Corrupt Binary Confounders*

Pierre Nguimkeu§

Georgia State University

August 2026

Abstract

Researchers facing mismeasured confounders must often choose between omitting the variable (risking omitted-variable bias) or including a noisy proxy (introducing measurement-error bias). While conventional wisdom favors proxy inclusion, recent work on nonclassical measurement error shows that a sufficiently poor proxy can sometimes exacerbate bias relative to omission. We show that this "fear of the proxy" is unfounded when the confounder is binary. Although binary misclassification is inherently nonclassical, including a corrupt binary proxy strictly reduces the bias in the coefficient of interest relative to omitting the confounder entirely. We then present two estimators that explicitly account for classification error in the confounder: a two-step procedure that estimates false-positive and false-negative rates before applying a corrected structural estimator, and a joint GMM estimator that combines the misclassification and the structural moment restrictions. Monte Carlo simulations quantify the bias from proxy inclusion relative to omission and demonstrate the finite-sample performance of the proposed estimators under varying degrees of misclassification and identification strength. An empirical application to household business data from Cameroon illustrates the practical relevance of these methods for estimating returns to education and gender earnings disparities when formality status is misclassified.

Keywords: Misclassification, Measurement Error, Omitted Variable Bias, Binary Confounders, GMM.
JEL Codes: C13, C20, O12.

*I thank Augustine Denteh, Desiré Kédagni, Tom Mroz, Rusty Tchernis and seminar participants at Georgia State University and the Georgia Econometrics Workshop for helpful comments.

§ Department of Economics, Georgia State University, 55 Park Place, Suite 600, Atlanta, GA 30303, USA. Email: nnguimkeu@gsu.edu.

1 Introduction

Estimating the causal effect of an explanatory variable on an outcome is a central goal in applied economics and business research. The validity of the workhorse multiple regression model hinges on the successful inclusion of relevant control variables, or confounders, to isolate the effect of interest and mitigate omitted variable bias (OVB) (Angrist & Pischke 2009). In practice, however, researchers rarely have access to perfectly measured data and must almost always work with corrupt or noisy ones, a challenge documented extensively in Bound, Brown & Mathiowetz (2001). Key confounders such as parental history, health status, informality or program participation, are often misreported in survey data, presenting applied researchers with a persistent dilemma: Should one include the imperfect, "noisy" proxy variable, or omit it entirely and accept the resulting OVB? For the case of classical measurement error (CME), where the error is uncorrelated with the true value and the regression error, the literature provides a relatively clear answer. Early work by Wickens (1972) and McCallum (1972) has formally shown that including a classically mismeasured control variable is generally the "lesser of two evils." It induces a smaller bias on the coefficient of interest than the OVB resulting from omission, leading to the widely accepted heuristic that "poor controls are better than no controls".

However, this wisdom has been challenged by literature focusing on non-classical measurement error (NCME), where the measurement error term is correlated with the true value. Studies such as Frost (1979), Bound et al. (2001), Bollinger (2003), Wolpin (2013), and, more recently De Luca, Magnus & Peracchi (2018) have shown that when measurement error in the confounder is non-classical, including a proxy can increase the bias in the coefficients of interest compared to simply omitting the confounder. In general, the classical measurement error assumption is frequently untenable in practice and there is no a priori reason to believe that it satisfies this condition. This is particularly true when dealing with bounded variables or sensitive self-reported data where errors are inherently tied to a respondent's true state (Bound et al. 2001). In such contexts, the measurement error term is not merely random noise but is endogenously determined, potentially leading to biases that differ significantly from the standard attenuation predicted by classical theory (Meyer & Mittag 2017). This counter-intuitive result has created legitimate anxiety among practitioners, leading to uncertainty about whether to control for variables known to have high levels of non-classical misreporting.

This paper formally addresses this dilemma specifically for the ubiquitous case of binary confounders. Misclassification in binary variables is, by definition, non-classical: if the true value is 1, the error can only be -1 or 0; if the true value is

0, the error can only be 1 or 0. Thus, the error is negatively correlated with the truth and is often correlated with other regressors or the outcome (Aigner 1973). Despite this non-classical nature, we prove a key theoretical result in this paper: in the case of a binary confounder, including a corrupt proxy cannot yield a larger asymptotic bias than omitting the variable entirely, provided the level of corruption is not so severe that the proxy is uninformative. This result restores the "include" wisdom for the binary case and provides a crucial, generalizable guideline that alleviates concerns about bias amplification, and insures a theoretical safety net for researchers. The result stems from the subtle, structural correlations inherent in binary misclassification which do not yield the same pessimistic conclusions as the continuous case that previous authors have discussed.

While our first finding confirms that including a binary proxy as confounder reduces bias in the coefficients of interest, it does not eliminate it. Therefore, our second major contribution is to present two practical, fully consistent estimators that correct the residual bias left by the naive OLS model when misclassification probabilities are unknown to the researcher. We propose a two-step procedure that estimates misclassification probabilities via a first-step maximum likelihood (Hausman, Abrevaya & Scott-Morton 1998) and uses them to correct the structural regression, and an efficient joint GMM Estimator that combines them. We establish the consistency and asymptotic normality of our proposed estimators, which are sufficiently flexible to also incorporate the unusual case in which misclassification probabilities are known to the researcher (e.g., through validation studies). We use Monte Carlo simulations to quantify the finite-sample costs of including a corrupt binary confounder and the performance of the proposed correction estimators, under various degrees of misclassification and identification strength. Finally, we illustrate the practical relevance of our findings with an empirical application examining the returns to education among household business owners in Cameroon.

This paper contributes to two related strands of the econometric literature: the analysis of measurement error in control variables, and the identification of models with misclassified binary regressors. While the literature on the measurement error in control variables has grown substantially (see, e.g., surveys by Griliches 1986, Bound et al. 2001, Chen, Hong & Nekipelov 2011), a critical distinction must be drawn between continuous and binary confounders. Early work by Wickens (1972) and McCallum (1972), and more recently by Pei, Pischke & Schwandt (2019) established that under classical measurement error, including a continuous proxy variable reduces the bias on the coefficient of interest relative to omitting the variable entirely. For continuous variables subject to non-random or differential

measurement error, the potential for getting a bias expansion instead is a genuine threat, as detailed by Frost (1979) and elaborated upon by Bollinger (2003), Wolpin (2013), and De Luca et al. (2018). These works show scenarios where the correlation structure of the continuous error term can destabilize OLS estimation of the coefficient of interest and make the naive estimator worse than simple omission.

The second strand examines the consequences of binary misclassification for identification. Aigner (1973) and Bollinger (1996) provided the seminal analysis of regression with misclassified binary regressors, showing that point identification is lost without further assumptions, often leading to bounding approaches (Card 1996, Kane, Rouse & Staiger 1999, Black, Berger & Scott 2000, Bollinger 2001, 2003, Kreider, Pepper, Gundersen & Jolliffe 2012, Kreider & Pepper 2007, Kreider 2010, van Hasselt & Bollinger 2012, Bollinger & van Hasselt 2017, Bollinger & Minier 2014, Tommasi & Zhang 2024). The modern literature, however, often pursues point identification via auxiliary information. This includes methods using instrumental variables (Frazis & Loewenstein 2003, Mahajan 2006, Lewbel 2007, Hu 2008, Hu & Shennach 2008, Bollinger & Minier 2014, Ura 2018, Yanagi 2019, Kasahara & Shimotsu 2022, Haider & Stephens Jr 2025) or repeated measures (Black et al. 2000, Schennach 2016, DiTraglia & García-Jimeno 2019). Another set of papers dive into potential bias amplification. This includes research by Bound et al. (2001) and Nguimkeu, Denteh & Tchernis (2019) highlighting that misclassification in treatment status can produce biases that sign-flip or exaggerate treatment effects. Wossen, Abdoulaye, Alene, Nguimkeu, Feleke, Rabbi, Haile & Manyong (2019), Meyer, Mittag & Goerge (2022), Haider & Stephens Jr (2025), Lamarche (2025) provide empirical evidence of this phenomenon using experimental, household survey, or administrative data.

Our paper moves beyond these frameworks by focusing on misclassification in a binary confounder. We show that the consequences of NCME in a binary control variable are fundamentally different. The structural relationship between the misclassification probabilities and the true binary values imposes a unique set of constraints on the asymptotic bias formula that prevent the bias-expansion mechanism documented in the continuous case from taking effect in the binary confounder case. This finding refines the understanding of NCME, establishing that the robust inclusion heuristic survives the violation of the classical error assumption specifically because of the discrete nature of the variable. It therefore bridges the gap between the "optimistic" classical measurement error literature and the "cautious" non-classical measurement error literature. We also derive conditions under which specific structural features of the binary error allow for point identification of the main coefficient of interest, specifically the non-linear

curvature and covariate variation in the misclassification process.

The rest of the paper is organized as follows: Section 2 describes the econometric framework and compares the various biases. Section 3 develops correction estimators for misclassified binary confounders. Section 4 presents the results of a Monte Carlo Simulation study. Section 5 illustrates the methods in an empirical application, and Section 6 concludes. Mathematical proofs are given the appendix.

2 The Econometric Framework

We consider the following regression model for observation i in a random sample of size n, given by

$$y_i = \alpha + \beta s_i + \gamma z_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i | s_i, z_i] = 0, \tag{1}$$

where y_i is the outcome variable (such as wage, hours of work, amount of loan), s_i is the explanatory variable of interest (such as years of education), which could potentially also be a vector of explanatory variables of interest, and $z_i \in \{0, 1\}$ is an unobserved binary confounder such that $\gamma \neq 0$ and $\text{Cov}(s_i, z_i) \neq 0$. In practice, there could be potentially many other correctly measured control variables added to the model. To simplify the exposition, we ignore them in our discussion, given that they can be incorporated in s_i without changing our reasoning. The main parameter of interest for the researcher is β, while γ is a nuisance parameter.

