NBER WORKING PAPER SERIES

MISCLASSIFICATION IN BINARY CHOICE MODELS

Bruce Meyer
Nikolas Mittag

Working Paper 20509
http://www.nber.org/papers/w20509

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2014

 We would like to thank Frank Limehouse for assistance with the data and participants at presentations at the Census Research Data Center Conference, UNCE at Charles University and Xiamen University for their comments.

Misclassification in Binary Choice Models
Bruce Meyer and Nikolas Mittag
NBER Working Paper No. 20509
September 2014
JEL No. C25,C81,H53

ABSTRACT

While measurement error in the dependent variable does not lead to bias in some well-known cases, with a binary dependent variable the bias can be pronounced. In binary choice, Hausman, Abrevaya and Scott-Morton (1998) show that the marginal effects in the observed data differ from the true ones in proportion to the sum of the misclassification probabilities when the errors are unrelated to covariates. We provide two sets of results that extend this analysis. First, we derive the asymptotic bias in parametric models allowing for correlation of the errors with both observables and unobservables. Second, we examine the bias in a prototypical application in two different datasets, using a variety of methods that differ in the amount of knowledge that is assumed about the error process. Our application is receipt of food stamps, the largest and most widely received welfare program in the U.S. Monte Carlo results and our empirical application show that the bias formulas accurately describe the bias in finite samples. Our results indicate that the robustness of signs and relative magnitudes of coefficients implied by the earlier proportionality results does not necessarily extend to estimated Probit coefficients, and does not apply when errors are correlated with covariates. Using administrative records linked to survey data as validation data, we evaluate estimators that are consistent under misclassification. Estimators based on the assumption that misclassification is independent of the covariates are sensitive to their functional form assumptions and aggravate the bias if the conditional independence assumption is invalid in all cases we examine. On the other hand, estimators that allow misreporting to be correlated with the covariates perform well if an accurate model of misreporting or validation data are available. Estimators that incorporate more information about the errors, such as aggregate underreporting rates, tend to be more robust to misspecification of the misreporting model.

Bruce Meyer
Harris School of Public Policy
University of Chicago
1155 E. 60th Street
Chicago, IL 60637
and NBER
bdmeyer@uchicago.edu

Nikolas Mittag
CERGE-EI
Politických vz 7
111 21 Praha 1
Czech Republic
nikolas.mittag@cerge-ei.cz

1 Introduction

Many important outcomes are binary such as program receipt, labor market status, and educational attainment. These outcomes are frequently misclassified in data sets due to interviewer or respondent error or other reasons. It is a common misconception that measurement error in a dependent variable does not lead to bias, but this result requires classical measurement error. Misclassification of a binary variable is necessarily non-classical measurement error, and thus leads to bias. However, there are few general results on the direction and magnitude of the bias in binary choice models. In this paper, we examine the properties of binary choice models with measurement error in the dependent variable. We then discuss the performance of several estimators designed to account for measurement error. We rely on a combination of analytical results, simulations, and results from an application to the Food Stamp Program.

Several papers have examined misreporting in surveys and have found high rates of misclassification in binary variables such as participation in welfare programs (Marquis and Moore, 1990; Meyer, Mok and Sullivan, 2009; Meyer, Goerge and Mittag, 2014), Medicaid enrollment (Call et al., 2008; Davern et al., 2009*a*,*b*) and education (Black, Sanders and Taylor, 2003). Bound, Brown and Mathiowetz (2001) provide an overview of misreporting in survey data. In the case of program reporting, false negatives, i.e. recipients that fail to report receiving program benefits, seem to be the main problem with rates of underreporting sometimes exceeding 50%. As our application to food stamp receipt shows, measurement error can badly bias substantive studies with binary outcomes such as those examining take-up of other programs (e.g. Bitler, Currie and Scholz, 2003; Haider, Jacknowitz and Schoeni, 2003), labor market status (e.g. Poterba and Summers, 1995) or educational attainment (e.g. Eckstein and Wolpin, 1999; Cameron and Heckman, 2001). Since our application deals with misreporting in survey data, we use the terms misclassification and misreporting interchangeably, but all our results remain valid if misclassification arises for other reasons. A frequent cause of error besides misreporting is subjective classification of a dependent

variable, for example whether there is a recession or not (e.g. Estrella and Mishkin, 1998) or the presence of an armed civil conflict (e.g Collier and Hoeffler, 1998; Fearon and Laitin, 2003). Similarly, a proxy variable is often used instead of the true variable of interest, such as when arrests or incarcerations are used instead of crimes (e.g. Levitt, 1998; Lochner and Moretti, 2004). Another reason for misclassification is prediction error. This problem may arise, for example, if some observations of the dependent variable are missing and imputed values are substituted. Moreover, there is no ex ante reason to believe that misclassification is random. This randomness assumption may be more likely to be true if misclassification stems from coding errors or failure to link some records.

A few papers have analyzed the consequences of misreporting for econometric models. For example, Bollinger and David (1997, 2001) and Meyer, Goerge and Mittag (2014) examine how misclassification affects estimates of food stamp participation and Davern et al. (2009*b*) analyze the demographics of Medicaid enrollment. These papers show that misclassification affects the estimates of common econometric models and distorts the conclusions drawn from them in meaningful ways. From these studies we know that misreporting can seriously alter estimates from binary choice models, but we know very little about the way misreporting affects estimates in general. This situation is aggravated by the scarcity of analytic results on bias in binary choice models. Carroll et al. (2006) and Chen, Hong and Nekipelov (2011) provide overviews of the literature on measurement error in non-linear models and there is a small literature on misspecification in binary choice models (e.g. Yatchew and Griliches, 1985; Ruud, 1983, 1986), but general results or formulas for biases are scarce and usually confined to special cases for specific models. In a key paper, Hausman, Abrevaya and Scott-Morton (1998) show that the marginal effects in the observed mismeasured data differ from the true ones in proportion to the sum of the misclassification probabilities when the errors are unrelated to covariates.

In the more general case, we present a closed form solution for the bias in the linear probability model and decompose the bias in non-linear binary choice models such as the Probit

model into four components. We present closed form expressions for three bias components and an equation that determines the fourth component. These results help to explain the bias found in the studies above and are informative about the likely size and direction of bias in cases where the "true" dependent variable is not available. We illustrate these results using simulations and data on food stamp receipt from Illinois and Maryland matched to the 2001 American Community Survey (ACS) and the 2002-2005 Current Population Survey (CPS). We show how these biases affect coefficients in a model of food stamp receipt. We then use our results to interpret biased coefficients and assess whether substantive conclusions obtained from misclassified data are likely to be valid. Some features of the true parameters are often robust to misclassification and we consider conditions under which one can use the biased coefficients to learn about features of the true coefficients such as signs, bounds and relative magnitudes.

While little is known about how misclassification biases estimates, several papers have attempted to correct estimates for misclassification or proposed estimators that are consistent in the presence of misclassification. In terms of binary choice models, Bollinger and David (1997) and Hausman, Abrevaya and Scott-Morton (1998) introduce consistent estimators for the Probit model. Unless the true parameters are known, it is impossible to test whether these estimators or other corrections improve parameter estimates. In most cases, a change indicates a problem with the original model, but by itself this does not imply that the correction is an improvement.

We use the same data and model of food stamp participation mentioned above to analyze the performance of several estimators for the Probit model that are consistent under certain forms of misclassification. We examine their performance and assess how sensitive they are to violations of their assumptions for consistency and how useful it is to incorporate additional information on the nature of misclassification. Our results suggest that some of the corrections work very well, but that making false simplifying assumptions on the misclassification may lead to results that are even worse than ignoring the problem altogether.

We find that a good model of misclassification can serve as a substitute for accurate data. However, a bad model of misclassification can make things worse than the naive Probit estimates. This result shows that it is important to know whether there is misclassification in the data and whether it is related to the covariates.

The next section introduces the models and discusses the bias in theory, section 3 examines them in practice using the matched survey data and simulation studies. Section 4.1 introduces the consistent Probit estimators, section 4.2 evaluates their performance when misclassification is unrelated to the covariates while section 4.3 analyzes their performance when misclassification is related to the covariates. Section 5 concludes.

2 Bias due to Misclassification of a Binary Dependent Variable

2.1 Model Setup and Previous Results

Throughout this paper, we are concerned with a situation in which a binary outcome y is related to observed characteristics X, but the outcome indicator is subject to misclassification. Let y_i^T be the true indicator for the outcome of individual i and y_i be the observed indicator that is subject to misclassification. The sample size is N and N_{MC} observations are misclassified, N_{FP} of which are false positives and N_{FN} are false negatives. We define the probabilities of false positives and false negatives conditional on the true response as

$$\Pr(y_i = 1|y_i^T = 0) = \alpha_{0i}$$
$$\Pr(y_i = 0|y_i^T = 1) = \alpha_{1i}$$

We refer to them throughout the paper as the conditional probabilities of misreporting. Additionally, we define a binary random variable M that equals one for individual i if

individual i's outcome is misreported

$$m_i = \begin{cases} 0 & \text{if } y_i^T = y_i \\ 1 & \text{if } y_i^T \neq y_i \end{cases}$$

We consider two cases, in one the true model is a linear probability model, in the other it is a Probit model. In case of the linear probability model, $E(y|X) = X\beta$, so the researcher would like to estimate the following OLS regression

$$y_i^T = x_i'\beta^{LPM} + \varepsilon_i^{LPM}$$

to obtain the K-by-1 vector $\hat{\beta}^{LPM}$, an estimate of the marginal effects of X on the true outcome. Since y_i^T is not observed this regression is not feasible. Using only the observed data yields the observed model

$$y_i = x_i'\tilde{\beta}^{LPM} + \tilde{\varepsilon}_i^{LPM}$$

Section 2.2 derives this bias in the observed model and shows that it is only zero in special cases.

If the true model is a Probit model, the specification can be motivated by the existence of a latent variable y_i^{T*} such that

$$y_i^T = 1\{y_i^{T*} = x_i'\beta + \varepsilon_i \geq 0\}$$

where ε_i is drawn independently from a standard normal distribution and β is the K-by-1 coefficient vector of interest. Extending our results to other binary choice models in which ε_i is drawn from a different distribution is straightforward. If there is no misreporting, a consistent estimate of the coefficient vector β can be obtained by maximum likelihood. Estimating a Probit model using the observed indicator y_i instead of y_i^T yields $\hat{\tilde{\beta}}$, which

is potentially biased. Little is known about bias due to measurement error in non-linear models (see Carroll et al., 2006). Yatchew and Griliches (1984, 1985) derive some results on misspecification in Probit models. We use their results on the effect of omitted variables and misspecification of the distribution of the error term in the derivation of the bias in section 2.3. The papers mentioned above that propose estimation strategies that are consistent in the presence of misreporting show that ignoring the problem leads to inconsistent estimates, but do not discuss the nature of this inconsistency. Hausman, Abrevaya and Scott-Morton (1998) provide the relation between marginal effects in the observed data and marginal effects in the true data if misreporting is not related to the covariates. They assume that the probabilities of false negatives and false positives conditional on the true response are constants for all individuals, i.e.

$$\begin{aligned} \alpha_{0i} &= \alpha_0 \\ \alpha_{1i} &= \alpha_1 \end{aligned} \qquad \forall i \tag{1}$$

We refer to this kind of misreporting as "conditionally random", because conditional on the true value, y_i^T, misreporting is independent of the covariates X. Hausman, Abrevaya and Scott-Morton (1998) show that under this assumption the marginal effects in the observed data are proportional to the true marginal effects

$$\frac{\partial \Pr(y=1|x)}{\partial x} = (1-\alpha_0-\alpha_1) f(x'\beta)\beta \tag{2}$$

where $f()$ is the derivative of the link function (e.g. the normal cdf in the Probit model and the identity function in the linear probability model), so that $f(x'\beta)\beta$ are the true marginal effects. The constant of proportionality is the same for all elements of β and is linearly decreasing in the two conditional probabilities of misreporting. If, as assumed in Hausman, Abrevaya and Scott-Morton (1998), $\alpha_0 + \alpha_1 < 1$, the marginal effects are attenuated: they are smaller in absolute value in the observed data than the true marginal effects, but retain the correct signs.

This result is informative about the differences between the observed and the true stochastic models, but it is a relationship between the true parameters that does not necessarily extend to estimates of these parameters from a misspecified model. However if one has consistent estimates of the marginal effects in the observed data, $\frac{\partial \widehat{\Pr(y=1|x)}}{\partial x}$, equation (2) implies that they are all attenuated proportionally, which suggests that inference based on coefficient ratios may be valid. Coefficient ratios are informative about the relative magnitude of the coefficients and, if the sign of one coefficient is known, their direction. If one also has consistent estimates of the probabilities of misreporting, $\hat{\alpha}_0$ and $\hat{\alpha}_1$, one can calculate $(1-\hat{\alpha}_0-\hat{\alpha}_1)^{-1}\frac{\partial \widehat{\Pr(y=1|x)}}{\partial x}$. Equation (2) implies that this approach consistently estimates the true marginal effects. However, as discussed further below, if the true model is a Probit, running a Probit on the observed data only yields a consistent estimate of the marginal effects in the observed data in special circumstances. Thus, using the Probit marginal effects from the observed data in (2) usually yields inconsistent estimates of true marginal effects. However, we argue that the inconsistency can be expected to be small in many applications. This problem does not arise in the linear probability model, because a linear probability model on the observed data yields consistent estimates of the marginal effects in the observed data. Additionally, the relation in (2) extends from marginal effects to coefficients, because they are equal in the linear probability model. Consequently, even though (2) is about true parameters and not about bias, it may still be useful to infer something from estimates that use the observed data. Section 3 examines in how far these implications are useful in practice.

