, I., Tchernis, R., 2016. What can we learn about the effects of food stamps on obesity in the presence of misreporting? Am. J. Agric. Econ..

Battistin, E., Sianesi, B., 2011. Misclassified treatment effects and treatment status: an application to returns to education in the United Kingdom. Rev. Econ. Stat. 93 (2), 495–509.

Black, D.A., Berger, M.C., Scott, F.A., 2000. Bounding parameter estimates with nonclassical measurement error. J. Amer. Statist. Assoc. 95 (451), 739–748.

Bollinger, C.R., 1996. Bounding mean regressions when a binary regressor is mismeasured. J. Econometrics 73 (2), 387–399.

Bollinger, C.R., van Hasselt, M., 2017. Bayesian moment-based inference in a regression model with misclassification error. J. Econometrics 200 (2), 282–294.

Bound, J., 1991. Self-reported versus objective measures of health in retirement models. J. Hum. Resour. 26 (1), 106–138.

Brachet, T., 2008. Maternal smoking, misclassification, and infant health. Available at SSRN: https://ssrn.com/abstract=1009781.

Chen, X., Hu, Y., Lewbel, A., 2008a. A note on the closed-form identification of regression models with a mismeasured binary regressor. Statist. Probab. Lett. 78, 1473–1479.

Chen, X., Hu, Y., Lewbel, A., 2008b. Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. Econom. Lett. 100, 381–384.

Davidson, R., MacKinnon, J.G., 2004. Econometric Theory and Methods, vol. 5. Oxford University Press New York.

DiTraglia, F.J., García-Jimeno, C., 2017. Mis-classified, Binary, Endogenous Regressors: Identification and Inference.

Feinstein, J.S., 1990. Detection controlled estimation. J. Law Econ. 33 (1), 233–276.

Frazis, H., Loewenstein, M.A., 2003. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. J. Econometrics 117 (1), 151–178.

Gallant, A., Nychka, D., 1987. Semi-Nonparametric maximum likelihood estimation. Econometrica 55, 363–390.

Gandelman, N., Rasteletti, A., 2017. Credit constraints, sector informality and firm investments: evidence from a panel of uruguayan firms. J. Appl. Econ. 20 (2), 351–372.

Gentle, J.E., 2002. Random Number Generation and Monte Carlo methods, second ed. Oxford University Press New York.

Gumbel, E., 1961. Bivariate logistic distributions. J. Amer. Statist. Assoc. 56 (294), 335–349.

Gundersen, C., 2015. SNAP and obesity. In: Bartfeld, J., Gunderson, C., Smeeting, M.T., Ziliak, P.J. (Eds.), SNAP Matters: How Food Stamps Affect Health and Well Being. Stanford University Press, Redwood City, CA.

Haider, S., Solon, G., 2006. Life-cycle variation in the association between current and lifetime earnings. Amer. Econ. Rev. 96 (4), 1308–1320.

van Hasselt, M., Bollinger, C., 2012. Binary misclassification and identification in regression models. Econom. Lett. 115, 81–84.

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the dependent variable in a discrete-response setting. J. Econometrics 87 (2), 239–269.

Hu, Y., Shiu, J.L., Woutersen, T., 2015. Identification and estimation of single-index models with measurement error and endogeneity. Econom. J. 18, 347–362.

Hu, Y., Shiu, J.L., Woutersen, T., 2016. Identification in nonseparable models with measurement errors and endogeneity. Econom. Lett. 144, 33–36.

Ichimura, H., Lee, L.F., 1991. Semiparametric least squares estimators of multiple index models. Nonparametric Semiparametric Methods Econom. 3–49.

Kane, T.J., Rouse, C.E., 1995. Labor-market returns to two-and four-year college. Am. Econ. Rev. 600–614.

Kane, T.J., Rouse, C.E., Staiger, D., 1999. Estimating Returns to Schooling When Schooling is Misreported. In: Working Paper Series, vol. 7235, National Bureau of Economic Research, Available at NBER: http://www.nber.org/papers/w7235.

Kreider, B., 2010. Regression coefficient identification decay in the presence of infrequent classification errors. Rev. Econ. Stat. 92 (4), 1017–1023.

Kreider, B., Pepper, J.V., Gundersen, C., Jolliffe, D., 2012. Identifying the effects of SNAP (Food Stamps) on child health outcomes when participation is endogenous and misreported. J. Amer. Statist. Assoc. 107 (499), 958–975.

Lewbel, A., 2007. Estimation of average treatment effects with misclassification. Econometrica 75 (2), 537–551.

Lewbel, A., Dong, Y., Yang, T.T., 2012. Comparing features of convenient estimators for binary choice models with endogenous regressors. Canad. J. Econ. 45 (3), 809–829.

Lynch, A.G., Marioni, J.C., Tavaré, S., 2007. Numbers of copy-number variations and false-negative rates will be underestimated if we do not account for the dependence between repeated experiments. Am. J. Human Genet. 81 (2), 4–18.

Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74 (3), 631–665.

Marquis, K.H., Moore, J.C., 1990. Measurement errors in sipp program reports. In: Proceedings of the 1990 Annual Research Conference. U.S. Bureau of the Census, Washington, DC, pp. 721–745.

Meyer, B.D., Mittag, N., Goerge, R.M., 2018. Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation. In: Working Paper Series, vol. 25143, National Bureau of Economic Research, Available at NBER: http://www.nber.org/papers/w25143.

Meyer, B.D., Mok, W.K.C., Sullivan, J.X., 2009. The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences. In: Working Paper Series, vol. 15181, National Bureau of Economic Research, Available at NBER: http://www.nber.org/papers/w15181.

Millimet, D.L., 2011. The elephant in the corner: a cautionary tale about measurement error in treatment effects models. In: Missing Data Methods: Cross-Sectional Methods and Applications. In: Advances in Econometrics, vol. 27, Part 1, Emerald Group Publishing Limited, pp. 1–39.

Mykerezi, E., Mills, B., 2010. The impact of food stamp program participation on household food insecurity. Am. J. Agric. Econ. 92 (5), 1379–1391.

Poirier, D.J., 1980. Partial observability in bivariate probit models. J. Econometrics 12, 209–217.

Rothenberg, T., 1971. Identification in parametric models. Econometrica 69 (3), 577–591.

Stephens, M., Unayama, T., 2018. Estimating the impacts of program benefits: using instrumental variables with underreported and imputed data. Rev. Econ. Stat. (forthcoming).

Tripathi, G., 1999. A matrix extension of the Cauchy-Schwarz inequality. Econom. Lett. 63, 1–3.

Ura, T., 2018. Heterogeneous treatment effects with mismeasured endogenous treatment. Quant. Econ. (forthcoming).