We assume that the researcher can only observe y_i, the dependent variable, s_i, the main regressor(s) of interest, and x_i, a corrupted version of the confounder z_i. Data corruption takes the form of a measurement error, so that we can write

$$x_i = z_i + u_i, \tag{2}$$

where u_i is the measurement error. When z_i is binary, u_i is necessarily correlated (negatively) with z_i and is therefore a non-classical measurement error. We assume throughout that the measurement error is exogenous to the outcome equation, that is, $\text{Cov}(\varepsilon_i, u_i) = 0$. We also assume, as is common in a major strand of the econometric literature (e.g., Aigner 1973, Bollinger 1996, Bollinger & David 1997, Kreider & Pepper 2007, Bollinger & van Hasselt 2017, van Hasselt & Bollinger 2012, Black et al. 2000, Kane et al. 1999, Hausman et al. 1998), that

$$\begin{aligned} \Pr[x_i = 1 | s_i, z_i, \varepsilon_i] &= p(1 - z_i) + (1 - q)z_i \\ \Pr[x_i = 0 | s_i, z_i, \varepsilon_i] &= qz_i + (1 - p)(1 - z_i), \end{aligned} \tag{3}$$

with $p + q < 1$. This means p is the probability of false positives (i.e., the probability of observing $x_i = 1$ when $z_i = 0$) and q is the probability of false negatives

(i.e. the probability of observing $x_i = 0$ when $z_i = 1$). Meyer & Mittag (2017) refer to this condition as the misclassification probabilities being "conditionally random". The condition $p + q < 1$ is equivalent to $\text{Cov}(x_i, z_i) > 0$ implying that measurement error did not completely corrupt the confounder x_i so that it is still an informative surrogate for z_i. If $p + q = 1$, this would mean $\text{Cov}(x_i, z_i) = 0$, which implies no signal; hence, the proxy is pure noise and is not of interest. We focus our discussion on "population" relationships so that all the sample statistics are discussed in terms of their asymptotic limits.

Linear regression models of the form of Equation (1) are a workhorse of empirical microeconometrics, where interest centers on identifying the causal effect (β) of a policy or treatment (s_i) while controlling for a key confounder (z_i). In practice, critical confounders, such as individual ability, household social status, or managerial quality, are often unobserved latent states. Applied researchers must then choose between omitting these latent variables entirely or substituting them with observable, discrete proxies (x_i), such as binary survey indicators or administrative classifications. In development economics, studies evaluating returns to infrastructure or microfinance (s_i) must control for latent household wealth or social capital (z_i), typically measured via noisy, self-reported survey proxies (x_i) subject to misreporting (e.g., Deaton 2019, Alderman, Behrman, Lavy & Menon 2001). Similarly, labor economists estimating wage returns to schooling (s_i) rely on binary indicators of worker skill or training (z_i) that suffer from reporting error (e.g., Kane & Rouse 1995, Kane et al. 1999, Card 1996, Black, Sanders & Taylor 2003, Black et al. 2000). However, a prominent strand of the measurement error literature warns that non-classical measurement error structures risk inducing expansion bias, making proxy inclusion worse than complete omission (e.g., Frost 1979, Bound et al. 2001, Bollinger 2003, De Luca et al. 2018). We first evaluate this cautionary narrative for a misclassified binary surrogate, assessing whether proxy inclusion exacerbates distortion or systematically dominates omission, before turning to full bias-correction estimators.

If the researcher decides to drop the confounder z_i completely from the model, then they estimate an omitted variable model of the form:

$$y_i = \alpha^o + \beta^o s_i + \varepsilon_i^o, \tag{4}$$

where $\varepsilon_i^o = \varepsilon_i + \gamma z_i$. The OLS estimation of β using Equation (4) would then lead to an omitted variable bias. But if the researcher has a candidate proxy x_i for z_i and decides to include it, then they would run an operational (i.e., a naive or proxy) model using the corrupt binary confounder x_i, given by

$$y_i^c = \alpha^c + \beta^c s_i + \gamma^c x_i + \varepsilon_i^c, \tag{5}$$

where $\varepsilon_i^c = \varepsilon_i - \gamma u_i$, and the estimation of β using Equation (5) would lead to a corrupt proxy bias. In the next section, we discuss the coefficient β obtained from estimating these two models, and we compare the biases of including the corrupt confounder versus omitting it entirely. To simplify the exposition of this comparison, we first treat s_i as a scalar, but the subsequent proposed correction estimators are discussed for the general case where s_i could be a vector of variables.

2.1 Bias from Omission and Bias from Inclusion

The estimation of β from the omitted variable model given by Equation (4) using ordinary least squares (OLS) would lead to an estimator $\hat{\beta}^o$ such that:

$$\operatorname{plim} \hat{\beta}^o = \operatorname{Var}(s_i)^{-1}\operatorname{Cov}(y_i, s_i) = \beta + \gamma\delta, \tag{6}$$

where $\operatorname{Var}(\cdot)$ and $\operatorname{Cov}(\cdot,\cdot)$ denote the population variance and covariance, respectively, and $\delta = \operatorname{Var}(s_i)^{-1}\operatorname{Cov}(s_i, z_i)$ is the population slope coefficient in the regression of z_i on s_i. The size of the omitted variable bias is given by $\operatorname{Bias}(\hat{\beta}^o) = \operatorname{plim} \hat{\beta}^o - \beta = \gamma\delta$ and depends on both the coefficient γ in the true model (1) as well as the coefficient δ in the (infeasible) regression of the omitted confounder z_i on the main regressor of interest s_i.

On the other hand, estimating β from the proxy model given by Equation (5) using least squares yields an estimator $\hat{\beta}^c$ such that:

$$\operatorname{plim} \hat{\beta}^c = \beta + \gamma\operatorname{Var}(s_i|x_i)^{-1}\operatorname{Cov}(s_i, z_i|x_i), \tag{7}$$

where $s_i|x_i$ and $z_i|x_i$ are the residuals of the variables s_i and z_i obtained from partialling out x_i. That is, $s_i|x_i$ and $z_i|x_i$ are the residuals of the linear regressions of s_i and z_i on x_i, respectively. These residuals' variances can be expressed in terms of standard variances as follows.

$$\operatorname{Var}(s_i|x_i) = \operatorname{Var}(s_i) - \frac{\operatorname{Cov}(s_i, x_i)\operatorname{Cov}(x_i, s_i)}{\operatorname{Var}(x_i)} = \operatorname{Var}(s_i)\left[1 - R^2_{s,x}\right],$$

where $R^2_{s,x}$ is the population R^2 of the regression of s_i on x_i. For the residuals' covariance, we can also write

$$\operatorname{Cov}(s_i, z_i|x_i) = \operatorname{Cov}(s_i, z_i) - \frac{\operatorname{Cov}(s_i, x_i)\operatorname{Cov}(x_i, z_i)}{\operatorname{Var}(x_i)}.$$

The expression of the bias in the estimation of β from the proxy model given in Equation (7) is quite general and covers many possible scenarios including those not necessarily specific to our setup. Our interest lies in the case where z_i is binary and the measurement error in x_i is a misclassification in z_i (non-classical error in the true binary variable), exogenous to the outcome.

2.2 Bias Comparison

To understand the leading component of the bias in $\hat{\beta}^c$ from Equation (7) for our binary proxy case, it is useful to rewrite the elements in $\text{Cov}(s_i, z_i|x_i)$ in terms of the observables x_i and s_i and the misclassification probabilities. As shown by previous authors (e.g., Aigner 1973, Frazis & Loewenstein 2003) we can write:

$$\text{Cov}(s_i, z_i) = k_1 \text{Cov}(s_i, x_i) \text{ and } \text{Cov}(x_i, z_i) = k_1 k_2 \text{Var}(x_i), \text{ where}$$

$$k_1 = \frac{1}{1-p-q} \quad \text{and} \quad k_2 = \frac{(p_x - p)(1 - p_x - q)}{p_x(1-p_x)}, \quad \text{with} \quad p_x = \Pr[x_i = 1]. \tag{8}$$

Hence, $\text{Cov}(s_i, z_i|x_i) = (1 - k_2)\text{Cov}(s_i, z_i)$, and it follows that

$$\text{plim}\, \hat{\beta}^c = \beta + \gamma\delta \frac{1 - k_2}{1 - R^2_{s,x}}. \tag{9}$$

This means the size of the proxy bias is $\text{Bias}(\hat{\beta}^c) = \text{plim}\, \hat{\beta}^c - \beta = \gamma\delta \frac{1-k_2}{1-R^2_{s,x}}$. The relative size of the bias therefore depends on the relative size of k_2 with respect to $R^2_{s,x}$. We know that $0 < R^2_{s,x} < 1$. Likewise, given that $0 \leq p < p_x$ and $0 \leq q < 1 - p_x$ as implied by the condition $p + q < 1$, we also have $0 < k_2 \leq 1$ (see, e.g., Bollinger 1996). For high degrees of corruption (i.e. p and q are large), k_2 is low, while for low levels of corruption (i.e. p and q are small), k_2 is high. Also note that we can alternatively write $k_2 = (1 - p - q)^2 \frac{\text{Var}(z_i)}{\text{Var}(x_i)}$, i.e., k_2 is proportional to the ratio of the variance of the true confounder over the variance of its corrupt proxy (often referred to as the 'reliability ratio'). The quantity k_2 can therefore be regarded as capturing the proportion of signal in the corrupt data. We can write

$$\text{Bias}(\hat{\beta}^c) = \lambda \text{Bias}(\hat{\beta}^o),$$

where the factor $\lambda = \frac{1-k_2}{1-R^2_{s,x}}$ quantifies the relative size of the corrupt proxy bias compared to the omitted variable bias. Unlike in the continuous cases of non-classical measurement error, the associated bias is necessarily smaller than the omitted variable bias. We formalize this fundamental property in the following theorem.

Theorem 1. *Let $\hat{\beta}^o$ be the least squares estimator of β in the model where the binary confounder is omitted (4), and $\hat{\beta}^c$ the least squares estimator of β in the model where a corrupt binary proxy is included (5). Then,*

(i) Directional Agreement: $\text{Bias}(\hat{\beta}^c)$ *has the exact same sign as* $\text{Bias}(\hat{\beta}^o)$.