2.2 Bias In The Linear Probability Model

Measurement error in binary variables is a form of non-classical measurement error (Aigner, 1973; Bollinger, 1996) and the bias in OLS models when the dependent variable is subject to non-classical measurement error is the coefficient in the (usually infeasible) regression of the measurement error on the covariates (Bound, Brown and Mathiowetz, 2001). In our case,

the dependent variable is binary, so the measurement error takes on the following simple form:

$$u_i = y_i - y_i^T = \begin{cases} -1 & \text{if } i \text{ is a false negative} \\ 0 & \text{if } i \text{ reported correctly} \\ 1 & \text{if } i \text{ is a false positive} \end{cases}$$

Consequently, the coefficient in an OLS regression of the measurement error on the covariates X (if it were feasible) would be:

$$\hat{\delta} = (X'X)^{-1}X'u \tag{3}$$

$\hat{\delta}$ can only be zero if the measurement error is not correlated with X, which is unlikely: If a variable is a relevant regressor, $\Pr(y = 1)$ is a function of X. Since $u = -1$ can only occur if $y = 1$, this creates a dependence between u and X[1]. Equation (3) implies that the coefficient in an OLS regression of the misreported indicator on X, $\hat{\tilde{\beta}}^{LPM}$, is

$$\hat{\tilde{\beta}}^{LPM} = \hat{\beta}^{LPM} + \hat{\delta}$$

Consequently, the bias is

$$\mathbb{E}(\hat{\tilde{\beta}}^{LPM}) - \beta^{LPM} = \mathbb{E}(\hat{\delta}) \tag{4}$$

This implies that a consistent estimate of the expectation of $\hat{\delta}$ is sufficient to consistently estimate the coefficients in the linear probability model with misreporting. Such an estimate could be available from a validation study, but it requires the assumption that misreporting is the same in the sample that is used to obtain the estimate of the expectation of $\hat{\delta}$ and the sample used to estimate β and that the same covariates are used. However, the measurement error only takes on three values, so the formula for $\hat{\delta}$ simplifies to

$$\hat{\delta} = (X'X)^{-1}X'u = (X'X)^{-1}\sum_{i=1}^{N} x_i u_i$$

[1]Note, however, that if misclassification probabilities conditional on y^T depend on X in a peculiar way, $X'u$ can still be 0.

$$
\begin{aligned}
&= (X'X)^{-1} \left(\sum_{\substack{i \text{ s.t. } y_i=1 \\ \& y_i^T=0}} x_i \cdot 1 + \sum_{i \text{ s.t. } y_i=y_i^T} x_i \cdot 0 + \sum_{\substack{i \text{ s.t. } y_i=0 \\ \& y_i^T=1}} x_i \cdot (-1) \right) \\
&= (X'X)^{-1}(N_{FP}\bar{x}_{FP} - N_{FN}\bar{x}_{FN}) \\
&= N(X'X)^{-1} \left(\frac{N_{FP}}{N}\bar{x}_{FP} - \frac{N_{FN}}{N}\bar{x}_{FN} \right)
\end{aligned}
$$

where $\bar{x}_{FP}$ and $\bar{x}_{FN}$ are the means of X among the false positives and false negatives. Consequently in expectation[2]

$$
\begin{aligned}
\mathbb{E}(\hat{\delta}) = N(X'X)^{-1}[&\Pr(y=1, y^T=0)\mathbb{E}(X|y=1, y^T=0) \\
&- \Pr(y=0, y^T=1)\mathbb{E}(X|y=0, y^T=1)] \qquad (5)
\end{aligned}
$$

That is, the bias in $\hat{\hat{\beta}}^{LPM}$ depends on the difference between the conditional means of X among false positives and false negatives where these conditional means are weighted by the probability of observing a false negative or positive. The bias is this vector of differences pre-multiplied by the inverse of the covariance matrix of the data.

Consequently, misclassifying the dependent variable from 1 to 0 at higher values of a particular variable, while holding everything else fixed, decreases the estimated coefficient on that particular variable while misclassifying it from 0 to 1 increases it. The opposite effect occurs at lower values of the variable. Note that this result can be overturned if misclassification also affects the other components of the vector in brackets. Misclassifying more observations (i.e. increasing one of the probabilities of misreporting) amplifies this effect. This discussion also illustrates that the bias will only be zero in knife-edge cases in which the expression in brackets is 0. Neither equal probabilities of misreporting nor (conditional) independence of X and misreporting are sufficient for the bias to be zero. Equation (5) only depends on the probabilities of misreporting and the conditional means of the covariates, so one only needs these quantities to correct the bias. Even if one does not

[2]This assumes that X is non-stochastic. The extension to the stochastic case is straightforward.

know them, one may have an idea about their magnitude from previous studies or theory. In such cases equation (5) can be used to assess the likely direction and size of the bias. It makes sense that the bias depends only on conditional means, because the linear probability model is estimated by OLS, so the parameters are estimated from the (conditional) means only.

If the conditional probabilities of misreporting are constants as in Hausman, Abrevaya and Scott-Morton (1998), the results above simplify to

$$\mathbb{E}(\hat{\tilde{\beta}}_k^{LPM}) = (1 - \alpha_0 - \alpha_1)\beta_k^{LPM} \qquad \forall k \neq 0 \tag{6}$$

for the slope coefficients[3]. This confirms that if the true model is a linear probability model and misreporting is not correlated with X, one can use OLS on the observed data to obtain unbiased estimates of the marginal effects on the observed indicator, $\frac{\partial \Pr(y=1|x)}{\partial x}$. Consequently, in such cases, one can use equation (2) to correct both coefficients and marginal effects for misreporting if one knows the conditional probabilities of misreporting.

2.3 Bias In the Probit Model

While deriving the bias for the linear probability model only required the modification of existing results to a special case, such general results do not exist for non-linear models. We first show that misreporting of the dependent variable is equivalent to a specific form of omitted variable bias and then use results on the effect of omitting variables to decompose the bias due to misreporting. The results presented below are for the Probit model, but the extension to other binary choice models such as the Logit is straightforward.

[3]For the intercept: $\mathbb{E}(\hat{\tilde{\beta}}_0^{LPM}) = \alpha_0 + (1 - \alpha_0 - \alpha_1)\beta_0^{LPM}$. See appendix A for a proof.

2.3.1 Transformation Into An Omitted Variable Problem

The true data generating process without misreporting is assumed to be

$$y_i^T = 1\{x_i'\beta + \varepsilon_i \geq 0\}$$

so that, with m_i the indicator of misreporting, the data generating process for the misreported data is

$$\begin{aligned} y_i &= \begin{cases} 1\{x_i'\beta + \varepsilon_i \geq 0\} & \text{if } m_i = 0 \\ 1\{x_i'\beta + \varepsilon_i \leq 0\} & \text{if } m_i = 1 \end{cases} \quad \Leftrightarrow \\ y_i &= \begin{cases} 1\{x_i'\beta + \varepsilon_i \geq 0\} & \text{if } m_i = 0 \\ 1\{-x_i'\beta - \varepsilon_i \geq 0\} & \text{if } m_i = 1 \end{cases} \end{aligned} \tag{7}$$

Thus, the true data generating process has the following latent variable representation:

$$\begin{aligned} y_i* &= \begin{cases} x_i'\beta + \varepsilon_i & \text{if } m_i = 0 \\ -x_i'\beta - \varepsilon_i & \text{if } m_i = 1 \end{cases} && \Leftrightarrow \\ y_i* &= (1 - m_i)(x_i'\beta + \varepsilon_i) + m_i(-x_i'\beta - \varepsilon_i) && \Leftrightarrow \\ y_i* &= \underbrace{x_i'\beta + \varepsilon_i}_{\text{Well-Specified Probit Model}} \quad \underbrace{-2m_i x_i'\beta - 2m_i\varepsilon_i}_{\text{Omitted Variable}} \end{aligned} \tag{8}$$

The first two terms form a well specified Probit, because ε_i is not affected by misreporting, so it still is a standard normal variable. This transformation into an omitted variable problem is helpful for the results below because Yatchew and Griliches (1984, 1985) discuss omitted variable bias in Probit models. Much of the analysis below follows their arguments applied to the special case of misreporting.

We can decompose each of the omitted variable terms into its linear projection on X and deviations from it:

$$\begin{aligned} -2m_i x_i'\beta &= x_i'\lambda + \nu_i \\ -2m_i\varepsilon_i &= x_i'\gamma + \eta_i \end{aligned} \tag{9}$$

Substituting this back into the equation (8) gives:

$$\begin{aligned} y_i* &= x_i' \underbrace{(\beta + \lambda + \gamma)}_{\text{biased coefficient}} + \underbrace{\varepsilon_i + \nu_i + \eta_i}_{\text{misspecified error term}} \quad \Leftrightarrow \\ y_i* &= x_i'\tilde{\beta} + \tilde{\varepsilon}_i \end{aligned} \tag{10}$$

An immediate implication of (10) is that the observed data do not conform to the assumptions of a Probit model unless $\tilde{\varepsilon}_i$ is drawn independently from a normal distribution that is identical for all i. While $\tilde{\varepsilon}$ is uncorrelated with X and has a mean of zero by construction, it is unlikely that it would have constant variance and cannot come from a normal distribution.[4] Consequently, running a Probit on the observed data does not yield consistent estimates of the marginal effects in the observed data, so that using equation (2) to obtain estimates of the true marginal effects is inconsistent.

In summary, equation (10) highlights three violations of the assumptions of the original Probit model, leading to three effects of misreporting. First, due to the omitted variables, the linear projection of the latent variable is $X\tilde{\beta}$ instead of $X\beta$. Second, the variance of the misspecified error term $\tilde{\varepsilon}$ is different from the variance of the true error term ε. Finally, the error term $\tilde{\varepsilon}$ is not drawn from a normal distribution that is identical for all observations. The next sections discuss the implications of the violation of these assumptions for maximum likelihood estimates of β. We start by deriving an expression for the estimate of $\tilde{\beta}$ that one would obtain if $\tilde{\varepsilon}$ were a regular iid normal error term. There are two reasons for doing this despite the fact that $\tilde{\varepsilon}$ is not iid normal: First, we argue below that this may often provide a useful and tractable approximation to the biased coefficients. Second, equation (10) describes a Probit model with parameter $\tilde{\beta}$ and misspecified functional form of the error term, so a standard Probit does not consistently estimate $\tilde{\beta}$ because of additional bias from the functional form misspecification. However, obtaining formulas for $\tilde{\beta}$ as an intermediate step allows us to apply results from the literature on functional form misspecification in binary

[4] That they cannot be normal can be seen from the fact that the omitted variables have point mass at 0.

choice models (Ruud, 1983, 1986; Yatchew and Griliches, 1985) to derive how parameter estimates of a Probit model will differ from $\tilde{\beta}$ in a second step. Such bias may arise from the difference in the variances of ε and $\tilde{\varepsilon}$, heteroskedasticity or the higher order moments of the distribution of the misspecified error term being different from those of the normal distribution.

2.3.2 Bias In the Linear Projection

The first component of the bias is the result of the coefficients on X incorporating the linear projection of the omitted terms. Two terms are omitted, so the linear projection has two parts that are analogous to the two bias terms Bound et al. (1994) derive for linear models. The first term arises from a relation between misreporting and the covariates X. The second part stems from a relation of misreporting and the error term ε. Equations (9) are linear projections, so they can be analyzed like regression equations except for the fact that they are only conditional expectations in special cases. The familiar linear projection formula gives

$$\hat{\lambda} = -2(X'X)^{-1}X'SX\beta \tag{11}$$

where S is an N-by-N matrix with indicators for misreporting on the diagonal. Equation (11) shows that $\hat{\lambda}$ can be interpreted as minus twice the coefficient on X when regressing a variable that equals the linear index $X\beta$ for misreported observations and 0 for correctly reported observations on X. Under the usual Probit assumptions, $N^{-1}X'X$ converges to the uncentered variance-covariance matrix of X. Let $\operatorname{plim}_{N\to\infty} N^{-1}X'X = Q$, which is positive definite. Additionally, we define the probability limit of the uncentered covariance matrix of X among the misreported observations as $\operatorname{plim}_{N\to\infty} N_{MC}^{-1}(X'X|M=1) = Q_{MC}$. A typical element (r,c) of $X'SX$ is $\sum_{i=1}^{N} x_{ri}m_i x_{ci}$, whereas a typical element (r,c) of $X'X$ is $\sum_{i=1}^{N} x_{ri}x_{ci}$. From the sums in $X'X$, S selects only the x_i that belong to misreported observations, so that

$$\operatorname*{plim}_{N\to\infty}\frac{1}{N}X'SX = \operatorname*{plim}_{N\to\infty}\frac{N_{MC}}{N}\operatorname*{plim}_{N\to\infty}N_{MC}^{-1}\,(X'X|M=1)$$

$$= \Pr(M = 1) Q_{MC}$$

i.e. the term converges to the uncentered covariance matrix of X among those that misreport multiplied by the probability of misreporting. Thus, the probability limit of $\hat{\lambda}$ is

$$\operatorname*{plim}_{N \to \infty} \hat{\lambda} = -2 \Pr(M = 1) Q^{-1} Q_{MC} \beta \tag{12}$$

Equation (12) shows that the bias from this source cannot be zero for all coefficients if there is any misreporting, i.e. if $\Pr(M = 1) \neq 0$ unless all misclassified observations have identical values of all covariates. Otherwise, both right hand side matrices are positive definite, so lambda has positive rank, i.e. it contains non-zero elements. Thus, while some elements of λ can be 0 in special cases, misclassification necessarily induces bias in some coefficients since not all of them can be 0. Multiplication by $-2 \Pr(M = 1)$ creates a tendency for the bias to be in the direction opposite from the sign of the coefficient, which reduces to the rescaling effect if misreporting is not related to X. This effect can be amplified or reduced by $Q^{-1} Q_{MC}$, which introduces an additional (but matrix valued) rescaling factor due to the relation of misreporting to X. Both matrices are positive definite, so the diagonal elements are positive, which creates a tendency for λ and β to have different signs, causing the bias to be in the opposite direction from the sign of the coefficient. However, unless the off-diagonal elements are zero, bias from other coefficients "spreads" and may reverse this tendency.