(ii) Bias Reduction: $|\text{Bias}(\hat{\beta}^c)| \leq |\text{Bias}(\hat{\beta}^o)|$.

This result comes from the fact that, as shown in the appendix, we must always have $0 \leq \lambda \leq 1$ under the model assumptions, which means $k_2 \geq R^2_{s,x}$. In other words, as long as x_i is an informative proxy for the unobserved true confounder z_i, the proportion of signal in the corrupt data will generally be higher than the amount of correlation between the proxy and the main regressor of interest. There are extreme cases in which these two biases would be the same. An example is when the binary proxy suffers from one-sided misreporting such that, say, $q = 0$, and $p^* = p_x(1 - R^2_{s,x})$. Then $k_2(p^*, 0) = R^2_{s,x}$, which implies that $\lambda = 1$ and hence $\text{Bias}(\hat{\beta}^c) = \text{Bias}(\hat{\beta}^o)$. This means that the corrupt binary proxy is extremely noisy if the proportion of false positives in the data approaches $p_x(1 - R^2_{s,x})$. Since this threshold is a sample statistic that can be measured, we can readily assess how bad the corrupt confounder in the regression model is if the proportion of false positives is known (e.g., through validation data) or can be estimated. By interchanging the roles of p and q, we can establish a similar condition for data where the proxy confounder has only false negatives, i.e., $p = 0$ and $q^* = (1 - p_x)(1 - R^2_{s,x})$, which leads to $k_2(0, q^*) = R^2_{s,x}$ and $\lambda = 1$ so that $\text{Bias}(\hat{\beta}^c) = \text{Bias}(\hat{\beta}^o)$.

Since misclassification is assumed nondifferential with respect to s_i such that $p = \Pr(x_i = 1|z_i = 0, s_i)$ and $q = \Pr(x_i = 0|z_i = 1, s_i)$ do not depend on s_i or other covariates, Theorem 1 aligns conceptually with the classical "intermediate value" property established in epidemiology and biostatistics by Greenland (1980), Ogburn & VanderWeele (2012), Ogburn & Vanderweele (2013), and Sjoelander, Peña & Gabriel (2022). These studies show that partially controlling for a nondifferentially misclassified binary confounder yields an effect estimate bounded strictly between the crude (unadjusted) and true parameters. Our result extends this intuition to a standard OLS linear regression framework, resolving a specific tension in the econometric literature. While modern studies on non-classical measurement in continuous confounders (e.g., Frost 1979, Bound et al. 2001, Bollinger 2003, De Luca et al. 2018) warn that non-classical error structures can induce severe expansion bias, making proxy inclusion worse than omission, Theorem 1 demonstrates that the discrete, bounded support of binary variables prevents such expansion. Thus, even though binary misclassification is inherently non-classical ($\text{Cov}(u_i, z_i) < 0$), Greenland's intermediate value property holds robustly in linear regression settings.

3 Correction Estimators

A central methodological challenge in estimating model (1) under non-classical measurement error is securing point identification of the misclassification parameters and, by extension, the primary causal effect β. In the existing literature, point identification is traditionally achieved through auxiliary validation data or external instrumental variables (e.g., Frazis & Loewenstein 2003, Mahajan 2006, Lewbel 2007, Hu 2008, Kreider et al. 2012, Meyer & Mittag 2017, DiTraglia & García-Jimeno 2019, Haider & Stephens Jr 2025). Where valid instruments exist, an IV-based strategy remains the preferred estimation approach that can provide a non-parametric foundation for identification without relying on specific functional-form restrictions. In practice, however, valid instrumental variables for measurement error processes are notoriously difficult to find, particularly when working with secondary survey data in developing economies where no suitable instrument or validation sub-sample exists. In such scenarios, insisting on an IV strategy leaves the practitioner with a stark choice between abandoning point identification entirely and settling for partial-identification bounds (Bollinger 1996, Black et al. 2000, van Hasselt & Bollinger 2012, Gregory & van Hasselt 2025) or ignoring covariate misclassification and suffering severe bias as shown above.

In this paper, we present practical alternative identification strategies that could be useful for precisely these data-constrained environments. The proposed solutions are based on either a two-step estimator that estimates misclassification probabilities as a first step, or a GMM estimator that jointly estimates misclassification probabilities along with model parameters. Estimating these probabilities requires specifying a misclassification model assuming knowledge of the conditional distribution and predictors for the true binary confounder. For the rest of the discussion in this section, we assume that s_i is an l-vector of correctly measured regressors and that all variables are in deviation from their sample means, so that we focus on slopes and can deal with the model intercept separately. For a sample of size n, we denote $\boldsymbol{y} = [y_1, \ldots, y_n]'$, $\boldsymbol{x} = [x_1, \ldots, x_n]'$, $\boldsymbol{z} = [z_1, \ldots, z_n]'$, and $\boldsymbol{s} = \begin{bmatrix} s_{11} & \ldots & s_{1l} \\ \vdots & \ddots & \vdots \\ s_{n1} & \ldots & s_{nl} \end{bmatrix}$. Using s_i and x_i, respectively, to build moment conditions from the true model given by Equation (1), we obtain the following population equations:

$$\begin{bmatrix} \text{Var}(s_i) & \text{Cov}(s_i, z_i) \\ \text{Cov}(x_i, s_i) & \text{Cov}(x_i, z_i) \end{bmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{bmatrix} \text{Cov}(s_i, y_i) \\ \text{Cov}(x_i, y_i) \end{bmatrix}$$

which, by the covariance relationships obtained earlier, yields:

$$\begin{bmatrix} \text{Var}(s_i) & k_1\text{Cov}(s_i, x_i) \\ \text{Cov}(x_i, s_i) & k_1k_2\text{Var}(x_i) \end{bmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{bmatrix} \text{Cov}(s_i, y_i) \\ \text{Cov}(x_i, y_i) \end{bmatrix}. \tag{10}$$

Equation (10) characterizes the relationship between the model's true parameters and sample statistics, and provides the basis for the methods we proposed to identify and consistently estimate β and γ. Notice that when $p = q = 0$ or if there is no measurement error in the confounder, then $k_1 = k_2 = 1$ and Equation (10) simplifies to the normal equations of the OLS estimators $\hat{\beta}^c$ and $\hat{\gamma}^c$ in the operational model given by Equation (5). We start by discussing the infeasible case where the misclassification probabilities (p and q) are known a priori.

3.1 Estimation with Known Misclassification Rates

Before presenting our main two-step and GMM estimators, it is instructive to first examine the oracle case where the misclassification probabilities (p and q) are known a priori. In data-constrained environments, such as household surveys in developing countries, these misclassification parameters are virtually never known to the researcher, rendering this benchmark impractical for direct empirical estimation. Nevertheless, analyzing it helps isolate the mechanics of the bias-correction formula under ideal identification conditions and provides a useful benchmark against which feasible estimators can be evaluated. It also serves as the conceptual foundation for our proposed two-step and GMM strategies, which operationalize this baseline by replacing the unknown p and q with estimates identified internally via functional-form and covariate variation in the conditional distribution of z_i.

Given knowledge of (p, q), the factors k_1 and k_2 are known and Equation (10) fully identifies γ and β. We can therefore define consistent estimators for γ and β by solving for them in Equation (10), and taking the sample analogs. The result is summarized in the following proposition.

Theorem 2. *Consider the estimation of Model (1) when only a corrupt binary proxy confounder is available. For given misclassification probabilities (p, q) in the corrupt proxy, define the estimators $\hat{\beta}$ and $\hat{\gamma}$ by:*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{bmatrix} \boldsymbol{s}'\boldsymbol{s} & k_1\boldsymbol{s}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{s} & k_1k_2\boldsymbol{x}'\boldsymbol{x} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{s}'\boldsymbol{y} \\ \boldsymbol{x}'\boldsymbol{y} \end{bmatrix} \tag{11}$$

where k_1 and k_2 are functions of (p, q) given in Equations (8). Then

1. *$\hat{\beta}$ and $\hat{\gamma}$ are consistent estimators of β and γ i.e.,* $\text{plim}\,\hat{\beta} = \beta$ *and* $\text{plim}\,\hat{\gamma} = \gamma$.

2. *$\hat{\beta}$ and $\hat{\gamma}$ are asymptotically normal, i.e.,* $\sqrt{n}\begin{pmatrix} \hat{\beta}-\beta \\ \hat{\gamma}-\gamma \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$*, where* $\mathbf{V}$ *is their asymptotic covariance matrix.*

Proof. Appendix. □

A sample-based consistent estimator for the asymptotic covariance matrix, $\mathbf{V}$, is provided in the appendix so that robust inference can be performed with the estimated coefficients using standard methods. This variance matrix is also sufficiently general to accommodate heteroskedasticity of unknown form in the model errors. When there is no measurement error, i.e., $p = q = 0$, these estimators are just the least squares estimators, i.e., $\hat{\beta} = \hat{\beta}^c$ and $\hat{\gamma} = \hat{\gamma}^c$. When there is measurement error and (p, q) can be observed, these consistent estimators can be obtained from the naive OLS estimator in probability limits as follows:

$$\begin{aligned} \hat{\gamma} &= \frac{1}{k_1\left[1-(1-k_2)\dfrac{\mathrm{Var}(x)}{\mathrm{Var}(x|s)}\right]}\hat{\gamma}^c, \\ \hat{\beta} &= \hat{\beta}^c - \hat{\gamma}k_1(1-k_2)\mathrm{Var}(s|x)^{-1}\mathrm{Cov}(s,x). \end{aligned}$$

This means that when misclassification probabilities are known, consistent estimators for γ and β are simply linear adjustments of the naive OLS estimators from the operational model, where the adjustment factors are sample statistics. This resembles the relationship between the OLS and the modified least squares (MLS) procedure of Aigner (1973) where the MLS is just a linear modification of the OLS. The bias adjusted least squares procedure of Nguimkeu, Rosenman & Tennekoon (2021) also bears a similar property. These approaches provide a way to correct for the bias resulting from misclassified regressors using only naive OLS estimation rather than resorting to complicated estimation procedures.

Although the main attention may not be on the intercept α, its consistent estimation can readily obtain from the consistent estimation of β and γ as follows:

Corollary 1. *Under the conditions of Theorem 2, define the estimator of α for Model (1) by*

$$\hat{\alpha} = \bar{y} - \hat{\beta}'\bar{s} - \hat{\gamma}k_1(\bar{x} - p). \tag{12}$$

Then $\hat{\alpha}$ is consistent and asymptotically normal, i.e., $\operatorname{plim}\hat{\alpha} = \alpha$ *and* $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, v_\alpha)$*, where v_α is its asymptotic variance.*

Proof. Appendix. □

These types of correction estimators for $(\alpha, \beta', \gamma)'$ are usually motivated by the availability of misclassification probabilities (p, q) that can generally be obtained through validation samples in some U.S. data (e.g., Aigner 1973, Freeman 1984, Card 1996, Savoca 2000, Battistin, Nadai & Sianesi 2014, Courtemanche, Denteh & Tchernis 2019). When these probabilities are not directly available, we can exploit useful information either in the form of distributional assumptions or available predictors to derive the misclassification probabilities as a first step of a two-step approach or within a generalized method of moments (GMM) approach.

3.2 Two-Step Estimator

A limitation of the above estimator is that it requires upfront knowledge of the misclassification probabilities of the corrupt confounder. In instances where validation studies exist, these probabilities are known. But more often they are unknown and this is likely the most common and important case. Therefore, they may need to be estimated in advance as a first step. We suggest to replace the unknown misclassifcation probabilities (p, q) with their consistent estimators $(\hat{p}, \hat{q})$ in Equations (11) and (12) above. The asymptotic properties of the ensuing estimators remain the same, i.e., they are consistent and asymptotically normal, although they are likely less efficient due to the sampling error that is carried over from the first stage estimation of the misclassification probabilities.

Let $(\hat{p}, \hat{q})$ be a consistent estimator of (p, q), i.e., $\operatorname{plim} \hat{p} = p$ and $\operatorname{plim} \hat{q} = q$, and define $\hat{k}_1 = k_1(\hat{p}, \hat{q})$ and $\hat{k}_2 = k_2(\hat{p}, \hat{q})$ using the functions given earlier by Equation (8). We have the following result:

Theorem 3. *Let the conditions of Theorem 2 hold and denote by $\hat{\gamma}$ and $\hat{\beta}$ the estimators defined by*

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{bmatrix} \boldsymbol{s}'\boldsymbol{s} & \hat{k}_1 \boldsymbol{s}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{s} & \hat{k}_1 \hat{k}_2 \boldsymbol{x}'\boldsymbol{x} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{s}'\boldsymbol{y} \\ \boldsymbol{x}'\boldsymbol{y} \end{bmatrix}, \tag{13}$$

where $\hat{k}_1$ and $\hat{k}_2$ are functions of $(\hat{p}, \hat{q})$ as defined earlier. Then

1. *$\hat{\gamma}$ and $\hat{\beta}$ are consistent estimators of γ and β, i.e., $\operatorname{plim} \hat{\gamma} = \gamma$ and $\operatorname{plim} \hat{\beta} = \beta$.*
2. *$\hat{\gamma}$ and $\hat{\beta}$ are asymptotically normal, i.e., $\sqrt{n} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\gamma} - \gamma \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}^*)$, where $\mathbf{V}^*$ is their asymptotic covariance matrix.*

Proof. Appendix. □

A consistent estimator for the asymptotic covariance matrix, $\mathbf{V}^*$, is provided in the appendix. In practice, there are a few options through which consistent misclassification probabilities $(\hat{p}, \hat{q})$ can be obtained. One approach is to assume a conditional probability function, $\Pr(z_i = 1|\boldsymbol{w}_i)$, for the true confounder z_i, where $\boldsymbol{w}_i$ is a vector of predictors for z_i, and use it to estimate the misclassification probabilities (p, q) by maximum likelihood (MLE) or nonlinear least squares (NLS) as suggested by Hausman et al. (1998). For instance, noticing that one can write $\Pr(x_i = 1|\boldsymbol{w}_i) = p + (1-p-q)\Pr(z_i = 1|\boldsymbol{w}_i)$, the estimators $(\hat{p}, \hat{q})$ can be obtained by maximizing the associated log-likelihood function

$$\sum_{i=1}^{n} x_i \log(\Pr(x_i = 1|\boldsymbol{w}_i)) + (1 - x_i) \log(1 - \Pr(x_i = 1|\boldsymbol{w}_i)) = \sum_{i=1}^{n} \ell_i(p, q, \boldsymbol{\pi}),$$

where the conditional probability function of $z = 1$ is given by $\Pr(z_i = 1|\boldsymbol{w}_i) = F(\boldsymbol{w}_i'\boldsymbol{\pi})$, for some known $F(\cdot)$, e.g., normal or logistic distribution. One can also estimate this model with nonlinear least squares by minimizing the function

$$\sum_{i=1}^{n} (y_i - p - (1 - p - q)F(\boldsymbol{w}_i'\boldsymbol{\pi}))^2 = \sum_{i=1}^{n} \ell_i(p, q, \boldsymbol{\pi}).$$

As explained in Hausman et al. (1998), conditions for identification of $\phi = (p, q, \boldsymbol{\pi}')'$ are similar to those of the traditional binary choice model discussed in Newey & McFadden (1994), with the additional monotonicity condition given earlier by $p + q < 1$. This means that identification of the model parameters is achieved by exploiting the nonlinearity of $F(\cdot)$ alongside variations in $\boldsymbol{w}_i$. The linear probability model (where $F(v) = v$) therefore cannot individually identify the parameters $(p, q, \boldsymbol{\pi})$ and is not recommended for applying the proposed method. For instance, parameter identification in the probit estimation that we use in our simulations and empirical application in the subsequent sections would require $F(\cdot)$ to be a normal CDF, $\mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_i']$ to be a nonsingular matrix, and $p + q < 1$.

Other methods to obtain $(\hat{p}, \hat{q})$ through parametric maximum likelihood, quasi-maximum likelihood or nonlinear least squares estimation can be found in Bollinger & David (1997) and Poterba & Summers (1995). One could also explore semiparametric procedures which do not require distributional assumptions, such as the semiparametric maximum likelihood estimator of Hausman et al. (1998) based on the maximum rank correlation estimator of Han (1987), the maximum score estimator of Manski (1985), or the semiparametric approach of Abrevaya & Hausman (1999) based on the monotone rank estimator of Cavanagh & Sherman (1998). The main drawback of nonparametric or semiparametric methods as a first stage, however, is that they are, in general, slower to converge than the parametric rate,

potentially yielding inflated variance in the second stage.

The sample analogue of the asymptotic variance $\mathbf{V}^*$ provides standard errors that are robust to both heteroskedasticity of any form and the sampling error of the estimated misclassification probabilities obtained in the first stage.

As with the known misclassification probabilities case, if one is interested in estimating the intercept in Model (1), a consistent and asymptotically normal estimator can use the estimated misclassification probabilities and be obtained similarly as follows:

Corollary 2. *Under the conditions of Theorem 3, define the estimator of α for Model (1) by*

$$\hat{\alpha} = \bar{y} - \hat{\beta}'\bar{s} - \hat{\gamma}\hat{k}_1(\bar{x} - \hat{p}). \tag{14}$$

Then $\hat{\alpha}$ is consistent and asymptotically normal, i.e., $\operatorname{plim}\hat{\alpha} = \alpha$ *and* $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, v_\alpha^*)$*, where v_α^* is its asymptotic variance.*

Proof. Appendix. □

Although the asymptotic variance and its sample-based analog for the two-step estimator of (α, β', γ) are given in the appendix, many researchers sometimes prefer to use bootstrapping for these types of models in practice (Gonçalves, Hounyo, Patton & Sheppard 2023). We suggest using a non-parametric bootstrap procedure that samples the data running both the Hausman et al. (1998) first step and the second step estimator in every iteration, so that the bootstrap automatically captures the transmission of sampling error from $(\hat{p}, \hat{q})$ to $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ without requiring the analytical derivation of the Jacobian.

3.3 Efficient GMM Estimator

The two-step estimator discussed above is consistent, numerically convenient and establishes a clear hierarchy between the misclassification model and the structural regression. It is, however, important to note two primary considerations for the practitioner: (i) By estimating p and q separately from the structural parameters (β, γ), we potentially sacrifice asymptotic efficiency. A joint GMM approach would stack the MLE scores or NLS gradient and the structural moments into a single vector. By solving this system simultaneously, one could achieve the semi-parametric efficiency bound. (ii) In the two-step framework, we assume information flows only from the misclassification model to the regression. In reality, the relationship between y_i and the proxy x_i can actually contain information on the probability of z_i being misclassified. A joint estimation would "leak" this structural information

back into the estimates of p and q, potentially improving their precision.

We consider the joint estimation of $\eta = (\beta', \gamma, \phi')'$, where $\phi = (p, q, \boldsymbol{\pi}')'$ is the parameter vector of the misclassification model. We define the stacked moment vector $g_i(\eta)$:

$$g_i(\eta) = \begin{bmatrix} \nabla_\phi \ell_i(\phi) \\ s_i(y_i - s_i'\beta - k_1\gamma x_i) \\ x_i(y_i - s_i'\beta - k_1 k_2 \gamma x_i) \end{bmatrix},$$

where $\nabla_\phi \ell_i(\phi)$ is the gradient of the misclassification objective function and the remaining components of $g_i(\eta)$ represent the structural regression moments implied by Equation (10). The efficient GMM estimator $\hat{\eta}$ minimizes the quadratic form:

$$\hat{\eta} = \arg\min_{\eta} \left(\sum_{i=1}^{n} g_i(\eta) \right)' \mathbf{W}_n \left(\sum_{i=1}^{n} g_i(\eta) \right),$$

where $\mathbf{W}_n$ is a symmetric positive-definite the weighting matrix. Since the system is exactly identified (i.e. the number of moment conditions equals the number of parameters), the sample estimator $\hat{\eta}$ satisfies the system of equations exactly (i.e., $\sum_{i=1}^{n} g_i(\hat{\eta}) = 0$) and the choice of $\mathbf{W}_n$ needs not be optimal to achieve efficiency. As above, identification relies on the nonlinearity of $F(\cdot)$ alongside variations in $\boldsymbol{w}_i$, as well as the rank condition for the matrix $[\boldsymbol{s}\, \boldsymbol{x}]$.