In summary, the magnitude of the bias depends on three things: all else equal, it is large if the probability of misclassification is large, misclassification comes from a wider range of X or is more frequent among extreme values of X. The second point follows from the fact that in such cases the conditional covariance matrix is large relative to the full covariance matrix. The third effect is due to the covariance matrices being uncentered, so if the mean of X among the misclassified observations differs a lot from that in the general sample, the bias will be larger. This is an intuitive leverage result: Observations that are further from

the center of the data have a larger influence on the linear projection.

The second component of the bias in the linear projection stems from the fact that misclassification may create a relation between X and the error term. Using the standard formula for linear projections, equation (9) and the definition of S from above, it is:

$$\hat{\gamma} = -2(X'X)^{-1}X'S\varepsilon \tag{13}$$

$\hat{\gamma}$ can also be interpreted as minus twice the regression coefficient on X when regressing a vector that contains ε_i for misreporters and zeros for all other observations on X. Using exactly the same arguments as above yields

$$\operatorname*{plim}_{N\to\infty}\hat{\gamma} = -2\Pr(M=1)Q^{-1}\operatorname*{plim}_{N\to\infty}N^{-1}(X'\varepsilon|M=1) \tag{14}$$

While $\operatorname*{plim}_{N\to\infty}N^{-1}X'\varepsilon = 0$ by assumption, this restriction does not determine the *conditional* covariance between X and the error term, $\operatorname*{plim}_{N\to\infty}N^{-1}(X'\varepsilon|M=1)$. The conditional covariance and thus $\operatorname*{plim}_{N\to\infty}\hat{\gamma}$ will be 0 if besides the assumed independence between X and ε it is also true that ε and M are independent as well. If X is independent of ε and M and the model includes an intercept, the slope coefficients will be unbiased. However, if the probability of misreporting depends on the true value y^T, because the determinants of false positives and false negatives differ, the bias is unlikely to be 0. In this case, M can only be independent of X *or* ε and is likely to depend on both.

2.3.3 Rescaling Bias

The second effect of misclassification is a rescaling effect that always occurs when misspecification affects the variance of the error term in Probit models. The coefficients of the latent variable model are only identified up to scale, so one normalizes the variance of the error term to one, which normalizes the coefficients to β/σ_ε. Consequently, misspecification that affects the variance of the error term leads to coefficients with the wrong scale. In the ab-

sence of the additional bias discussed below (i.e. if $\tilde{\varepsilon}$ were iid normal), estimating (10) by a Probit model gives

$$\operatorname*{plim}_{N\to\infty}\hat{\tilde{\beta}} = \frac{\tilde{\beta}}{SD(\tilde{\varepsilon})} = \frac{\beta+\lambda+\gamma}{SD(\varepsilon+\nu+\eta)} \equiv \bar{\beta} \tag{15}$$

One may expect the error components due to misreporting to increase the variance of the error term, i.e. $SD(\tilde{\varepsilon}) > SD(\varepsilon)$, so the rescaling will tend to result in a bias towards zero. However, cases in which the variance decreases are possible if misreporting depends on ε. The rescaling factor is the same for all coefficients, so it does not affect their relative magnitudes.

2.3.4 Bias Due To Misspecification of the Error Distribution

If $\tilde{\varepsilon}$ were iid normal, estimating equation (10) by a Probit model would yield a consistent estimate of $\bar{\beta}$ as given by (15). However, as was discussed above, $\tilde{\varepsilon}$ will not have the normality and homoskedasticity of ε, so that (10) additionally suffers from misspecification of the error term. This causes bias in addition to the one given by (15), i.e. one cannot obtain a consistent estimate of $\bar{\beta}$ by running a Probit on the observed data. Ruud (1983, 1986) characterizes this bias and discusses special cases in which the bias is proportional for all coefficients, but closed form solutions for the bias due to misspecification of the error distribution do not exist. Adapting a result from Yatchew and Griliches (1985) to our case provides an implied formula for the exact bias. Taking the probability limit of N^{-1} times the first order conditions of the log-likelihood function, the parameter estimate converges to the vector b that solves

$$\int f_X(x_j)\frac{x_j'\phi(x_j'b)}{\Phi(x_j'b)(1-\Phi(x_j'b))}\left[F_{\tilde{\varepsilon}|X=x_j}(-x_j'\bar{\beta}) - \Phi(-x_j'b)\right]\,\mathrm{d}x_j = 0 \tag{16}$$

where $F_{\tilde{\varepsilon}|X=x_j}$ is the conditional cumulative distribution function of $\tilde{\varepsilon}/Var(\tilde{\varepsilon})$, i.e. the misspecified error term normalized to have (unconditional) variance 1, evaluated at $X = x_j$. Consequently, $F_{\tilde{\varepsilon}|X=x_j}(-x_j'\bar{\beta})$ provides the probability that $\tilde{\varepsilon}_i/Var(\tilde{\varepsilon})$ is smaller than $-x_j'\bar{\beta}$ in the sub-population with a specific value of the covariates ($X = x_j$). Thus, it provides the

probability of observing $y = 1$ when drawing from the sub-populations with covariates equal to x_j. Note that the left hand side of (16) is the first derivative of a concave function, so the equation has a unique solution.

If $F_{\tilde{\varepsilon}|X}$ is a normal cdf with the same variance for all values of X, $b = \bar{\beta}$ solves (16) so that (15) gives the exact bias. Unfortunately, (16) has no closed form solution and can only be solved numerically for specific cases of $F_{\tilde{\varepsilon}|X}$ which is usually unknown. Note, however, that the unconditional distributions, $F_{\tilde{\varepsilon}}$ and Φ have the same first and second moments by construction. Consequently, (asymptotic) deviations of the parameter estimates from $\bar{\beta}$ only occur due to a dependence of the first two moments of $F_{\tilde{\varepsilon}}$ on X (e.g. heteroskedasticity) and differences in higher order moments of the two distributions (so we refer to this bias component as the "higher order bias").

We find the bias due to misspecification of the error term to be small in the Monte Carlo simulations in the next section, but these biases can become large if misclassification depends heavily on ε or the true value of y. Both are likely to induce asymetries (such as skewness) in $F_{\tilde{\varepsilon}}$, which creates differences between $F_{\tilde{\varepsilon}|X}$ and Φ that are unlikely to even out over the sample. If functional form misspecification can be largely ignored, the expressions for the bias are easier to analyze. In this case, one would also be justified using the result from Hausman, Abrevaya and Scott-Morton (1998) to obtain an approximation to the true marginal effects by estimating the observed marginal effects using a Probit model and scaling them up according to equation (2). The observed marginal effects will not be consistent, but if misreporting is conditionally random and the higher order bias is small, this may still provide a good approximation. Appendix B discusses how to further assess the likely direction and severity of this component of the bias and conditions under which one can expect (15) to provide a good approximation to the bias.

3 The Bias in Practice

This section uses administrative data matched to two major surveys to illustrate the applicability of the analytic results from the previous section. Misreporting is correlated with the covariates in the matched data, but we also examine misreporting that is conditionally random by conducting a Monte Carlo study in which we induce misreporting in the data using the probabilities of misreporting we observe in our matched sample. This exercise leaves the remainder of the data unchanged, so it allows us to focus on misreporting that is uncorrelated with the explanatory variables. Finally, we perform a Monte Carlo exercise using simulated data in order to assess the size of the components of the bias in the Probit model with correlation in more detail, particularly conditions under which the higher order bias tends to be small.

We use the data employed in Meyer, Goerge and Mittag (2014): the 2001 American Community Survey (ACS) and the 2002-2005 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) matched to food stamp administrative data from Illinois and Maryland. See Meyer, Goerge and Mittag (2014) for details on the data and the matching process as well as an analysis of the determinants of misreporting. We estimate simple Probit and linear probability models of food stamp take-up at the household level with three covariates: a continuous poverty index (income/poverty line) as well as dummies for whether the head of the household is 50 or older and whether the household is in Maryland. We restrict the sample to matched households with income less than 200% of the federal poverty line and adjust the survey weights for the non-random match probability as in Meyer, Goerge and Mittag (2014). Throughout this paper, we treat the administrative food stamp indicator as truth, even though it may contain errors such as those due to an imperfect match to the survey data. Given that there should be few mistakes in the administrative records and the match rate is high, this assumption seems plausible. Additionally, most of the analysis below does not require this assumption - the survey data can be considered as a misreported version of the administrative data even if neither of them represent "truth".

Table 1: Bias in the Linear Probability Model, ACS

	(1) Matched Data	(2) Survey Data	(3) Bias MC Study (random MR)	(4) Bias Survey Data	(5) Bias due to correlation
Poverty index	-0.0018	-0.0019	-27.94%	5.56%	33.50%
	(0.0001)	(0.0001)			
Age≥50	-0.1166	-0.1046	-28.93%	-10.29%	18.64%
	(0.0145)	(0.0133)			
Maryland	-0.0034	-0.0217	-10.03%	538.24%	548.26%
	(0.0157)	(0.0141)			

Note: Sample size: 5945 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All biases are in % of the coefficient from matched data. In the MC design, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.02374 and Pr(FN)=0.2596). 500 replications are performed and we report the average bias in %.

Table 2: Bias in the Linear Probability Model, CPS

	(1) Matched Data	(2) Survey Data	(3) Bias MC Study (random MR)	(4) Bias Survey Data	(5) Bias due to correlation
Poverty index	-0.0023	-0.0021	-42.18%	-8.70%	33.48%
	(0.0002)	(0.0001)			
Age≥50	-0.1264	-0.0985	-41.79%	-22.07%	19.72%
	(0.0174)	(0.0151)			
Maryland	-0.0937	-0.0706	-42.60%	-24.65%	17.95%
	(0.0184)	(0.0156)			

Note: Sample size: 2791 matched households from IL and MD with income less than 200% of the federal poverty line. All analyses include a constant term (not reported) and are conducted using household weights adjusted for match probability. All biases are in % of the coefficient from matched data. In the MC design, the dependent variable is administrative FS receipt with misreporting induced with the misreporting probabilities observed in the actual sample (Pr(FP)=0.03271 and Pr(FN)=0.3907). 500 replications are performed and we report the average bias in %.

For the linear probability model we have obtained closed form solutions for the bias that are straightforward to analyze in both the conditionally random and the correlated case. The results for both cases are presented in table 1 for the ACS and table 2 for the CPS and conform to the expectations from section 2.2. The results from the Monte Carlo study in

column (3) confirm that if misreporting is conditionally random, all slopes are attenuated by the same factor. In both surveys, this factor is close to its expectation given by equation (2): $1 - \alpha_0 - \alpha_1$. The obvious exception is the coefficient on the Maryland dummy in the ACS, for which the rescaling factor is clearly different. This is due to the fact that this coefficient is very imprecisely estimated and basically indistinguishable from 0. As is evident from column (4), this proportionality does not hold in the actual survey data, where misreporting is related to the covariates. The bias in the correlated case is smaller for all coefficients except for the imprecise Maryland dummy in the ACS, indicating that the biases partly cancel. Since the only difference in the data used for column (3) and (4) of table 1 and 2 is the correlation between the misreporting and the covariates, the difference in the biases is an estimate of the bias induced by correlation. This difference is presented in column (5) and in our case biases all coefficients away from 0. In both the random and the correlated case, the bias is always equal to $\hat{\delta}$ as defined by equation (3) (results not presented).

Table 3 examines in how far the implications of equation (2) allow us to learn something from the biased coefficients as suggested in section 2.1. The first three columns present coefficient ratios, since equation (2) suggests that if misclassification is conditionally random the constant of proportionality may cancel. The table contains estimates of the "true" coefficient ratios from the matched data (Column 1), the bias in the ratios from the survey data (Column 3, in percent of the true ratio) as well as the average bias in percent of the true ratio from simulating conditionally random misreporting as described above (Column 2). If the conditional probabilities of misreporting are known, one can multiply the coefficients by the constant of proportionality, $(1 - \alpha_0 - \alpha_1)^{-1}$ to obtain estimates of the true coefficients. The last two columns examine whether this improves the estimates in simulations where misreporting is conditionally random (Column 4) and in the estimates from the survey data, where misreporting is related to the covariates. Considering the bias found in table 1 and 2, the results are not surprising: Columns (2) and (5) show that if misreporting is indeed not related to the covariates, their ratios and scaled up versions indicate the right relative

magnitudes and signs of the coefficients. However, a small bias remains for all coefficients, which becomes sizable for the imprecisely estimated coefficients. As one would expect, the results in columns (3) and (5) show that the bias becomes much worse if the assumption that misreporting is conditionally random fails. As we have seen above, the correlation of misclassification with the covariates partly cancels the attenuation effect in our application, so that the rescaling factor under the assumption of no correlation induces an upward bias.

Table 3: What Can Be Learned From Survey Coefficients, LPM

	(1)	(2)	(3)	(4)	(5)
	Coefficient Ratios (relative to age coefficient)			Bias Rescaled Marginal Effects	
	Matched Data	Bias MC Study	Bias in Survey	MC Study	Survey
			ACS		
Poverty index	0.0154	1.41%	17.67%	0.37%	36.84%
Age≥50	-	-	-	-0.85%	39.48%
Maryland	0.0292	12.82%	611.46%	25.95%	39.63%
			CPS		
Poverty index	0.0182	3.51%	17.17%	-0.29%	71.43%
Age≥50	-	-	-	0.97%	73.50%
Maryland	0.7413	5.56%	-3.31%	-0.44%	73.51%

Note: See note Table 1 and 2

The results for the Probit models are presented in table 4 for the ACS and in table 5 for the CPS. Column (3) shows that, as in the linear probability model, coefficients are attenuated by the same factor if misreporting is conditionally random and the coefficient is reasonably precisely estimated. The rescaling factor is different from $1 - \alpha_0 - \alpha_1$, because coefficients and marginal effects are not equal in the Probit model. As was discussed above, due to the higher order bias in the Probit model, (2) does not hold between Probit estimates, but only between true parameters. So contrary to the linear probability model, this proportionality is not necessarily the case and may not generalize to other applications.

Column (4) underlines that both the proportionality and the attenuation only apply if misreporting is conditionally random: The biases are different and some of them are positive,

Table 4: Bias in the Probit Model, ACS

	(1) Matched Data	(2) Survey Data	(3) Bias MC Study (random MR)	(4) Bias Survey Data	(5) Bias due to correlation
Poverty index	-0.0060	-0.0071	-20.51%	18.33%	38.84%
	(0.0004)	(0.0005)			
Age≥50	-0.4062	-0.4167	-22.15%	2.58%	24.74%
	(0.0512)	(0.0543)			
Maryland	-0.0187	-0.0978	-9.08%	422.99%	432.07%
	(0.0548)	(0.0582)			

Note: See note Table 1

Table 5: Bias in the Probit Model, CPS

	(1) Matched Data	(2) Survey Data	(3) Bias MC Study (random MR)	(4) Bias Survey Data	(5) Bias due to correlation
Poverty index	-0.0074	-0.0083	-32.47%	12.16%	44.63%
	(0.0005)	(0.0006)			
Age≥50	-0.4297	-0.3992	-32.17%	-7.10%	25.07%
	(0.0614)	(0.0682)			
Maryland	-0.3338	-0.3189	-33.15%	-4.46%	28.69%
	(0.0736)	(0.0808)			

Note: See note Table 2

indicating a bias away from zero. The bias in the survey data is again smaller in absolute value than the bias without correlation. Column (5) confirms that the bias induced by correlation is in the opposite direction of the bias in the conditionally random case, so the two biases partly cancel. This explains why Meyer, Goerge and Mittag (2014) find that the bias in a model of program take-up with the same data but many controls is relatively small given the extent of misreporting in the data. As should be evident from equations (12) and (14), whether correlation between misclassification and the covariates reduces or exacerbates the bias depends on the sign and magnitude of the correlations and the true coefficients.

Table 6 examines the implications of equation (2) for Probit models. Columns (1)-(3) and (6)-(7) are the equivalent of table 3, i.e. they present coefficient ratios and rescaled

marginal effects. Columns (4) and (5) contain Probit estimates of the observed marginal effects in the matched data and the survey data, since contrary to the LPM, coefficients and marginal effects differ. Contrary to the LPM, using a Probit on the survey data does not yield consistent estimates of the observed marginal effects in the presence of misclassification due to the rescaling and higher order bias. Nonetheless, the results in table 6 are qualitatively similar to the results for the LPM in table 3: only a small bias remains in the conditionally random cases in columns (2) and (6), which supports our conjecture that the bias from functional form misspecification are small here. However, both ratios and rescaled coefficients are substantively biased if misclassification is related to X. Taken together, the results in tables 3 and 6 show that scaling-up may be a promising strategy if there is no correlation, but that it fails if there is correlation.

Table 6: What Can Be Learned From Survey Coefficients, Probit

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Coefficient Ratios (relative to age coefficient)			Observed Marginal Effects		Bias Rescaled Marginal Effects	
	Matched Data	Bias MC Study	Bias in Survey	Matched Data	Survey Data	MC Study	Survey Data
				ACS			
Poverty index	0.0148	2.19%	15.35%	-0.0018	-0.0017	1.39%	38.89%
Age≥50	-	-	-	-0.1033	-0.1164	-0.58%	39.50%
Maryland	0.0460	9.40%	409.82%	-0.0242	-0.0053	3.66%	39.67%
				CPS			
Poverty index	0.0172	4.84%	20.73%	-0.0019	-0.0021	1.98%	73.68%
Age≥50	-	-	-	-0.0902	-0.1214	2.76%	73.50%
Maryland	0.7768	6.32%	2.84%	-0.0721	-0.0943	0.95%	73.37%

Note: See note Table 1 and 2

It is difficult to interpret the bias in the correlated case, because the formulas derived above for the Probit model offer few general insights on the relative magnitude of the components of the bias. For example, it seems likely that attenuation is more pronounced in the CPS survey coefficients because the rate of misreporting is higher than in the ACS, but this result could be altered by differences in the other bias components. The numbers in

column (5) of tables 4 and 5 suggest that these components are similar, but not the same in the two surveys. In order to obtain more evidence on the determinants and relative sizes of these components, we perform a simulation study that allows us to control for the factors that cause the bias components.

We chose parameters that generate data with a similar structure as the observed data. In particular, we generate two continuous covariates (x_1 and x_2). x_1 is drawn from a standard normal distribution, x_2 is generate from a normal distribution with mean $0.1x_1$ and variance 1, so it is mildly correlated with x_1. The dependent variable is generated according to the Probit model $y^T = 1\{a + 0.5x_1 + 0.5x_2 + e \geq 0\}$ where e is drawn from a standard normal distribution. The intercept a is chosen such that the mean of y^T is always 0.25. We generate misclassification according to $m = 1\{c + bx_1 + e_{MC} \geq 0\}$ where e_{MC} is another standard normal variable.[5] Consequently, the misreporting model is the same for false negatives and false positives in order to keep the results simple and tractable by isolating certain sources of bias. We also ran simulations in which misreporting depends on y^T, but they make m depend on e and induce an additional source of dependence between m and X. This makes the results hard to interpret, since they confound multiple sources of bias. We briefly discuss how the results from these and other more complex models differ from the simpler models presented here at the end of this section, the complete results are available upon request.

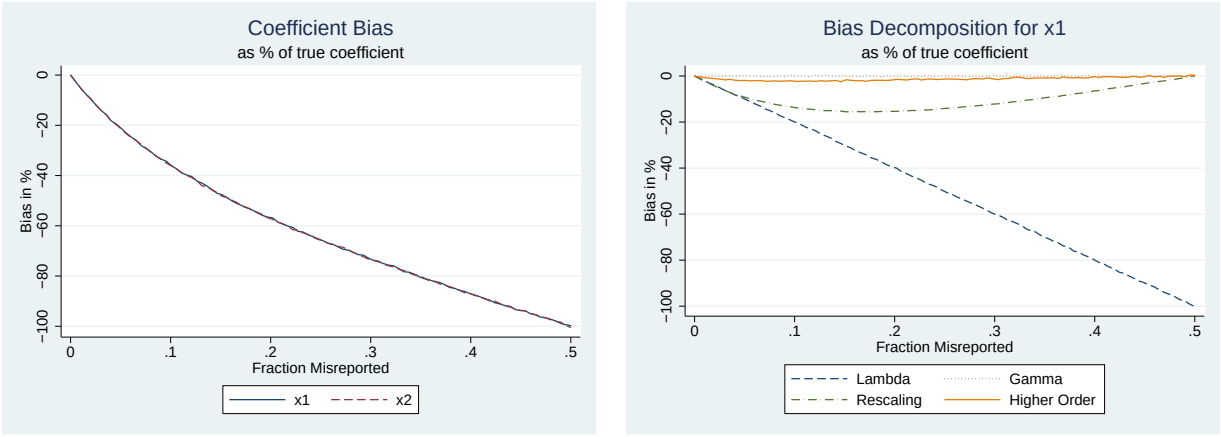
We use this MC design to examine three cases. In the first and second case, we increase the level of misreporting from 0% to 50% ($\alpha_0 = \alpha_1 = .5$) holding everything else constant. In the first case, b equals 0 so misreporting is not related to x_1, while in the second case b is constant at 0.25 creating a modest correlation. In the third case, we hold the level of misreporting fixed at 30% and increase b from -1 to 1, so that the correlation between x_1 and m increases from roughly -0.5 to 0.5. In all cases we take 100 equally spaced grid points over the parameter space. At each point, we draw 100 samples of 10,000 observations and

[5]The results are similar when generating x_1 and x_2 from uniform or χ^2 distributions or making one of them a dummy variable. They are extremely close when drawing e_{MC} from other distributions such as the uniform, t- or χ^2_1- distribution. Note that misclassification is related to x_2 only through the correlation between x_1 and x_2.

run a Probit on y^T to obtain $\hat{\beta}$ and on the misreported data (yielding $\hat{\beta}^{MC}$).[6] We record the bias as the difference between the two estimated coefficient vectors $(\hat{\beta} - \hat{\beta}^{MC})$ as well as the bias due to $\hat{\lambda}$ as given by equation (11), $\hat{\gamma}$ as given by (13) and the rescaling bias implied by (15): $(\hat{\beta}^{MC} + \hat{\lambda} + \hat{\gamma})/SD(\hat{\varepsilon} + \hat{\nu} + \hat{\eta}) - (\hat{\beta}^{MC} + \hat{\lambda} + \hat{\gamma})$. We are particularly interested in the higher order bias, in order to assess when it may be small enough to be ignored. Calculating this bias analytically requires solving (16) numerically, so we calculate it as the "residual" bias, i.e. $(\hat{\beta}^{MC} - \hat{\lambda} - \hat{\gamma}) \cdot SD(\hat{\varepsilon} + \hat{\nu} + \hat{\eta}) - \hat{\beta}$.

Figures 1-3 show the bias of the two coefficients in the left panel and the decomposition of the bias in the coefficient on x_1 into the four components in the right panel. The results are as expected based on the analytic results in the previous section. Our misreporting model generates misreporting that is independent of ε, so $\hat{\gamma}$ is zero as expected. The results confirm that coefficients are attenuated with the correct sign for reasonable levels of misreporting and correlation, but that this does not have to hold in general. We find that the higher order bias is small in all cases, but discuss circumstances under which this will not be the case below.

Figure 1: MC Design 1 - Uncorrelated Misreporting (b=0)

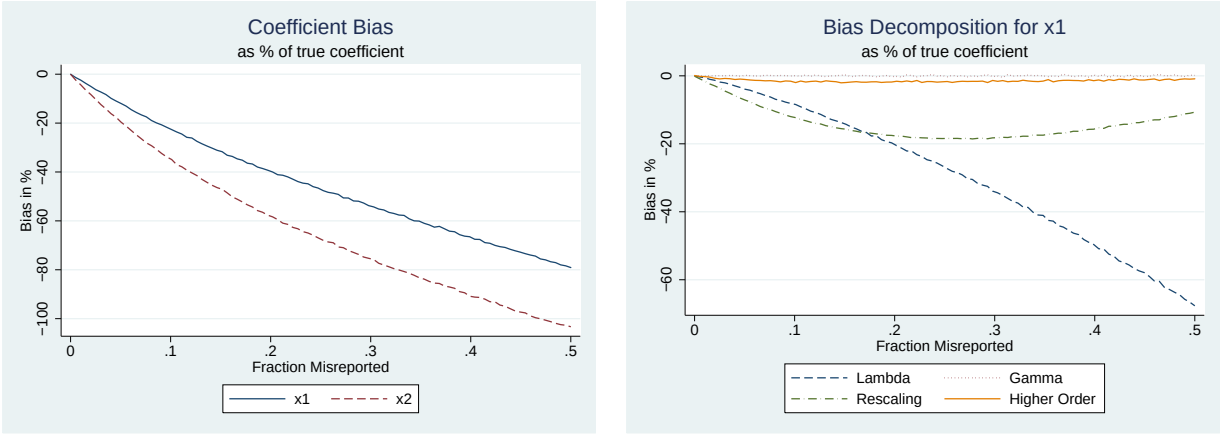


The left panel of figure 1 shows that in the uncorrelated case, the bias is the same for both coefficients (in relative terms) and always between 0% and -100%, i.e. both coefficients

[6]Throughout, we compute the bias as deviations from $\hat{\beta}$ rather than β to reduce the effect of sampling variation in X and m.

are always attenuated. As one would expect, it increases continuously until both coefficients are exactly 0 when 50% of all observations are misreported, implying that the dependent variable has no information content. It also shows that the bias is not increasing linearly in the sum of the conditional probabilities of misreporting, α_0 and α_1, which underlines that equation (2) is a relation between true values and not between Probit estimates. The bias decomposition in the right panel shows why this is the case: The bias is mainly due to $\hat{\lambda}$, which is linear and has a slope of -2 as predicted by (12). Additionally, there is a rescaling bias that is non-linear in the level of misreporting. As misreporting increases, the variance of $\tilde{\varepsilon}$ increases, but the absolute value of $\hat{\tilde{\beta}}$ decreases. As can be seen from (15), the former makes the rescaling bias larger, the latter decreases it again. Overall, the rescaling bias is small relative to the whole bias and the higher order bias remains negligible for all levels of misreporting, which suggests that corrections based on equation (2) would not be far off here.

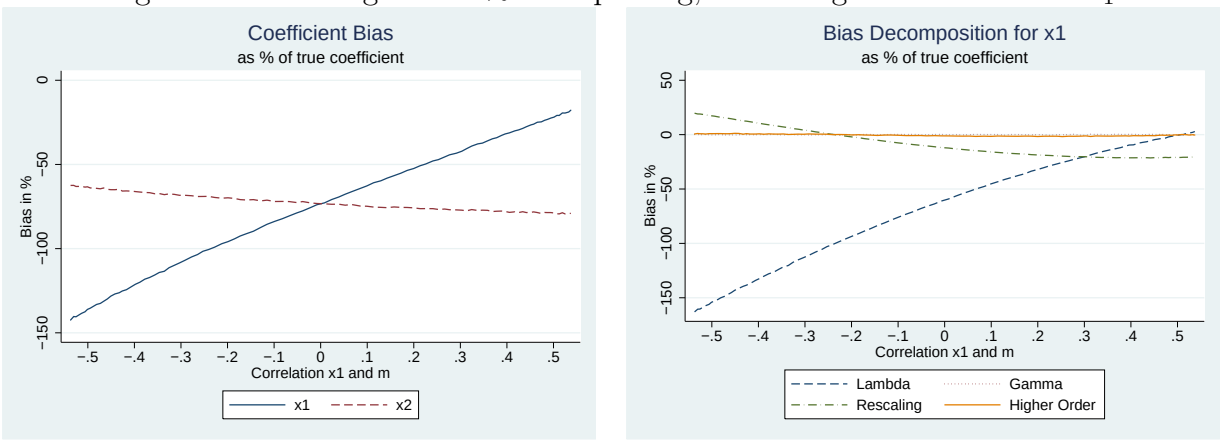
Figure 2: MC Design 2 - Correlated Misreporting (b=-0.005)



The only difference in the simulations behind figure 2 is that the coefficient on x_1 in the misreporting model is held fixed at 0.25, i.e. misreporting is modestly related to x_1. It is impossible to hold both the coefficient and the correlation constant as the level of misreporting increases. We chose to hold the coefficient fixed, which causes the correlation to decrease from 0 to -0.2 as misreporting increases from 0% to 50%. The main difference from the uncorrelated case shown in figure 1 is that the coefficient on x_1 is less severely

biased. As in the food stamp data, the bias due to correlation reduces the bias from $\hat{\lambda}$, but is not strong enough to override it. Consequently, the coefficient is attenuated over the whole range, but does not go to zero even when misreporting reaches 50%. This finding does not generalize. If one chooses b such that the (negative) correlation is stronger, the coeffcient can be biased away from zero and if b is positive, the bias is more severe than in figure 1 so that the coeffcient can change its sign. The coefficient on x_2 is almost unaffected. The bias decreases by 0-1.5 percentage points, which we would expect given that the correlation is positive and low. The bias decomposition is similar to the previous case: the effect is mainly driven by $\hat{\lambda}$ and the higher order bias is small. Contrary to the previous case, $\hat{\lambda}$ is not linear, but concave. This pattern is due to the fact that the correlation between x_1 and m is increasing at a decreasing rate, so its tendency to reduce the bias increases at a decreasing rate. The bias due to rescaling looks different, but this difference is mainly due to its denominator (which is less attenuated and thus bigger in absolute value). The variance of $\tilde{\varepsilon}$ is very similar in the two cases.