Theorem 4. *Under the model assumptions and standard regularity conditions (compact parameter space, continuous moments, identification),*

1. *$\hat{\eta}$ is consistent, i.e., $\hat{\eta} \xrightarrow{p} \eta_0$.*
2. *$\hat{\eta}$ is asymptotically normal, i.e., $\sqrt{n}(\hat{\eta} - \eta_0) \xrightarrow{d} N\left(0, (G'\Omega^{-1}G)^{-1}\right)$, where the components of the asymptotic variance, $(G'\Omega^{-1}G)^{-1}$, are given by $\Omega = E[g_i(\eta_0)g_i(\eta_0)']$ and $G = E[\nabla_\eta g_i(\eta_0)]$.*

Proof. Appendix. □

A sample-based consistent estimator of the asymptotic variance, $(G'\Omega^{-1}G)^{-1}$, is provided in the appendix. Under this specification, the asymptotic variance of $\hat{\theta} = (\hat{\beta}, \hat{\gamma})$ reaches the semi-parametric efficiency bound, as it utilizes the full covariance structure between the misclassification score and the structural error term. The standard errors obtained from the sample analog of the asymptotic variance are robust to any form of heteroskedasticity in the structural error term. The estimator of the intercept, $\hat{\alpha}$, is given by Equation (14) and has the same asymptotic properties, that is, consistent and asymptotically normal. As with the Two-Step estimator above, bootstrapping can also be employed to obtain standard errors for the GMM estimator of $(\alpha, \beta', \gamma, p, q.\boldsymbol{\pi})$ and used for statistical inference (e.g,. Hahn 1996, Brown & Newey 2002, Dovonon & Gonçalves 2017).

4 Monte Carlo Simulation Study

This section provides finite-sample validation for both the problem and the recommended solution. We compute the cost of a corrupt confounder at various degrees of misclassification, we evaluate the biases in naive estimation, and we assess the finite sample performance of the proposed estimators with Monte Carlo simulations.

4.1 Simulation Design

The data generating process is simulated as follows. The main predictor of the binary confounder is $w_i \sim \chi^2_{(1)}$, the true confounder is $z_i = \mathbf{1}[-1 + \pi w_i + v_i > 0]$, where $v_i \sim \mathcal{N}(0,1)$. The main explanatory variable is $s_i = \kappa z_i + \xi_i$, $\xi_i \sim \chi^2_{(1)}$, with $\kappa \in \{-0.8, 0.8\}$. The parameter κ controls the level and sign of the correlation between the variable of interest, s_i, and the true binary confounder, z_i. The true outcome equation is given by:

$$y_i = 1 - 0.2s_i + 4z_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0,2)$$

The researcher does not observe z_i but rather a corrupt proxy x_i, with misclassification rates $p, q \in [0,1)$, such that

$$x_i = z_i \mathbf{1}[u_i > q] + (1 - z_i)\mathbf{1}[u_i < p].$$

This means p is the rate of false positives and q is the rate of false negatives. The operational or proxy model that the researcher can estimate is therefore

$$y_i = \alpha + \beta s_i + \gamma x_i + \epsilon_i$$

We first fix $\pi = 1$ and we consider a set of different combinations values of misclassification probabilities (p, q), that is, $(0.15, 0.15)$; $(0.15, 0.30)$; $(0.30, 0.15)$, which allows us to evaluate the bias and performance of the proxy and proposed estimators at different levels of corruption in the misclassified binary confounder, as well as when the researcher chooses to omit the confounder entirely.

The estimates of the model parameters, (β, γ), are obtained for the infeasible OLS estimator using the correct confounder (Oracle OLS), the pessimistic OLS estimator that omits the confounder entirely (Omitted OLS), the naive OLS estimator that includes the corrupt binary proxy (Proxy OLS), and the proposed correction estimators: a two-step estimator that estimates misclassification probabilities with the Hausman et al. (1998) approach in a first step (Two-Step) and a GMM estimator that jointly estimates all parameters together (Joint GMM). We

only report the estimates of our parameter of interest, β. The estimated probabilities of misclassification assume the correct specification of the distribution of the correct confounder.[1] All our results are based on $N = 1000$ simulations with a generated sample of $n = 5000$ observations.

4.2 Finite Sample Performance of Alternative Estimators

The Monte Carlo simulation results in Table 1 provide comprehensive finite sample validation for our main theoretical claims across various degrees of correlation and corruption. We use the Oracle OLS estimator, which utilizes the true confounder z_i, as the ideal benchmark, and the Omitted OLS estimator as the measure of the maximum possible bias, which is the Omitted Variable Bias (OVB).

Our primary finding is the empirical confirmation that, for a misclassified binary confounder, including the corrupt binary proxy is strictly preferable to omitting it. As expected from our DGP, Omitted OLS estimator consistently exhibits severe omitted variable bias in the estimate of the coefficient of interest whose true value is set at $\beta = -0.20$. When $\kappa = -0.8$, the $\hat{\beta}$ estimate from Omitted OLS is near -0.322 and when $\kappa = 0.8$, it is near 0.500, a complete sign switch. This large and systematic bias, including possible sign-flipping, reflects the cost of ignoring the possibly strong correlation between s_i and z_i. In every simulation scenario, the $\hat{\beta}$ estimate from Proxy OLS estimator is demonstrably closer to the true value than the Omitted OLS estimate. When $\kappa = -0.8$, the omitted OLS estimator has a bias of approximately -0.12, whereas the naive Proxy OLS estimator has biases ranging from about -0.095 to -0.109. The same result is even more striking when $\kappa = 0.8$, and remains true even when the estimated coefficients have the wrong sign as in this case. Hence, including the corrupt proxy always achieves partial bias reduction. This pattern empirically validates our main theoretical result that $|\text{Bias}(\hat{\beta}_{\text{Include}})| < |\text{Bias}(\hat{\beta}_{\text{Omit}})|$. It confirms that the expansion bias mechanism observed in misclassified treatment variables (e.g. Nguimkeu et al. 2019, Wossen et al. 2019) does not dominate the bias reduction achieved for a coefficient of interest by including a corrupt binary confounder.

Although the Proxy OLS estimator is better than the omitted OLS estimator, it still systematically suffers from residual asymptotic bias. When $\kappa = -0.8$, this bias increases with higher levels of corruption, and when $\kappa = 0.8$, the bias is so severe that the estimator completely flips sign as well. This persistent bias is substantial

[1] We do not report the estimator that takes misclassification probabilities as known since it is similar to the case of known z, the Oracle OLS. We also do not report the case of misclassification implied by a misspecified model using logistic v_i, as it gives similar results as though it was correctly specified.

Table 1: Monte Carlo Performance of Alternative Estimators

ρ	(p, q)	Estimator	Mean	Bias	SD	RMSE
		Oracle OLS	-0.199	0.001	0.014	0.014
		Omitted OLS	-0.320	-0.120	0.016	0.121
	$(0.15, 0.15)$	Proxy OLS	-0.295	-0.095	0.016	0.096
		Two-Step	-0.211	-0.011	0.063	0.064
		Joint GMM	-0.213	-0.013	0.046	0.056
		Oracle OLS	-0.200	0.000	0.014	0.014
		Omitted OLS	-0.322	-0.122	0.016	0.123
-0.9	$(0.15, 0.30)$	Proxy OLS	-0.305	-0.105	0.016	0.106
		Two-Step	-0.197	0.003	0.051	0.051
		Joint GMM	-0.222	-0.022	0.043	0.048
		Oracle OLS	-0.199	0.001	0.014	0.014
		Omitted OLS	-0.320	-0.120	0.016	0.121
	$(0.30, 0.15)$	Proxy OLS	-0.309	-0.109	0.016	0.110
		Two-Step	-0.196	0.004	0.086	0.086
		Joint GMM	-0.202	-0.002	0.070	0.082
		Oracle OLS	-0.199	0.001	0.016	0.016
		Omitted OLS	0.498	0.698	0.027	0.698
	$(0.15, 0.15)$	Proxy OLS	0.212	0.412	0.024	0.413
		Two-Step	-0.231	-0.031	0.254	0.255
		Joint GMM	-0.233	-0.033	0.046	0.056
		Oracle OLS	-0.200	0.000	0.017	0.017
		Omitted OLS	0.498	0.698	0.028	0.698
0.9	$(0.15, 0.30)$	Proxy OLS	0.321	0.521	0.026	0.522
		Two-Step	-0.215	-0.015	0.110	0.111
		Joint GMM	-0.212	-0.012	0.090	0.091
		Oracle OLS	-0.200	-0.000	0.017	0.017
		Omitted OLS	0.498	0.698	0.028	0.698
	$(0.30, 0.15)$	Proxy OLS	0.333	0.533	0.026	0.534
		Two-Step	-0.212	-0.012	0.071	0.072
		Joint GMM	-0.211	-0.011	0.063	0.064

Notes: The true coefficient of interest is $\beta = -0.20$, and the coefficient on the binary confounder is $\gamma = 4.00$. "Oracle OLS" uses the true binary confounder z_i. "Omitted OLS" excludes the confounder entirely. "Proxy OLS" uses the misclassified proxy x_i in place of z_i. "Two-Step" estimates the misclassification probabilities in a first step and applies the proposed correction in the second step. "Joint GMM" estimates the misclassification and structural moment restrictions jointly. Bias is defined as $\mathbb{E}[\hat{\beta}] - \beta$. SD denotes the Monte Carlo standard deviation, and RMSE denotes the root mean squared error.

and could lead to dramatic policy prescriptions, especially with sign-switching estimates. The proposed correction procedures then address the residual bias left after proxy inclusion as both the Two-Step and Joint GMM estimators successfully eliminate the asymptotic bias inherent in OLS. Across all regimes, empirical means for both estimators closely track the true parameter value ($\beta = -0.20$), with bias generally remaining below -0.033. Increasing false positive (p) or false negative (q) rates inflates estimation variance (SD and RMSE) compared to the baseline Oracle OLS. The variance penalty reflects the information loss from estimating p and q in the first step or jointly. However, Joint GMM systematically outperforms Two-Step estimator in finite-sample efficiency (lower SD and RMSE). For instance, under $\kappa = 0.8$ and $(p, q) = (0.15, 0.15)$, Two-Step exhibits a higher standard deviation at SD $= 0.254$ (RMSE $= 0.255$) due to first-stage error propagation. In contrast, Joint GMM dramatically stabilizes estimation, cutting the standard deviation to SD $= 0.046$ (RMSE $= 0.056$). This represents a five-fold increase in estimation precision, bringing efficiency within close proximity of the unobservable Oracle OLS benchmark (SD $= 0.016$).