Figure 3: MC Design 3 - 30% misreporting, increasing correlation with x_1



In the third case misreporting stays constant at 30%, but we strengthen the correlation between x_1 and m by raising b from -1 to 1, which makes the correlation increase from -0.54 to 0.54. The key insight from the results in figure 3 is that most of the regularities we have stressed can be overturned for extreme cases: With extreme negative correlation, the bias exceeds -100% so that the coefficient changes sign and hence the rescaling bias is away from

zero. For high positive values of the correlation the bias from $\hat{\lambda}$ is away from zero. If the off-diagonal elements in (12) are negative and large relative to the diagonal elements, some components of $\hat{\lambda}$ can become positive and cause a bias away from 0. If we further increase this correlation, e.g. by allowing the models for false positives and false negatives to differ, this effect can become strong enough to override the attenuation bias, so that the estimate is biased away from zero. This simulation shows that while not always true, coefficients tend to retain their sign but are attenuated for a range of reasonable correlations even at a relatively high level of misreporting (30%). As in the two previous cases, the higher order bias is small and the changes in the rescaling bias are mainly driven by the changes in its numerator, as the variance of $\tilde{\varepsilon}$ only decreases from 1.46 to 1.25.

The results from the last simulation raise the question in how far the patterns in these simulations are artifacts of the simplified simulation setups. In order to assess this question, we conducted several additional simulation designs in which we made m depend on ε, X and ε, and varied the predictive power of the outcome model as well as $\Pr(y = 1)$. We repeated all simulations with false positives or false negatives only (as an extreme case of models of misclassification that depends on y^T) and generated the covariates from other distribution and with non-linear relationships between x_1 and x_2. Results are available upon request. While one can never generalize from such exercises, they underline some features of the results above that are simplifications to ease exposition and that do not hold more generally. It should be clear from the setup that $\hat{\gamma}$ is zero in the simulations above by design. If misclassification depends on both X and ε, for example by generating only false positives or false negatives, $\hat{\gamma}$ becomes non-zero and behaves as one would expect based on equation (13). Similarly, the higher order bias is not small in all cases. Noticeable higher order bias usually arises when misclassification depends on ε, for example when the models for false positives and false negatives differ. If becomes sizable in our simulations when misclassification has an asymmetric effect on the distribution of ε, i.e. if the probability of misclassification for large values of the error term in the outcome equation is higher or lower

than for small values. This is in line with the symmetric weighting function in equation (16), which suggests that symmetric deviations of $F_{\varepsilon|X}$ and Φ may average out over the sample. Under similar circumstances, we find several cases in which the rescaling bias is not toward zero. While it is unclear in how far the results regarding the higher order bias generalize, the rescaling bias increases the coefficient in absolute value whenever misclassification decreases the variance of ε. This is likely to happen when misclassification is monotonically related to ε, since this shifts weight from one tail of the distribution of the error to the other, which usually decreases the variance of symmetric distributions.

While we find that the general patterns of the simulations are robust to changes in the data generating process, the usual caveat of simulations as special and simplified cases applies. However, one of the most robust findings is that slope coefficients only change sign if misreporting is strongly related to the covariates or ε in particular ways: none of the coefficients in any of our applications changes sign, and it only happens for some extreme cases in the Monte Carlo studies. While one of the components of the bias sometimes biases the coefficients away from zero, an overall bias away from zero only seems to arise in cases where M is highly correlated with X or ε. In addition, the coefficient estimates tend to be attenuated, i.e. lie between 0 and the true coefficient. This is always the case when misclassification is conditionally random, but can be overturned if it is strongly related to X. If one can rule out such cases, the estimates can be interpreted as lower bounds for the true coefficients and one may be able to infer the sign of the coefficients, which is often of key interest.

In this section, we have illustrated how the bias components derived above affect coefficients in real data both if misreporting is conditionally random and if it is related to the covariates. In our application to food stamp take up, the correlation with covariates reduces the bias, which explains why Meyer, Goerge and Mittag (2014) find relatively small bias and shows that correlation with the covariates is not necessarily bad. We have found that the linear probability model and the Probit model bias can be assessed by the analytic formulas,

even though this requires some insight into the likely magnitude of the different components if misreporting is related to the covariates. We have lined out conditions when the higher order bias in the Probit model tends to be small, which would allow one to focus on the more tractable parts of the bias, although one should be cautious in generalizing from MC studies. Overall, the results underscore the conjectures that the signs of coefficients are robust to a wide range of misreporting mechanisms and that there is a tendency for the coefficients to be attenuated even though the results in section 2.3 shows that none of these results should be expected to hold in all cases.

4 Consistent Estimators

This section introduces estimators for the Probit model that are consistent under different assumptions and tests their performance. We focus on the Probit model, because it is the most common parametric model and other maximum likelihood estimators can be corrected in similar ways. Some semiparametric estimators that relax the normality assumptions have been proposed (e.g Hausman, Abrevaya and Scott-Morton, 1998), but misspecification and misclassification are different problems and we focus on the latter here, so we do not examine the performance of semiparametric estimators. Section 3 showed that corrections for the linear probability model work well if an estimate of $\hat{\delta}$ is available and to our knowledge no other consistent estimators have been proposed, so we do not examine corrections for the linear probability model.

We use six different estimators for the Probit model: three estimators are only consistent under conditionally random misreporting, and three are still consistent if misreporting is related to X. Section 4.1 describes the estimators, which are all variants of estimators that have been proposed elsewhere, section 4.2 evaluates their performance under conditionally random misreporting, section 4.3 allows misreporting to be related to X.

4.1 Theory

All estimators can be derived from equation (7), which implies that the probability that the observed outcome y_i is one is $\Pr(y_i = 1) = \alpha_{0i}(1 - \Phi(x_i'\beta)) + (1 - \alpha_{1i})\Phi(x_i'\beta)$. Thus, the probability distribution of the observed outcome y_i can be written as:

$$\Pr(y_i) = [\alpha_{0i} + (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)]^{y_i} + [1 - \alpha_{0i} - (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)]^{1-y_i}$$

This immediately implies the log-likelihood function of the observed data:

$$\begin{aligned}\ell(\alpha, \beta) = &\sum_{i=1}^{N} y_i \ln\left(\alpha_{0i} + (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)\right) + \\ &(1 - y_i)\ln\left(1 - \alpha_{0i} - (1 - \alpha_{0i} - \alpha_{1i})\Phi(x_i'\beta)\right)\end{aligned} \tag{17}$$

The parameters of this likelihood are not identified, as there are $2N + K$ parameters (two αs for each observation plus the K-by-1 vector β). Hausman, Abrevaya and Scott-Morton (1998) assume that the conditional probabilities of misclassification (α_{0i} and α_{1i}) are constants as in (1), which reduces the log-likelihood function in (17) to

$$\begin{aligned}\ell(\alpha_0, \alpha_1, \beta) = &\sum_{i=1}^{N} y_i \ln\left(\alpha_0 + (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta)\right) + \\ &(1 - y_i)\ln\left(1 - \alpha_0 - (1 - \alpha_0 - \alpha_1)\Phi(x_i'\beta)\right)\end{aligned} \tag{18}$$

This likelihood function only includes $K+2$ parameters: α_0, α_1 and β. They show that these parameters are identified due to the non-linearity of the normal cdf as long as $\alpha_0 + \alpha_1 < 1$. The parameters can be consistently estimated by maximum likelihood or non-linear least squares. We refer to this estimator as the HAS-Probit. Their assumption that the probabilities of misreporting are constants implies that α_0 is the population probability of a false positive and α_1 is the population probability of a false negative. These probabilities or estimates of them may be known from validation data or other out of sample information.

Let $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ denote such estimates. As in Imbens and Lancaster (1994) this information can be incorporated in the estimation procedure. In our application below, the probabilities can be considered known, because they are either calculated within sample or known from data on the whole population. Thus, this approach simplifies to plugging $\tilde{\alpha}_0$ and $\tilde{\alpha}_1$ into (18), yielding the likelihood of our second estimator:

$$\ell(\beta) = \sum_{i=1}^{N} y_i \ln\left(\tilde{\alpha}_0 + (1 - \tilde{\alpha}_0 - \tilde{\alpha}_1)\Phi(x_i'\beta)\right) + (1 - y_i)\ln\left(1 - \tilde{\alpha}_0 - (1 - \tilde{\alpha}_0 - \tilde{\alpha}_1)\Phi(x_i'\beta)\right) \tag{19}$$

Poterba and Summers (1995) take a similar approach to a model of labor market transitions. If it is unreasonable to assume that α_0 and α_1 are known, one should use one of the procedures Imbens and Lancaster (1994) suggest to obtain standard errors, because the usual standard errors are inconsistent as Hausman, Abrevaya and Scott-Morton (1998) point out. Meyer, Mok and Sullivan (2009) provide an alternative estimate of α_1 that we denote as $\tilde{\alpha}_1'$. They assume that $\tilde{\alpha}_0$ is small and can be ignored and calculate $\tilde{\alpha}_1'$ as the ratio of the population weighted number of people who report receipt of a program to the number of people who receive it according to administrative totals. While assuming $\tilde{\alpha}_0 = 0$ is a likely misspecification, estimates of the needed ratio are often available when separate estimates of α_0 and α_1 are not. Thus, we also examine the performance of the estimator that maximizes (18) with α_0 constrained to 0 and α_1 constrained to $\tilde{\alpha}_1'$:

$$\ell(\beta) = \sum_{i=1}^{N} y_i \ln\left((1 - \tilde{\alpha}_1')\Phi(x_i'\beta)\right) + (1 - y_i)\ln\left(1 - (1 - \tilde{\alpha}_1')\Phi(x_i'\beta)\right) \tag{20}$$

All three estimators assume that conditional on truth, misreporting is independent of the covariates. The unconstrained HAS-Probit teases out $\hat{\alpha}_0$ and $\hat{\alpha}_1$ from the observed binary responses, while the other two estimators constrain these parameters based on outside information. The main benefit of using outside information is the reduction in computational

complexity and increased robustness in the presence of functional form misspecification, because separating the probabilities of participation and misreporting no longer heavily depends on the functional form assumptions. These approaches may or may not increase the efficiency of the estimators (Imbens and Lancaster, 1994).

In many cases, the assumption that misreporting is conditionally random does not hold. One can allow the misreporting probabilities to depend on X if one can predict α_{0i} and α_{1i} based on outside information. Such predictions could be obtained by using the parameters from models of misreporting that use validation data (e.g. Meyer, Goerge and Mittag, 2014; Marquis and Moore, 1990) to predict $\hat{\alpha}_{0i}$ and $\hat{\alpha}_{1i}$ in the misreported data. As Bollinger and David (1997) show, these predicted probabilities can be used in the pseudo-likelihood

$$\ell(\beta) = \sum_{i=1}^{N} y_i \ln\left(\hat{\alpha}_{0i} + (1 - \hat{\alpha}_{0i} - \hat{\alpha}_{1i})\Phi(x_i'\beta)\right) + (1 - y_i)\ln\left(1 - \hat{\alpha}_{0i} - (1 - \hat{\alpha}_{0i} - \hat{\alpha}_{1i})\Phi(x_i'\beta)\right) \tag{21}$$

They show that maximization of this likelihood yields consistent estimates of β. We refer to this estimator as the predicted probabilities estimator. Bollinger and David (1997) also correct the standard errors for the estimation error in the parameters used to predict the probabilities of misreporting. This correction requires the samples used to estimate the misreporting parameters and β to be independent, which is not the case in our application, so we choose to bootstrap the standard errors. Our results confirm their finding that the correction has a very small effect on the SEs.

The predicted probabilities estimator does not require the researcher to have access to the validation data used to estimate the probabilities of misreporting, but if both the validation data and the data used to estimate the outcome model are available, one could estimate the misreporting model and the outcome model jointly. Assuming that misreporting can be

described by single index models, the two models imply a system of 3 equations:

$$\begin{aligned}
\Pr(y_i|y_i^T=0,x_i^{FP}) &= \left[F^{FP}(x_i^{FP\prime}\gamma^{FP})\right]^{y_i} + \left[1-F^{FP}(x_i^{FP\prime}\gamma^{FP})\right]^{1-y_i} \\
\Pr(y_i|y_i^T=1,x_i^{FN}) &= \left[F^{FN}(x_i^{FN\prime}\gamma^{FN})\right]^{y_i} + \left[1-F^{FN}(x_i^{FN\prime}\gamma^{FN})\right]^{1-y_i} \\
\Pr(y_i^T|x_i) &= \Phi(x_i'\beta)^{y_i^T} + \left[1-\Phi(x_i'\beta)\right]^{1-y_i^T}
\end{aligned} \tag{22}$$

The first equation gives the model for false positives, which depend on covariates X^{FP} through the parameters γ^{FP} and the link function F^{FP}. Similarly, the second equation gives the model for false negatives and the third equation the model for the true outcome of interest, which depends on the parameters of interest, β. In the application below, we assume that the misreporting models are Probit models, i.e. F^{FP} and F^{FN} are standard normal cumulative distribution functions. This assumption yields a fully specified parametric system of equations that can be estimated by maximum likelihood. The likelihood function is given in appendix C and depends on three components. Which components an observation contributes to depends on whether it contains $\left[y_i, y_i^T, x_i^{FP}, x_i^{FN}\right]$, $\left[y_i, x_i, x_i^{FP}, x_i^{FN}\right]$ or both. The set of observations with the first variables identifies the misreporting models, while the set of observations with the second group of variables identifies the outcome equation in the predicted probabilities estimator. The intersection of the two sets of observations identifies both the misreporting model and the outcome model, so in principle it could be used to estimate the outcome model directly. One may still want to estimate the full model, either because one is interested in the misreporting model or because one considers the observations in the intersection to be insufficient to estimate the parameters of interest (e.g. for reasons of efficiency or sample selection). Such cases often arise if a small subset of the observations has been validated, so that the union of the two sets is much larger than the intersection. In this case the validated observations allow estimation of the true outcome model and the misreporting model while those that were not validated only identify the observed outcome model. We examine an estimator for this setting that we refer to as the

joint estimator with common observations. In other cases, like those discussed by Bollinger and David (1997, 2001), observations that identify the true outcome model by themselves are not available, so we also consider an estimator in which the intersection of the two sets of observations is empty: Some observations identify the misreporting model and others the observed outcome model, but none identify both. Consequently, this estimator has to rely on less information than the previous estimator. We refer to it as the joint estimator without common observations.

4.2 Performance of Estimators when Misreporting is Conditionally Independent

This section and the following one discus the performance of the estimators that we introduced above. What is unique about our analysis is that we use the actual data for a common use of binary choice models, and, more importantly, we know the true value of the dependent variable from administrative data. We apply the estimators to the matched survey data used in part 3 and we begin by examining their performance if misreporting is conditionally random by conducting Monte Carlo simulations. In these simulations, we induce false positives and false negatives at the rate actually observed in the real data, but misreporting is unrelated to X conditional on y_i^T. We run 500 replications of this MC exercise and record the average bias and MSE. For the Predicted Probabilities estimator, we obtain estimates of the probabilities of misreporting (the first stage) from the same sample that we use for the outcome model (the second stage). The joint estimators allow estimation with two different samples with different information, so we split the sample in half randomly and use the two halves for the two samples required for the joint estimators. This choice implies that comparing the joint estimators to other estimators in terms of efficiency is not straightforward because the samples used are different. With the exception of the estimator that fixes α_0 at 0, all estimators are consistent in this setting if the Probit assumption holds in the matched data.

The results in table 7 show that the HAS-Probit greatly improves upon the uncorrected Probit estimator (in columns 2 and 6) if one has estimates of α_0 and α_1 from other sources that one can employ. Columns 4 and 8 use the "true" probabilities of misreporting which results in estimates that have little bias (with the exception of the imprecise Maryland coefficient in the ACS). However, such estimates may not be available, so it is useful to know whether using biased estimates of the probabilities of misreporting still improves the results. Inaccurate estimates may be available from other data sources or time periods. For reports of program receipt, a particularly interesting case is to use the net underreporting rate for α_1 and constrain α_0 to 0, which means that the estimates of α_0 and α_1 are biased, but not too far off. The results in columns 3 and 7 show that biased estimates of the probabilities of misreporting affect the slope coefficients. Choosing α_0 and α_1 to be lower than the true probabilities leads to a partial correction: the corrected estimates are between the estimates from the survey and the matched data. In our case, the corrected estimates still substantially improve the estimates from survey data, but it is unclear in how far this still holds when the estimates of α_0 and α_1 are further off.

Table 7: Comparing Estimators - Conditional Random Misreporting

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ACS				**CPS**			
	Coef.	Percent Bias			Coef.	Percent Bias		
	Matched Data	Standard Probit	$\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .19$	$\tilde{\alpha}_0 = .024$ $\tilde{\alpha}_1 = .26$	Matched Data	Standard Probit	$\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .29$	$\tilde{\alpha}_0 = .033$ $\tilde{\alpha}_1 = .39$
Poverty index	-0.0060	-20.51%	-11.12%	1.34%	-0.0074	-32.47%	-18.43%	3.30%
Age≥50	-0.4062	-22.15%	-14.45%	-2.30%	-0.4297	-32.17%	-20.20%	0.00%
Maryland	-0.0187	-9.08%	15.22%	13.20%	-0.3338	-33.15%	-24.90%	5.84%
Constant	0.0987	-321.02%	-87.58%	10.63%	0.4108	-147.65%	-47.69%	7.22%

Note: Sample sizes: 5945 (ACS), 2791 (CPS) matched households, based on 500 replications. See notes Table 1 and 2 for further details on the samples and MC design. The conditional probabilities of misreporting in column 4 and 8 are based on the actual probabilities, column 3 and 7 use the (expected) net under count as the probability of false negatives.

While the results in table 7 underline that the HAS-Probit can yield substantial improvements if misclassification is conditionally random and one constrains the probabilities

of misclassification based on outside information, we find that the HAS-Probit performs poorly when these probabilities are left unconstrained. In the MC study using the linked data, we faced problems of convergence[7] in many replications and in about 1 percent of the cases convergence seemed to be achieved, but the results were very far from the true parameters (by several hundred percent of the true coefficient).[8] However, even excluding replications where convergence may be an issue does not improve over the naive estimators and leads to large bias in α_0 and α_1. Sufficiently "trimming" the replications by the absolute value of the coefficients can lead to smaller bias than an uncorrected Probit, but MSE remains higher and such a procedure is only feasible in MC studies, but not in practice. The identification of $\hat{\alpha}_0$ and $\hat{\alpha}_1$ in the HAS-Probit comes from the functional form assumptions on the underlying binary choice model, so the problems are likely due to a violation of these assumptions.

Due to these problems, we do not report the results from the linked data, but we examined the problem further using simulated data and the ACS public use files.[9] In summary, we find evidence that misspecification of the functional form of ε (skewness and heavy/light tails) leads to bias, but does not replicate the pattern we find in the real data. While the bias is larger than the bias when estimating a Probit on the data without misclassification, the bias is negligible compared to the uncorrected Probit when there is misclassification even for substantial departures from normality. This indicates that the HAS-Probit is not sensitive to

[7]The likelihood is only globally concave in the parameters under certain conditions that may not hold in practice. See Hausman and Scott-Morton (1994) for the exact conditions.

[8]While such problems may be due to our implementation of the estimator, we have extensively examined the issue using simulated data and the ACS public use files. The results, which are available upon request, indicate that the problems are not due to our implementation. We find similar problems when we use simulated data where we can start the algorithm at the true values and when using a variety of optimization routines. Our simulations also indicate that the difficulties with convergence stem from the inclusion of dummy variables, but the difficulties of convergence are only a computational problem and not associated with failures to converge. The algorithm reliably indicates convergence failures and rarely fails to converge in simulated data. We do not find the rare cases of extreme bias in the simulated data, but they persist in MC studies using the public use data when convergence is carefully monitored. A potential explanation for this is that they arise from replications in which the likelihood function is not concave, but they may also be rare cases of unnoticed convergence failures. In practice, they should be easy to recognize and avoid.

[9]All results are available upon request. We used the ACS public use files instead of the linked data because they are sufficiently similar and we no longer have access to the confidential linked data.

misspecification of the functional form of the error term, which suggests that it is unlikely that semiparametric extensions would solve our problem. However, we find that if one leaves the probabilities of misclassification unconstrained, the estimator is sensitive to violations of the assumption that ε is independent of the covariates. In simulations with an omitted variable that is correlated with the covariates, the HAS-Probit suffers from large bias[10] and a similar pattern as we observe in the real data emerges: the estimates of α_0 and α_1 are substantially biased, but relatively stable across replications while the estimates of the coefficients are both biased and variable. As in the real data, both bias and MSE are small when fixing α_0 and α_1 at their expected values. These results reinforce our belief that the main problem with the HAS-Probit is its dependence on the functional form assumption to identify $\hat{\alpha}_0$ and $\hat{\alpha}_1$. If the probabilities of misreporting are poorly estimated, the estimates of the slope coefficients can be severely biased, due to the fact that misreporting is not completely corrected (or has been overcompensated). Knowing these probabilities from outside information fixes this fragility, and greatly improves the performance of the estimator even if the estimates of α_0 and α_1 are biased as in the MC study using the linked data.

Results from the estimators that remain consistent under correlated misreporting in the same MC setup are as expected, so we do not present them. The Predicted Probabilities estimator has a high MSE, which suggests that one should avoid variables that mainly introduce noise in the first stage. The two joint estimators do well, particularly the joint estimator with common observations, but a small bias remains which may be due to the violation of the functional form assumptions.

In conclusion, these results suggest that if misreporting is conditionally random, estimators that are able to account for misreporting can greatly improve the estimates. However, unless one has great faith that the underlying binary choice model is correctly specified, it is useful to have external estimates of the probabilities of misreporting. Otherwise, since the error rates are hard to estimate, their bias and imprecision can lead to bias in the coefficients

[10]Note that "bias" in this case indicates deviations from the misspecified Probit without misclassification.

and convergence problems.

4.3 Performance of Estimators when Misreporting is Not Conditionally Independent

That it is feasible to estimate the parameters of the true model if misreporting is conditionally random makes it attractive to assume that this is the case. This assumption clearly fails in our data, so one should be concerned with the performance of the estimators when this assumption is violated. Our matched data contains both the "true" dependent variable y^T as well as the misreported indicator y. This enables us to compare estimates of the true coefficients from using the administrative dependent variable to estimates of the biased coefficients from using the survey reports that suffer from misclassification.

Table 8: Comparing Estimators - Correlated Misreporting

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			ACS					**CPS**		
	Matched Data	Standard Probit	HAS Probit	$\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .19$	$\tilde{\alpha}_0 = .024$ $\tilde{\alpha}_1 = .26$	Matched Data	Standard Probit	HAS Probit	$\tilde{\alpha}_0 = 0$ $\tilde{\alpha}_1 = .29$	$\tilde{\alpha}_0 = .033$ $\tilde{\alpha}_1 = .39$
Poverty index	-0.0060	-0.0071	-0.0188	-0.0082	-0.0094	-0.0074	-0.0083	-0.0082	-0.0107	-0.0146
	(0.0004)	(0.0005)	(0.0019)	(0.0006)	(0.0007)	(0.0005)	(0.0006)	(0.0006)	(0.0008)	(0.0015)
Age≥50	-0.4062	-0.4167	-0.8267	-0.4622	-0.5287	-0.4297	-0.3992	-0.3879	-0.4739	-0.6270
	(0.0512)	(0.0543)	(0.1097)	(0.0599)	(0.0713)	(0.0614)	(0.0682)	(0.0679)	(0.0812)	(0.1170)
Maryland	-0.0187	-0.0978	-0.3187	-0.1140	-0.1238	-0.3338	-0.3189	-0.3058	-0.3604	-0.5760
	(0.0548)	(0.0582)	(0.1137)	(0.0640)	(0.0757)	(0.0736)	(0.0808)	(0.0805)	(0.0977)	(0.1655)
$\hat{\alpha}_0$			0.0000					0.0011		
			(0.0000)					(0.0000)		
$\hat{\alpha}_1$			0.6114					0.0000		
			(0.0233)					(0.0000)		

Note: Sample sizes: 5945 (ACS), 2791 (CPS), SEs in parentheses. All analyses conducted using household weights adjusted for PIK probability. The conditional probabilities of misreporting in column 5 and 10 are based on the actual probabilities, column 4 and 9 use the (expected) net undercount as the false negative probability.

We find that all estimators that are consistent only under conditionally random misreporting fare poorly if misreporting is related to the covariates. Table 8 shows that in both surveys, all estimators suffer from larger bias than the naive estimators. One may have expected this given the finding that the biases partly cancel for the survey estimates. Therefore, one should be cautious with the assumption that misclassification is conditionally

independent of the covariates.

Table 9: Comparing Estimators - Correlated Misreporting, ACS

	Survey Data	Matched Data	Pred. Prob.	JE 1: no common observations	JE 2: common observations
Poverty index	-0.0071	-0.0060	-0.0059	-0.0076	-0.0063
	(0.0005)	(0.0004)	(0.0006)	(0.0013)	(0.0005)
Age≥50	-0.4167	-0.4062	-0.3950	-0.5086	-0.3660
	(0.0543)	(0.0512)	(0.0530)	(0.1229)	(0.0615)
Maryland	-0.0978	-0.0187	0.0050	0.1322	-0.0366
	(0.0582)	(0.0548)	(0.0574)	(0.1387)	(0.0662)
Constant	0.0686	0.0987	0.1199	0.3088	0.1568
	(0.0605)	(0.0583)	(0.0721)	(0.1489)	(0.0701)
Weighted Distance	15.672	0	0.694		0.779
Precision	230.824	272.011	220.563	66.650	209.915

Note: Sample size 5945 matched households. The first stage model for (3)-(5) includes age≥50, a MD dummy, the poverty index and its square. The model for false negatives also includes a cubic term in poverty. SEs for (3) are bootstrapped to account for the estimated first stage parameters. All analyses conducted using household weights adjusted for match probability. A mistake prevented the distance statistic for column (4) from being disclosed.

Tables 9 and 10 present the results from the estimators that are consistent if misreporting is related to the covariates: The predicted probabilities estimator from Bollinger and David (1997) and the two joint estimators. The last two rows contain two measures to evaluate their overall performance compared to the estimates from the survey only and matched data. The row labeled "Weighted Distance" gives the average distance to the coefficients from the matched data weighted by the inverse of the variance matrix of the estimates from the matched data. We only use the variance matrix from the matched data in order to use a metric of the difference in point estimates that does not depend on the differences in the efficiency of the estimators. The number in the last row is the F-Statistic of the coefficients from the matched data using the variance matrix of the estimator in that column. This can be interpreted as a measure of efficiency with higher values being better. We use the coefficient from the matched data rather than the estimates in each column in order to avoid

confounding efficiency with estimates that are larger in absolute value.[11] The values from the joint estimators are not directly comparable to the other estimators, since the sample definitions differ.