4.3 Robustness under Identification Breakdown

The Monte Carlo simulation results in Table 2 demonstrate how parameter identification degrades as the first-stage strength (π) approaches zero, illustrating classic weak identification dynamics within non-linear misclassification models.[2]. Both standard OLS benchmarks display substantial positive bias across all levels of π. While their bias decreases mechanically as the underlying dependence weakens, Proxy OLS consistently achieves lower bias than Omitted OLS (e.g., 0.256 vs. 0.363 at $\pi = 0.3$), reaffirming that inclusion of an imperfect proxy reduces error relative to total omission even under weak identification, though it remains heavily biased. When $\pi = 0.3$, both corrective estimators still successfully pull the parameter estimates back toward the true parameter value, though both carry substantial variance penalties (SD $\approx$ 0.186–0.195) compared to Oracle OLS. As the first-stage signal weakens further ($\pi = 0.1$), first-stage estimation error in the misclassification parameters $(\hat{p}, \hat{q})$ begins to contaminate structural estimation. This is consistent with the fact that the correction relies on estimated misclassification parameters, so weak identification in the first-stage model translates into additional uncertainty in the structural estimator. Both estimators exhibit severe variance amplification, resulting in a positive bias shift ($\hat{\beta} \approx -0.138$ for Two-Step and -0.129 for Joint GMM) as the non-linear identification bounds weaken.

[2]The assumption is that the researcher is able to choose a proper non-linear $F(\cdot)$ so that identification largely relies on variation in $\boldsymbol{w}_i$

Table 2: Performance Under Weak Identification

π	Identification	Estimator	Mean	Bias	SD	RMSE
0.3	Weak	Oracle OLS	-0.200	-0.000	0.022	0.022
		Omitted OLS	0.163	0.363	0.043	0.366
		Proxy OLS	0.056	0.256	0.037	0.259
		Two-Step	-0.241	-0.041	0.195	0.199
		Joint GMM	-0.154	0.046	0.186	0.191
0.1	Very Weak	Oracle OLS	-0.199	0.001	0.021	0.021
		Omitted OLS	0.033	0.233	0.036	0.236
		Proxy OLS	-0.022	0.178	0.033	0.181
		Two-Step	-0.138	0.062	0.158	0.169
		Joint GMM	-0.129	0.071	0.128	0.146
0.0	No Variation	Oracle OLS	-0.200	0.000	0.023	0.023
		Omitted OLS	0.008	0.208	0.036	0.211
		Proxy OLS	-0.037	0.163	0.032	0.166
		Two-Step	-0.432	-0.232	0.550	0.597
		Joint GMM	-0.077	0.123	0.282	0.307

Notes: The true coefficient of interest is $\beta = -0.20$. The experiment sets $\kappa = 0.8$, $p = 0.15$, and $q = 0.30$, and varies π to alter the strength of identification in the misclassification equation. "Oracle OLS" uses the true confounder. "Omitted OLS" excludes the confounder entirely. "Proxy OLS" uses the misclassified proxy directly. "Two-Step" estimates the misclassification parameters in a first-stage nonlinear model and applies the proposed correction in the structural equation. "Joint GMM" estimates the misclassification and structural restrictions jointly. Bias is defined as $\mathbb{E}[\hat{\beta}] - \beta$.

When $\pi = 0$, the covariate w_i contains zero information on the latent confounder z_i, making the misclassification probabilities p and q non-identifiable from the first-stage equation $\Pr[x_i = 1 \mid w_i]$. The deterioration in the proposed estimators is therefore expected and is not a numerical anomaly. The Two-Step estimator collapses, producing severe negative bias of -0.232 ($\hat{\beta} = -0.432$) and high sampling variance (SD = 0.550, RMSE = 0.597, while Joint GMM also exhibits severe bias (0.123) and inflated variance (SD = 0.282), although with relatively greater finite-sample robustness. The comparison between the two-step and joint GMM estimators is also informative. It shows that across weak-identification designs, joint GMM continues to have lower variance and RMSE than the two-step procedure, a pattern consistent with the efficiency advantage of estimating the misclassification and structural equations jointly rather than treating first-stage estimates as fixed inputs.

In conclusion, these simulations support the two central tenets of our paper. First, the naive approach using a corrupt binary proxy is a necessary first step

that correctly reduces bias compared to omitting the confounder entirely. Second, the proposed correction estimators are robust and effective solutions for achieving full consistency when identification is adequate. They substantially reduce bias compared to both omission and naive proxy inclusion. When identification is weak, the correction methods remain potentially valuable but become noisier. If the first-stage model lacks identifying variation altogether, the variance cost can dominate. By contrast, the theoretical result on proxy inclusion remains robust across these regimes: even when the researcher cannot correct the misclassification, including the proxy still produces less bias than omitting the confounder entirely.

5 Returns to Education in Cameroon

The purpose of this empirical application is primarily methodological. We apply the proposed methods in a real data example of returns to education and gender earnings gaps among household businesses in Cameroon. We use this context to illustrate how misclassification of an economically important binary covariate can affect substantive empirical conclusions, and how the proposed estimators can be implemented in a policy-relevant setting. Consequently, estimates should not be interpreted as definitive causal effects, as such an interpretation would require additional assumptions concerning, among other things, selection, omitted heterogeneity, and endogeneity that are beyond the scope of the present exercise.

5.1 Context and Data

Household businesses dominate the labor force in Cameroon and across Sub-Saharan Africa, yet their income dynamics remain poorly understood due to widespread informality and measurement challenges. In particular, business formality status, typically self-reported in household and enterprise surveys, is prone to misreporting, either because entrepreneurs misunderstand legal definitions, strategically misstate their status, or operate in gray zones between formality and informality. This application re-examines two canonical relationships in the development and labor economics literature: returns to education and gender earnings gaps, by explicitly accounting for misclassification in self-reported business formality. Ignoring such misclassification can bias estimates of human capital returns and distort conclusions about the role of formalization in income generation.

Our empirical application uses data from the 2005 Employment and Informal Sector Survey (EESI 2005) conducted by the National Institute of Statistics of Cameroon (INS-Cameroon 2005, Nguimkeu 2014). The EESI is a nationally representative household survey designed to capture employment outcomes, labor

market participation, and characteristics of informal economic activity. We focus on household business owners, for whom the survey collects detailed information on business income, owner characteristics, and self-reported formality status. This population is of particular interest because household businesses account for a large share of non-agricultural employment in Cameroon, yet operate largely outside the reach of administrative data (INS-Cameroon 2005). The outcome variable is business income, while key covariates include education, age, age-squared, and a female indicator. These variables capture standard human capital and demographic dimensions emphasized in both the labor and development literatures.

A central variable in our analysis is business formality status, denoted z in our framework, which captures whether a household business operates formally or informally. In the survey, formality is measured through a self-reported binary indicator, denoted x in our framework, based on registration or legal status as perceived by the respondent. We treat z as a key binary confounder in the income equation. Business formality is correlated with education, gender, and age, and is also strongly associated with income. Omitting formality therefore induces omitted-variable bias, while including it as a regressor is standard practice in empirical work on household businesses. Self-reported formality status is likely to be misclassified, however. Several features of the institutional environment in Cameroon make accurate reporting difficult. First, enforcement of registration and tax obligations is uneven, and many businesses operate in gray zones between formal and informal activity (Nguimkeu 2022, Kede Ndouna & Tsafack Nanfosso 2023). Second, entrepreneurs may be uncertain about whether their business meets legal thresholds for formality. Third, strategic misreporting may arise due to fear of taxation or regulatory scrutiny (Benjamin & Mbaye 2012, Adom 2025). As a result, the observed indicator x may differ from the true latent state z.

Table 3 reports summary statistics for household businesses in Cameroon, distinguishing between self-reported formal and informal enterprises. Only 11% of businesses report being formal, underscoring the dominance of informality in the economy. Self-reported formal businesses appear sharply different from informal ones along income, education, and gender dimensions. Formal enterprises earn substantially higher income, are run by more educated owners, and are disproportionately male-owned. Although these differences are striking, they are based on self-reported formality status, which may be mismeasured. In settings with weak enforcement and ambiguous regulatory thresholds, businesses may misreport formality or misclassify their legal status. As a result, observed differences between formal and informal enterprises may conflate true productivity gaps with measurement error and selection. This concern is consistent with earlier work emphasizing

Table 3: Summary Statistics: Household Businesses in Cameroon

	All Businesses	Formal	Informal
Panel A: Income and Owner Characteristics			
Income (1000 CFA)	106.70	353.30	75.30
	(140.36)	(229.99)	(82.69)
Education (years)	7.98	12.54	7.40
	(4.30)	(4.40)	(3.93)
Age (years)	37.54	41.67	37.02
	(10.92)	(7.74)	(11.16)
Female (=1)	0.34	0.09	0.37
	(0.47)	(0.29)	(0.48)
Panel B: Formality Status			
Formality (=1)	0.11	1.00	0.00
	(0.32)		
Observations	478	54	424

Notes: Income is monthly income from household businesses in 1000 CFA. Formal indicates self-reported registration or formal status of the business. Standard deviations in parentheses. Formal and informal groups are defined based on self-reported formality.

that informality is a latent economic state rather than a directly observed characteristic (Nguimkeu 2014, 2022). Treating self-reported formality as error-free may therefore bias estimates of returns to education and gender earnings gaps, motivating the application of the proposed misclassification-corrected estimators.

5.2 Results

To estimate returns to education and gender gaps in earnings, we apply the proposed correction procedures, which allow us to estimate misclassification probabilities directly from the data based on Hausman et al. (1998) approach, and to recover consistent estimates of the earnings equation. The results are reported in Table 4. Across all specifications, education is positively associated with log income from household businesses and is significant for all estimators. However, the magnitude of returns varies sharply across estimators. The omitted OLS specification implies an average return to an additional year of education of about 9.1 percent, while treating observed formality status as an error-free proxy lowers this estimate to 5.9 percent. By contrast, both corrected estimators yield substantially larger and remarkably similar estimates: 22.8 percent for the Two-Step estimator and 22.6 percent for Joint GMM. The close correspondence between the Two-Step and Joint GMM estimates is reassuring and suggests that the correction is not

driven by the particular estimation procedure. This pattern indicates that part of what is often interpreted as a low return to education among household businesses actually reflects selection into (true) formality, rather than productivity gains from schooling alone.

Table 4: Returns to Education and Gender Earnings Gaps in Cameroon

Variable	Omitted OLS	Proxy OLS	Two-Step	Joint GMM
Panel A: Structural Log Income Equation Results				
Education	0.091***	0.059***	0.228***	0.226***
	(0.010)	(0.009)	(0.022)	(0.022)
Age	0.127***	0.111***	0.192***	0.191***
	(0.022)	(0.021)	(0.037)	(0.037)
$\text{Age}^2/100$	-0.135***	-0.123***	-0.183***	-0.182***
	(0.027)	(0.026)	(0.045)	(0.044)
Female	-0.747***	-0.677***	-1.047***	-1.042***
	(0.094)	(0.091)	(0.172)	(0.170)
Formality (z)	—	1.171***	-3.201***	-3.199***
	—	(0.109)	(0.903)	(0.895)
Intercept	0.893**	1.396***	-1.827**	-1.785**
	(0.424)	(0.401)	(0.808)	(0.799)
Panel B: Misclassification Parameters (Formality Status)				
False Positive ($\hat{p}$)			0.113***	0.113***
			(0.015)	(0.015)
False Negative ($\hat{q}$)			0.248***	0.238***
			(0.007)	(0.006)

Notes: In Panel A, the dependent variable is log income from household businesses. The Proxy OLS treats self-reported formality status as error-free. The corrected estimators (Two-Step and GMM) account for misclassification in formality using estimated misclassification probabilities following Hausman et al. (1998). Panel B reports misclassification probabilities (p, q) for the Joint GMM and the first-stage of Two-Step. The first stage and score function use all model covariates as predictors of true formality status. Bootstrap standard errors in parentheses. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

The life-cycle earning profile follows a standard concave pattern across all specifications, with income rising with age and declining at older ages. Importantly, these effects remain stable after correction, suggesting that life-cycle earnings dynamics are robust to misclassification in formality status. Again, the nearly identical Two-Step and Joint GMM estimates provide evidence that these substantive conclusions are robust across the two correction procedures. Female-owned household businesses earn substantially less than male-owned businesses in all models.

The raw gender gap is relatively small when formality is omitted (-0.747), or when it is controlled for with a self-reported status (-0.677). After correcting for misclassification, the gap becomes large, approximately -1.05 under both Two-Step and Joint GMM. This suggests that formalization is an important factor in explaining gender income disparities in household businesses. When accounting for true formality status, women face sizeable income penalties, pointing to deeper structural constraints such as differential access to capital, networks, digital technologies or market opportunities.

The most striking result concerns the coefficient on formality itself. When self-reported formality is treated as error-free, Proxy OLS produces a large positive coefficient of 1.171. Once misclassification is explicitly modeled, however, the estimated coefficient becomes approximately -3.20 under both corrected estimators. This is not simply an attenuation correction: accounting for classification error reverses the estimated sign of the relationship between latent true formality status and household-business earnings. This result implies that, conditional on education, age, gender, and the other included covariates, the latent formality state identified by the model is associated with substantially lower earnings. This result is consistent with a weak value proposition of formalization for small-scale enterprises in Africa (Nguimkeu & Okou 2021). Firms that formalize may incur taxes and compliance costs without receiving sufficiently large compensating benefits, while some may enter formality primarily because scale and visibility make continued informality difficult (Nguimkeu 2022, Adom 2025). This suggests that policies centered only on enforcement or registration are unlikely to succeed. Sustainable formalization requires changing the payoff structure by making formal status valuable.

Panel B provides direct evidence that self-reported formality is noisy. The estimated false-positive rate ($p \approx 11.3\%$) under both correction methods implies that about 11 percent of observations classified as formal are truly informal. The estimated false-negative probability is substantially larger ($q \approx 24\%$ to 25%), implying that 24-25 percent of truly formal businesses report themselves as informal. Misclassification is therefore asymmetric: true formal businesses appear considerably more likely to be recorded as informal than true informal businesses are to be recorded as formal. Importantly, the Two-Step and Joint GMM procedures produce very similar estimates of both p and q, providing additional evidence that the estimated misclassification process is stable across estimation methods. This asymmetric misclassification is consistent with institutional realities in Cameroon, where entrepreneurs may underreport formality because of weak enforcement or ambiguity about legal status.

Although this empirical exercise is not designed to identify the causal effects in this setting, the sharp change in some of the estimated coefficients demonstrates that accounting explicitly for misclassification can lead to economically meaningful differences in estimated relationships relative to both omitting the confounder and treating an imperfect proxy as error-free. In the context of small household enterprises in Cameroon, the corrected pattern seems consistent with the local business and institutional environment, but these mechanisms remain suggestive rather than identified in this application.

6 Conclusion

This paper addresses a fundamental dilemma in regression analysis: whether to omit a mismeasured confounder or include a corrupt proxy. While the recent literature on non-classical measurement error has fostered a "fear of the proxy"—suggesting that including imperfect measures may occasionally be more damaging than exclusion—we prove that this caution is unnecessary when the confounder is binary. Our theoretical results demonstrate that including a misclassified binary proxy, such as a business's formality status in an income equation, strictly reduces the bias in the coefficient of interest relative to omitting the variable entirely. This finding provides a robust justification for researchers to utilize available indicators, even when their accuracy is known to be imperfect.

To move beyond mere bias reduction toward consistent estimation, we proposed two methodological solutions: a sequential two-step estimator and a joint GMM framework. By integrating a misclassification model into the structural equation, we derived analytical correction factors that recover the true structural parameters. These estimators are consistent and asymptotically normal, with a sample-based variance estimator that correctly propagates first-stage uncertainty into the final standard errors. Crucially, these corrections rely on two fundamental requirements: sufficient variation in the underlying predictors of the latent state and the essential non-linear shape of their probability distribution. Our Monte Carlo simulations confirm that these methods perform well in finite samples, successfully eliminating the attenuation and contamination biases that plague naive linear models. Our empirical findings from household businesses in Cameroon demonstrate the critical need for these corrections. Naive proxy models severely distort the economic reality by overstating the benefits of formalization while underestimating both the returns to human capital and the gender earnings gap. Once corrected, we find that returns to education and gender disparities are far more pronounced, and the expected gains from formalization turn negative.

In summary, this study provides both a theoretical safety net for using mismeasured binary proxies and a practical toolkit for correcting them in regression analysis. For practitioners, this means that rather than reverting to the known hazards of omitted variable bias out of fear of misclassification error in a binary confounder, one should include the binary proxy and apply the appropriate GMM or two-step corrections proposed in this paper. This ensures credible structural inference and provides a more accurate basis for policy recommendations. Future work might extend this analysis to mismeasured categorical confounders with more than two values, examine the cases of covariate-dependent misclassification, or explore the complex interactions that arise when multiple binary confounders are misclassified.

A Proofs of Theorems

This section contains the proofs of all our theoretical results, where some proofs, especially standard arguments, are deliberately heuristic for space.

A.1 Proof of Theorem 1

Proof. We need to prove that $\lambda \leq 1$, that is, $k_2 \geq R^2_{s,x}$. The regressor of interest, s_i, is assumed to be a scalar here for simplicity. By the Cauchy-Schwarz inequality, we will always have $\text{Cov}(s_i, z_i)^2 \leq \text{Var}(s_i)\text{Var}(z_i)$. Because $\text{Cov}(s_i, z_i) = k_1\text{Cov}(s_i, x_i)$ and $\text{Var}(z_i) = k_1^2 k_2 \text{Var}(x_i)$, this means $\text{Cov}(s_i, x_i)^2 \leq k_2 \text{Var}(s_i)\text{Var}(x_i)$, and hence $k_2 \geq \dfrac{\text{Cov}(s_i, x_i)^2}{\text{Var}(s_i)\text{Var}(x_i)} = R^2_{s,x}$. □

A.2 Proof of Theorem 2

Proof. We prove the consistency and asymptotic normality of $(\hat{\beta}, \hat{\gamma})$ when (p, q) is known. Taking all variables in deviation from their sample means, the true model is given by $y_i = s_i'\beta + \gamma z_i + \varepsilon_i$. Let $\theta = (\beta', \gamma)'$. The proposed estimator $\hat{\theta}$ satisfies the sample moment condition $\dfrac{1}{n}\sum_{i=1}^n \psi_i(\hat{\theta}) = 0$, where:

$$\psi_i(\theta) = \begin{pmatrix} s_i(y_i - s_i'\beta - k_1\gamma x_i) \\ x_i(y_i - s_i'\beta - k_1 k_2 \gamma x_i) \end{pmatrix}$$

i. Consistency

By the Weak Law of Large Numbers (WLLN), the estimator converges to the value θ that satisfies $E[\psi_i(\theta)] = 0$. Substituting the true model into the moment conditions:

$$E[\psi_i(\theta)] = \begin{pmatrix} E[s_i(s_i\beta + \gamma z_i + \varepsilon_i - s_i\beta - k_1\gamma x_i)] \\ E[x_i(s_i\beta + \gamma z_i + \varepsilon_i - s_i\beta - k_1k_2\gamma x_i)] \end{pmatrix} = \begin{pmatrix} \gamma(E[s_iz_i] - k_1E[s_ix_i]) + E[s_i\varepsilon_i] \\ \gamma(E[x_iz_i] - k_1k_2E[x_i^2]) + E[x_i\varepsilon_i] \end{pmatrix}$$