Table 10: Comparing Estimators - Correlated Misreporting, CPS

	(1) Survey Data	(2) Matched Data	(3) Pred. Prob.	(4) JE 1: no common observations	(5) JE 2: common observations
Poverty index	-0.0083	-0.0074	-0.0070	-0.0044	-0.0070
	(0.0006)	(0.0005)	(0.0007)	(0.0028)	(0.0008)
Age≥50	-0.3992	-0.4297	-0.4109	-0.3368	-0.3693
	(0.0682)	(0.0614)	(0.0616)	(0.1775)	(0.0777)
Maryland	-0.3189	-0.3338	-0.3733	-0.4076	-0.3347
	(0.0808)	(0.0736)	(0.0819)	(0.1958)	(0.0931)
Constant	0.1892	0.4108	0.3688	0.1741	0.3454
	(0.0735)	(0.0723)	(0.0836)	(0.2462)	(0.0998)
Weighted Distance	27.197	0	0.318		0.481
Precision	148.198	189.154	153.357	37.682	131.888

Note: Sample size 2791 matched households. The first stage model for (3)-(5) includes age≥50, a MD dummy and the poverty index. SEs for (3) are bootstrapped to account for the estimated first stage parameters. All analyses conducted using household weights adjusted for match probability. A mistake prevented the distance statistic for column (4) from being disclosed.

The results show that all three estimators work well. The joint estimator without common observations is less efficient than the joint estimator with common observations, but we cannot reject the hypothesis that it is unbiased. Its lack of precision suggests that it is only an attractive option in large datasets if the joint estimator with common observations is not feasible. Both the predicted probabilities estimator and the joint estimator with common observations work extremely well. The predicted probabilities estimator works a little better in our applications, but at least in terms of efficiency, we have stacked the deck in its favor given that we had to split the sample for the joint estimator. One would expect the joint estimator to be more efficient when the same data are used, as it is the maximum likelihood

[11] As any summary measure, these two statistics measure a particular aspect of the performance and may not capture other aspects well. For example, the first statistic is not a test of equality, but is the χ^2_4 statistic of a test that the coefficients from the matched data are equal to the values in a given column. The test statistic from a test of equality may rank the estimators differently.

estimator.[12] The main drawback of the joint estimator with common observations is that it requires observations that identify the entire outcome model. Such observations are rarely available and when they are available a good case can often be made for using *only* those observations in a regular Probit model. On the other hand, the predicted probabilities estimator only requires a consistent estimate of the parameters of the misreporting model, which can often be obtained from other studies, as in Bollinger and David (1997) and does not require the linked data to be available.

An important concern for these estimators besides bias and efficiency is their robustness to misspecification. One will usually not be able to assess whether one has actually reduced bias by using a correction for misreporting since, unlike in this study, validation data are not generally available. The results from applying estimators that are only consistent if misreporting is conditionally random indicate that when this assumption fails increased bias is certainly possible. Informal evidence from using subsamples (such as one of the two states) to identify the misreporting model suggest that neither of the estimators is particularly sensitive to minor misspecification, but the joint estimator with common observations is more robust than the predicted probabilities estimator. In the MC study where misreporting is conditionally random, both joint estimators produced estimates that were less biased than the uncorrected Probit estimates in both surveys for all coefficients except for the insignificant and imprecisely estimated Maryland dummy in the ACS. The predicted probabilities estimator, on the other hand, produced estimates that were more biased than the survey estimates in this setting and fared worse in terms of mean squared error than the joint estimator with common observations for all coefficients. This suggests that the predicted probabilities estimator is sensitive to the inclusion of irrelevant variables in the first stage. One can often avoid such misspecification by doing rigorous specification tests on the misreporting model, which consequently is more important when using the predicted probabilities estimator.

[12]The results for the joint estimator with common observations support this as the SEs are only slightly larger than those of the predicted probabilities estimator despite the fact that the joint estimator only uses half of the sample to estimate the outcome model. The SEs from the joint estimator without common observations are surprisingly large compared to the predicted probabilities estimator.

The importance of sensitivity to misspecification of the misreporting model becomes very concrete when considering validation studies: The results above show that the information obtained from validation studies can improve survey based estimates a lot, but validation studies are costly. This raises the question how much one loses from correcting estimates based on validation data from previous years, a subset of the population or even a different survey. Nothing is lost if the misreporting models are the same in the two datasets, but if they are slightly different the loss depends on the robustness to misspecification of the misreporting model. We examine this issue by estimating the misreporting model on the IL sample of the ACS and using it to see how well it corrects food stamp take up in MD. The misreporting models are statistically different in the two states, but qualitatively similar, so one may be tempted to use them to correct estimates if validation data are only available for some states, even though estimates will not be consistent. The results are in line with our previous findings: The joint estimator with common observations performs best both in terms of bias and efficiency. The joint estimator without common observations still suffers from a lack of precision. The predicted probabilities estimator does well in terms of bias (as measured by the distance metric defined above) compared to an uncorrected Probit, but contrary to the previous case it does worse than the joint estimator with common observations. The advantage of the joint estimator with common observations seems to increase with the degree of misspecification in the misreporting model. All estimators do better than the naive survey estimates in terms of the distance metric used above, so if similar data have been validated or parameter estimates from similar data are available, using them to correct the survey coefficients may be worth trying.[13]

[13]We also correct estimates of food stamp take-up in the ACS using the misreporting model we observe in the CPS and vice versa to see how extrapolating from a different survey works. The results, which are available upon request, underline that the joint estimator with common observations is more efficient than the other two estimators, but do not provide conclusive evidence on the performance. They seem to improve over the naive estimator, particularly for the coefficients on the continuous variable.

5 Conclusion

We analyze the properties of common binary choice estimators when the dependent variable is subject to misclassification and then evaluate the performance of several estimators designed to take such misclassification into account. In the first part of the paper, we derive analytic results for the bias due to misclassification in the linear probability and the Probit model for the general case where misclassification is allowed to depend on the covariates in arbitrary ways. These results are informative about the likely size and direction of biases in practice. They also help to explain the results of previous studies, for example why Meyer, Goerge and Mittag (2014) find the signs of coefficient estimates to be remarkably robust to misclassification. We provide conditions under which certain features of the true parameters, such as their signs and relative magnitudes, are robust to misclassification. Under these conditions, some inference can still be based on the biased coefficients. We illustrate these results using simulations and models of food stamp take-up using two unique data sets that include "true" food stamp receipt from administrative data besides the survey reports. These applications show that the asymptotic results are useful to interpret the biases found in practice. They underline that the bias due to misclassification of the dependent variable can be substantial, but also confirm that there is a tendency for coefficients to retain their sign. However, the asymptotic results and simulations show that this tendency does not have to hold in general.

Our evaluation of the estimators that take misclassification into account shows that the true parameters can be estimated from noisy data if the misclassification model is well-specified. Which estimator performs best depends on the nature of the misclassification, in particular whether it is related to covariates or not. We find that if misclassification is conditionally random, the HAS Probit works well if one fixes the conditional probabilities of misclassification using estimates from outside information. In our case of program participation, where the probability of false positives is low, even using only the net underreporting rate (which is often easy to obtain) yields improvements over the naive estimator. However,

one should be cautious to assume that misclassification is conditionally random. While the estimators are not very sensitive to small misspecifications, we find that they can easily perform worse than the naive model if key assumptions such as conditional independence of the covariates are erroneously made.

Additional information on the nature of the misclassification process not only helps to avoid invalid assumptions, but can also help to increase the robustness of the estimators by placing additional restrictions on the model. We show that even if misclassification is not conditionally random, it is still possible to obtain consistent estimates. This requires additional information on the misclassification model, such as parameter estimates from which one can predict the probabilities of misclassification. Among the estimators we evaluate that do not require misclassification to be conditionally random, one should use the joint estimator with common observations if it is feasible. If it is not feasible the predicted probabilities estimator is an attractive alternative. As the main downside of the predicted probabilities estimator is its greater sensitivity to the misclassification model, its specification should be tested. In general, our results show the value of validation data to understand the nature of misclassification and its effects.

Appendix A: Proof of Equation 6

Equation (6) gives the expectation of the coefficients in the linear probability model when the conditional probabilities of misreporting are constants as in Hausman, Abrevaya and Scott-Morton (1998), i.e. when

$$\begin{aligned} \Pr(y_i = 1|y_i^T = 0) &= \alpha_{0i} = \alpha_0 \\ \Pr(y_i = 0|y_i^T = 1) &= \alpha_{1i} = \alpha_1 \end{aligned} \qquad \forall i$$

By the assumptions of the linear probability model $\Pr(y_i^T = 1|X) = x_i'\beta^{LPM}$ and $\Pr(y_i^T = 0|X) = 1 - x_i'\beta^{LPM}$, so that the probability mass function of the measurement error, U, conditional on X is

$$\Pr(U = u_i|X = x_i) = \begin{cases} \Pr(y_i = 1|y_i^T = 0) \cdot \Pr(y_i^T = 0|X) = \alpha_0(1 - x_i'\beta^{LPM}) & \text{if } u_i = 1 \\ 1 - \alpha_0 + (\alpha_0 - \alpha_1)x_i'\beta^{LPM} & \text{if } u_i = 0 \\ \Pr(y_i = 0|y_i^T = 1) \cdot \Pr(y_i^T = 1|X) = \alpha_1 x_i'\beta^{LPM} & \text{if } u_i = -1 \end{cases}$$

where the probability for $u_i = 0$ follows immediately from the fact that the probabilities have to sum to 1. Consequently, the conditional expectation of the measurement error is

$$\begin{aligned} \mathbb{E}(u_i|X = x_i) &= 1 \cdot [\alpha_0(1 - x_i'\beta^{LPM})] + 0 \cdot [1 - \alpha_0 + (\alpha_0 - \alpha_1)x_i'\beta^{LPM}] - 1 \cdot [\alpha_1 x_i'\beta^{LPM}] \\ &= \alpha_0 - (\alpha_0 + \alpha_1)x_i'\beta^{LPM} \end{aligned}$$

Assuming that X is non-stochastic[14], this implies that the bias, $\mathbb{E}(\hat{\delta})$, is

$$\begin{aligned} \mathbb{E}(\hat{\delta}) &= \mathbb{E}(X'X)^{-1}(X'u) \\ &= (X'X)^{-1}\mathbb{E}(X'u) \\ &= (X'X)^{-1}\sum_i x_i'\mathbb{E}(u_i|x_i) \end{aligned}$$

[14]The extension to stochastic X follows from the law of iterated expectation.

$$
\begin{aligned}
&= (X'X)^{-1} \sum_i x_i'(\alpha_0 - (\alpha_0 + \alpha_1) x_i' \beta^{LPM}) \\
&= (X'X)^{-1} \sum_i x_i' \alpha_0 - (\alpha_0 + \alpha_1)(X'X)^{-1} \sum_i x_i x_i' \beta^{LPM} \\
&= (X'X)^{-1} X' \alpha_0 - (\alpha_0 + \alpha_1)(X'X)^{-1}(X'X)\beta^{LPM} \\
&= (\alpha_0, 0, \ldots, 0)' - (\alpha_0 + \alpha_1)\beta^{LPM}
\end{aligned}
$$

The first term is the coefficient vector from a regression of a constant, α_0, on X, so the intercept will be equal to α_0 and all other coefficients will be zero. The expectation of the biased coefficient vector is $\mathbb{E}(\hat{\tilde{\beta}}^{LPM}) = \beta^{LPM} + \mathbb{E}(\hat{\delta})$. Consequently, the expectation of the (biased) intercept is

$$\mathbb{E}(\hat{\tilde{\beta}}_0^{LPM}) = \alpha_0 + (1 - \alpha_0 - \alpha_1)\beta_0^{LPM}$$

and the expectation of the (biased) slope parameters is

$$\mathbb{E}(\hat{\tilde{\beta}}_{1\ldots k}^{LPM}) = (1 - \alpha_0 - \alpha_1)\beta_{1\ldots k}^{LPM}$$

Appendix B: Further Analysis of the Higher Order Bias

We have shown that the bias in the Probit model depends on four components: two components due to the linear projection, a rescaling bias and bias due to misspecification of the higher order moments of the distribution function. The bias due to misspecification of the error distribution is harder to assess than the other components. If one has enough information about $F_{\tilde{\varepsilon}|X}$ to take random draws from it, one can simulate the exact bias or even obtain an exact solution of (16). In practice, however, such detailed information will rarely be available, so we discuss some factors that influence the size and direction of the bias. They may allow the researcher to informally assess whether this bias is likely to be small, which justifies considering $\bar{\beta} - \beta$ a good approximation to the full bias.

Given that the left hand side of equation (16) is the derivative of a Probit likelihood, which is globally concave in b, the left hand side of each equation in (16) crosses 0 only once and does so from above. In the absence of bias due to functional form misspecification, it does so at $b = \bar{\beta}$. In the univariate case, this implies that if the left hand side of (16) is positive at this point, the additional bias will be positive, while the additional bias will be negative if the left hand side is negative at $b = \bar{\beta}$. In the multivariate case, this can in principle be offset for some, but not all coefficients by the fact that the bias "spreads" between coefficients. For the multivariate case, note that

$$f_X(x_i)\frac{\phi(x_i'b)}{\Phi(x_i'b)(1-\Phi(x_i'b))} > 0 \tag{23}$$

so (16) can be interpreted as a weighted average of $x_i' \left[F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta})\right]$ with the weights given by (23). Consequently, observations for which $sign(x_i) = sign(F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta}))$ tend to cause a positive bias in the coefficient on x, while observations with opposing signs tend to cause a negative bias. The weight function has a minimum at 0 and increases in either direction, so differences at more extreme values of $x'b$ have a larger impact. Larger values of x also tend to make $x_i' \left[F_{\tilde{\varepsilon}|X=x_i}(x_i'\bar{\beta}) - \Phi(x_i'\bar{\beta})\right]$ larger, because x enters it multi-

plicatively. The expression at each value of x is weighted by its density $f_X(x)$, so differences at frequent values of x have a larger impact.

Consequently, one can get an idea of the direction of the bias if one knows how $F_{\tilde{\varepsilon}_i}$ and Φ differ. If the former is larger in regions where the sample density of x is high, $|x'b|$ is high or $|x|$ is large, the bias will tend to be positive if x is positive in this region and negative if x is negative in this region.

Appendix C: Likelihood of the Joint Estimators

As discussed in section 4.1, the sample used for estimation can be divided into three disjoint subsamples. The first subsample, S_1 contains observations that are in the validation data, but not in the data used to identify the outcome model and thus contains $(y_i, y_i^T, x_i^{FP}, x_i^{FN})$. This sample identifies the parameters of the misreporting equations, but not the outcome equation. The second subsample, S_2, contains all observations that have been validated and include all variables in the outcome model, so that it contains $(y_i, y_i^T, x_i, x_i^{FP}, x_i^{FN})$. This sample identifies all parameters. The third subsample, S_3, contains the observations used to estimate the outcome model that have not been validated. Thus, it contains $(y_i, x_i, x_i^{FP}, x_i^{FN})$, which by itself identifies none of the parameters. Frequently, one of the first two samples will be empty. For the joint estimator with common observations, S_1 is empty; For the joint estimator with common observations, S_2 is empty as in Bollinger and David (1997).

Since the subsamples are disjoint and independent, the log-likelihood of the entire sample is the sum of the three log-likelihoods of the subsamples:

$$\begin{aligned}\ell(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = &\sum_{i \epsilon S_1} \ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \\ &\sum_{i \epsilon S_2} \ell_i^{S_2}(y_i, y_i^T, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) + \\ &\sum_{i \epsilon S_3} \ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN})\end{aligned} \tag{24}$$

Equation (22) and the assumption that the error terms in all three equations are independent draws from standard normal distributions imply the following (conditional) probabilities

$$\begin{aligned}\Pr(y_i | y_i^T = 0, x_i^{FP}) &= \left[\Phi(x_i^{FP\prime}\gamma^{FP})\right]^{y_i} + \left[1 - \Phi(x_i^{FP\prime}\gamma^{FP})\right]^{1-y_i} \\ \Pr(y_i | y_i^T = 1, x_i^{FN}) &= \left[\Phi(x_i^{FN\prime}\gamma^{FN})\right]^{y_i} + \left[1 - \Phi(x_i^{FN\prime}\gamma^{FN})\right]^{1-y_i} \\ \Pr(y_i^T | x_i) &= \Phi(x_i'\beta)^{y_i^T} + \left[1 - \Phi(x_i'\beta)\right]^{1-y_i^T}\end{aligned} \tag{25}$$

The first probability is the likelihood contribution of an observation in S_1 with $y_i^T = 0$, while the second probability is the liklihood contribution of an observation in S_1 with $y_i^T = 1$. Consequently, the likelihood contribution of an observation from S_1 is

$$\ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) = (1 - y_i^T)[y_i \ln \Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - y_i) \ln \Phi(-x_i^{FP\prime}\gamma^{FP})] + y_i^T[(1 - y_i) \ln \Phi(x_i^{FN\prime}\gamma^{FN}) + y_i \ln \Phi(-x_i^{FN\prime}\gamma^{FN})]$$

This is the sum of the likelihoods of probit models for false positives and false negatives.

The likelihood contribution of an observation in sample S_2 is the probability of the observed combination of y_i and y_i^T, which is

$$\Pr(y_i, y_i^T | x_i, x_i^{FP}, x_i^{FN}) = \begin{cases} \Pr(y_i | y_i^T = 0, x_i^{FP}) \Pr(y_i^T = 0 | x_i) & \text{if } y_i^T = 0 \\ \Pr(y_i | y_i^T = 1, x_i^{FN}) \Pr(y_i^T = 1 | x_i) & \text{if } y_i^T = 1 \end{cases}$$

Using the probabilities from (25) yields the likelihood contribution of an observation from sample S_2:

$$\ell_i^{S_2}(y_i, y_i^T, x, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = (1 - y_i^T) \ln \left([y_i \Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - y_i)\Phi(-x_i^{FP\prime}\gamma^{FP})]\Phi(-x_i'\beta) \right) + y_i^T \ln \left([(1 - y_i)\Phi(x_i^{FN\prime}\gamma^{FN}) + y_i\Phi(-x_i^{FN\prime}\gamma^{FN})]\Phi(x_i'\beta) \right)$$

Equation (25) defines the probabilities of false positives as $\alpha_{0i} = \Phi(x_i^{FP\prime}\gamma^{FP})$ and of false negatives as $\alpha_{1i} = \Phi(x_i^{FN\prime}\gamma^{FN})$. Using these probabilities of misreporting in equation (17) yields the contribution of an observation from S_3:

$$\ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = y_i \ln[\Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] + (1 - y_i) \ln[1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)]$$

Using the three likelihood contributions $\ell_i^{S_1}, \ell_i^{S_2}$ and $\ell_i^{S_3}$ in equation (24) and maximizing

it with respect to $(\beta, \gamma^{FP}, \gamma^{FN})$ yields consistent estimates of the three parameter vectors by the standard arguments for the consistency of maximum likelihood. Standard errors of all parameters can be obtained as usual. The joint estimator with common observations used above assumes that S_1 is empty, so the log-likelihood reduces to

$$\begin{aligned}
&\ell^{JE1}(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = \\
&\quad \sum_{i \epsilon S_2} \big[(1-y_i^T) \ln \big([y_i \Phi(x_i^{FP\prime}\gamma^{FP}) + (1-y_i)\Phi(-x_i^{FP\prime}\gamma^{FP})]\Phi(-x_i'\beta) \big) + \\
&\qquad\quad y_i^T \ln \big([(1-y_i)\Phi(x_i^{FN\prime}\gamma^{FN}) + y_i \Phi(-x_i^{FN\prime}\gamma^{FN})]\Phi(x_i'\beta) \big) \big] + \\
&\quad \sum_{i \epsilon S_3} \big[y_i \ln[\Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] + \\
&\qquad\quad (1-y_i) \ln[1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] \big]
\end{aligned}$$

The second joint estimator we use above, the joint estimator without common observations, assumes that S_2 is empty, so the log-likelihood reduces to

$$\begin{aligned}
&\ell^{JE2}(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = \\
&\quad \sum_{i \epsilon S_1} \big[(1-y_i^T)[y_i \ln \Phi(x_i^{FP\prime}\gamma^{FP}) + (1-y_i) \ln \Phi(-x_i^{FP\prime}\gamma^{FP})] + \\
&\qquad\quad y_i^T [(1-y_i) \ln \Phi(x_i^{FN\prime}\gamma^{FN}) + y_i \ln \Phi(-x_i^{FN\prime}\gamma^{FN})] \big] + \\
&\quad \sum_{i \epsilon S_3} \big[y_i \ln[\Phi(x_i^{FP\prime}\gamma^{FP}) + (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] + \\
&\qquad\quad (1-y_i) \ln[1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - (1 - \Phi(x_i^{FP\prime}\gamma^{FP}) - \Phi(x_i^{FN\prime}\gamma^{FN}))\Phi(x_i'\beta)] \big]
\end{aligned}$$

Note that the likelihood contribution of S_2 can be re-written as the sum of the likelihood contribution if the observation were in sample S_1 and the likelihood contribution to a standard probit likelihood:

$$\begin{aligned}
\ell_i^{S_2}(y_i, y_i^T, x, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = &\ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \\
&y_i^T \ln \Phi(x_i'\beta) + (1-y_i^T) \ln \Phi(-x_i'\beta)
\end{aligned}$$

This can be used to re-write the log-likelihood function as

$$\ell(y, y^T, x, x^{FP}, x^{FN}; \beta, \gamma^{FP}, \gamma^{FN}) = \sum_{i \epsilon S_1 \cup S_2} \ell_i^{S_1}(y_i, y_i^T, x_i^{FP}, x_i^{FN}; \gamma^{FP}, \gamma^{FN}) + \sum_{i \epsilon S_2} \ell_i^P(y_i^T, x_i; \beta) + \sum_{i \epsilon S_3} \ell_i^{S_3}(y_i, x_i, x_i^{FP}, x_i^{FN}; \beta, \gamma^{FP}, \gamma^{FN})$$

where ℓ^P is the log-likelihood function of a standard probit model. This shows more clearly which observations contribute to the identification of the parameters. In particular, it shows the value of observations in S_2, because in addition to the likelihood contribution of an observation in S_1, they add a term that directly identifies the parameters of the outcome model, β. It also shows that observations in S_3 contribute to the identification of the parameters of false positives and false negatives even though these observations only contain information on the observed dependent variable.

References

Aigner, Dennis J. 1973. "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics*, 1(1): 49–59.

Bitler, Marianne P., Janet Currie, and John K. Scholz. 2003. "WIC Eligibility and Participation." *The Journal of Human Resources*, 38: 1139–1179.

Black, Dan A., Seth Sanders, and Lowell Taylor. 2003. "Measurement of Higher Education in the Census and Current Population Survey." *Journal of the American Statistical Association*, 98: 545–554.

Bollinger, Christopher R. 1996. "Bounding mean regressions when a binary regressor is mismeasured." *Journal of Econometrics*, 73(2): 387–399.

Bollinger, Christopher R., and Martin H. David. 1997. "Modeling Discrete Choice With Response Error: Food Stamp Participation." *Journal of the American Statistical Association*, 92(439): 827–835.

Bollinger, Christopher R., and Martin H. David. 2001. "Estimation with Response Error and Nonresponse: Food-Stamp Participation in the SIPP." *Journal of Business & Economic Statistics*, 19(2): 129–41.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement error in survey data." In *Handbook of Econometrics*. Vol. 5, , ed. James J. Heckman and Edward Leamer, Chapter 59, 3705–3843. Amsterdam:Elsevier.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers. 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data." *Journal of Labor Economics*, 12(3): 345–368.

Call, Kathleen T., Gestur Davidson, Michael Davern, and Rebecca Nyman. 2008. "Medicaid undercount and bias to estimates of uninsurance: new estimates and existing evidence." *Health Services Research*, 43(3): 901–914.

Cameron, Stephen V., and James J. Heckman. 2001. "The Dynamics of Educational Attainment for Black, Hispanic, and White Males." *Journal of Political Economy*, 109(3): 455–499.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective. Monographs on Statistics and Applied Probability.* 2nd ed., Boca Raton:Chapman & Hall/CRC.

Chen, Xiaohong, Han Hong, and Denis Nekipelov. 2011. "Nonlinear Models of Measurement Errors." *The Journal of Economic Literature*, 49(4): 901–37.

Collier, Paul, and Anke Hoeffler. 1998. "On Economic Causes of Civil War." *Oxford Economic Papers*, 50(4): 563–73.

Davern, Michael, Jacob A. Klerman, David K. Baugh, Kathleen T. Call, and George D. Greenberg. 2009*a*. "An examination of the Medicaid undercount in the current population survey: preliminary results from record linking." *Health Services Research*, 44(3): 965–987.

Davern, Michael, Jacob A. Klerman, Jeanette Ziegenfuss, Victoria Lynch, and George Greenberg. 2009*b*. "A partially corrected estimate of medicaid enrollment and uninsurance: Results from an imputational model developed off linked survey and administrative data." *Journal of Economic & Social Measurement*, 34(4): 219–240.

Eckstein, Zvi, and Kenneth I. Wolpin. 1999. "Why Youths Drop out of High School: The Impact of Preferences, Opportunities, and Abilities." *Econometrica*, 67(6): 1295–1339.

Estrella, Arturo, and Frederic S. Mishkin. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics*, 80(1): 45–61.

Fearon, James D., and David D. Laitin. 2003. "Ethnicity, insurgency, and civil war." *American Political Science Review*, 97(01): 75–90.

Haider, Steven J., Alison Jacknowitz, and Robert F. Schoeni. 2003. "Food Stamps and the Elderly: Why Is Participation so Low?" *The Journal of Human Resources*, 38: 1080–1111.

Hausman, Jerry A., and Fiona M. Scott-Morton. 1994. "Misclassification of the dependent variable in a discrete-response setting." Department of Economics, MIT Working Paper 94-19.

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.

Imbens, Guido W., and Tony Lancaster. 1994. "Combining Micro and Macro Data in Microeconometric Models." *The Review of Economic Studies*, 61(4): 655–680.

Levitt, Steven D. 1998. "Juvenile Crime and Punishment." *Journal of Political Economy*, 106(6): 1156–1185.

Lochner, Lance, and Enrico Moretti. 2004. "The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports." *The American Economic Review*, 94(1): 155–189.

Marquis, Kent H., and Jeffrey C. Moore. 1990. "Measurement Errors in SIPP Program Reports." In *Proceedings of the 1990 Annual Research Conference*. 721–745. Washington, D.C.:U.S. Bureau of the Census.

Meyer, Bruce D., Robert Goerge, and Nikolas Mittag. 2014. "Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation." Unpublished Manuscript.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan. 2009. "The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences." Harris School of Public Policy Studies, University of Chicago Working Paper 0903.

Poterba, James M., and Lawrence H. Summers. 1995. "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification." *The Review of Economics and Statistics*, 77(2): 207–16.

Ruud, Paul A. 1983. "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models." *Econometrica*, 51(1): 225–228.

Ruud, Paul A. 1986. "Consistent estimation of limited dependent variable models despite misspecification of distribution." *Journal of Econometrics*, 32(1): 157–187.

Yatchew, Adonis, and Zvi Griliches. 1984. "Specification Error in Probit Models." Institute for Policy Analysis, University of Toronto Working Paper 8429.

Yatchew, Adonis, and Zvi Griliches. 1985. "Specification Error in Probit Models." *The Review of Economics and Statistics*, 67(1): 134–139.