Given $\text{Cov}(s, z) = k_1\text{Cov}(s, x)$, $\text{Cov}(x, z) = k_1k_2\text{Var}(x)$, and $E[\varepsilon|s, z] = 0$, all terms vanish at the true θ. Thus, $\hat{\theta} \xrightarrow{p} \theta$.

ii. Asymptotic Normality

Under standard regularity conditions, $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, \mathbf{V})$, where $\mathbf{V} = A^{-1}B(A')^{-1}$. The Jacobian matrix A is:

$$A = E\left[\frac{\partial\psi_i}{\partial\theta'}\right] = -\begin{bmatrix} E[s_is_i'] & k_1E[s_ix_i] \\ E[x_is_i'] & k_1k_2E[x_i^2] \end{bmatrix}$$

The meat matrix B for the heteroskedastic case is:

$$B = E[\psi_i(\theta)\psi_i(\theta)'] = \begin{bmatrix} E[s_is_i'\xi_{1,i}^2] & E[s_ix_i\xi_{1,i}\xi_{2,i}] \\ E[x_is_i'\xi_{1,i}\xi_{2,i}] & E[x_i^2\xi_{2,i}^2] \end{bmatrix}$$

where $\xi_{1,i} = \varepsilon_i + \gamma(z_i - k_1x_i)$ and $\xi_{2,i} = \varepsilon_i + \gamma(z_i - k_1k_2x_i)$.

iii. Sample-Based Robust Estimator

Let $\hat{\xi}_{1,i} = y_i - s_i'\hat{\beta} - k_1\hat{\gamma}x_i$ and $\hat{\xi}_{2,i} = y_i - s_i'\hat{\beta} - k_1k_2\hat{\gamma}x_i$. The empirical sandwich estimator for the variance is:

$$\widehat{\text{Var}}(\hat{\theta}) = \frac{1}{n}\hat{A}^{-1}\hat{B}(\hat{A}')^{-1}$$

where $\hat{A}$ is the design matrix and $\hat{B} = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}_i\hat{\psi}_i'$ using residuals $\hat{\xi}_{j,i}$.

□

A.3 Proof of Theorem 3

Proof. We prove the consistency and asymptotic normality of $(\hat{\beta}, \hat{\gamma})$ when (p, q) is unknown and consistently estimated as $(\hat{p}, \hat{q})$.

i. Consistency

Let $\theta = (\beta', \gamma)'$ and $\hat{\phi} = (\hat{p}, \hat{q}, \hat{\boldsymbol{\pi}})'$ be the first-step MLE estimates from the Hausman et al. (1998) procedure. Since $\hat{\phi} \xrightarrow{p} \phi_0$ and the moment function $\psi_i(\theta, \phi) = \begin{pmatrix} s_i(y_i - s_i'\beta - k_1\gamma x_i) \\ x_i(y_i - s_i'\beta - k_1 k_2 \gamma x_i) \end{pmatrix}$ is continuous in ϕ, by Slutsky's Theorem:

$$\frac{1}{n}\sum_{i=1}^{n} \psi_i(\theta, \hat{\phi}) \xrightarrow{p} E[\psi_i(\theta, \phi_0)]$$

As established in the case of known (p, q), the unique solution to $E[\psi_i(\theta, \phi_0)] = 0$ is θ_0. Thus, $\hat{\theta} \xrightarrow{p} \theta_0$.

ii. Asymptotic Normality

Let $\nabla_\phi \ell_i(\phi)$ be the gradient of the first-step estimation problem. The joint estimator satisfies:

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \nabla_\phi \ell_i(\hat{\phi}) \\ \psi_i(\hat{\theta}, \hat{\phi}) \end{pmatrix} = 0$$

A first-order Taylor expansion of $\frac{1}{n}\sum \psi_i(\hat{\theta}, \hat{\phi})$ around (θ_0, ϕ_0) yields:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -A^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_i(\theta_0, \phi_0) + F\sqrt{n}(\hat{\phi} - \phi_0)\right] + o_p(1)$$

where $A = E[\nabla_\theta \psi_i(\theta_0, \phi_0)]$ and $F = E[\nabla_\phi \psi_i(\theta_0, \phi_0)]$ from standard MLE theory. Substituting the influence function of the MLE, $\sqrt{n}(\hat{\phi} - \phi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L_i + o_p(1)$, where $L_i = -E\left[\nabla_{\phi\phi} \ell_i(\phi_0)\right]^{-1} \nabla_\phi \ell_i(\phi_0)$, we obtain:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathbf{V}^*),$$

where $\mathbf{V}^* = A^{-1}B^*(A')^{-1}$. The corrected variance matrix is $B^* = \mathrm{Var}(\psi_i + FL_i)$.

iii. Sample-Based Robust Estimator

To implement the estimator with n observations, let $\hat{\xi}_{1,i} = y_i - s_i'\hat{\beta} - k_1\hat{\gamma}x_i$ and $\hat{\xi}_{2,i} = y_i - s_i'\hat{\beta} - k_1 k_2 \hat{\gamma} x_i$ the residuals. The empirical sandwich components are:

- **Jacobian Matrix:** $\hat{A} = -\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} s_i s_i' & \hat{k}_1 s_i x_i \\ x_i s_i' & \hat{k}_1 \hat{k}_2 x_i^2 \end{bmatrix}$

- **Adjustment Matrix:** $\hat{F} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} -\hat{\gamma}\frac{\partial \hat{k}_1}{\partial p} s_i x_i & -\hat{\gamma}\frac{\partial \hat{k}_1}{\partial q} s_i x_i \\ -\hat{\gamma}\frac{\partial (\hat{k}_1 \hat{k}_2)}{\partial p} x_i^2 & -\hat{\gamma}\frac{\partial (\hat{k}_1 \hat{k}_2)}{\partial q} x_i^2 \end{bmatrix}$

- **Total Influence:** $\hat{m}_i = \hat{\psi}_i + \hat{F}\hat{L}_i$, where $\hat{L}_i$ is the estimated influence function of the first-step parameters.

The robust covariance matrix for $\hat{\theta}$ is then $\widehat{V}_\theta = \frac{1}{n}\hat{A}^{-1}\hat{B}^*(\hat{A}')^{-1}$, where $\hat{B}^* = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i \hat{m}_i'$ using residuals $\hat{\xi}_{j,i}$. □

A.4 Proof of Theorem 4

Proof. We establish the consistency and asymptotic normality of the joint GMM estimator $\hat{\eta}$, which estimates the structural parameters $\theta = (\beta', \gamma)'$ and the misclassification model parameters $\phi = (p, q, \boldsymbol{\pi})'$ simultaneously. Let $\eta = (\theta', \phi')' = (\beta', \gamma, p, q, \boldsymbol{\pi})'$ be the full vector of parameters. The estimator is based on a stacked vector of moment conditions $g_i(\eta)$: $g_i(\eta) = \begin{pmatrix} \nabla_\phi \ell_i(\phi) \\ \psi_i(\phi, \theta) \end{pmatrix}$, where:

- $\nabla_\phi \ell_i(\phi)$ is the score vector from the misclassification likelihood in Hausman et al. (1998), ensuring the identification of ϕ.

- $\psi_i(\phi, \theta)$ represents the structural regression moments as defined earlier above.

The GMM estimator minimizes the objective function:

$$Q_n(\eta) = \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\eta)\right)' \mathbf{W}_n \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\eta)\right)$$

where $\mathbf{W}_n$ is a positive semi-definite weighting matrix.

i. Consistency

We assume the set of parameters is a compact set.

Identification: The parameter ϕ is identified by the non-linear functional form of the likelihood moments $E[\nabla_\phi \ell_i(\phi)] = 0$ (Hausman et al. 1998). Given identified ϕ, the structural parameters θ are identified by the moments condition $E[\psi_i(\phi, \theta)] = 0$, provided the rank condition holds for the set $[s_i \quad x_i]$.

Uniform Convergence: By the Uniform Weak Law of Large Numbers (UWLLN), the sample objective function $Q_n(\eta)$ converges uniformly to the population objective function $Q_0(\eta) = E[g_i(\eta)]'\mathbf{W}E[g_i(\eta)]$, where $\mathbf{W}$ is the probability limit of $\mathbf{W}_n$.

Since $Q_0(\eta)$ is uniquely minimized at η_0 (due to identification) and $\hat{\eta}$ minimizes $Q_n(\eta)$, it follows that $\hat{\eta} \xrightarrow{p} \eta_0$.

ii. Asymptotic Normality

Because the model is exactly identified the sample estimator satisfies the system of equations exactly:

$$\bar{g}_n(\hat{\eta}) = \mathbf{0}.$$

Taking a first-order Taylor expansion of $\bar{g}_n(\hat{\eta})$ around the true value η_0 yields:

$$\mathbf{0} = \bar{g}_n(\eta_0) + G_n(\bar{\eta})(\hat{\eta} - \eta_0),$$

where $G_n(\bar{\eta}) = \frac{\partial \bar{g}_n(\eta)}{\partial \eta'}\Big|_{\eta=\bar{\eta}}$ is the sample Jacobian matrix evaluated at an intermediate point $\bar{\eta}$ between $\hat{\eta}$ and η_0. Assuming $G_n(\bar{\eta})$ is non-singular and invertible as $n \to \infty$, solving for $\sqrt{n}(\hat{\eta} - \theta_0)$ yields:

$$\sqrt{n}(\hat{\eta} - \eta_0) = -\left[G_n(\bar{\eta})\right]^{-1} \sqrt{n}\bar{g}_n(\eta_0).$$

By the Central Limit Theorem, $\sqrt{n}\bar{g}_n(\eta_0) \xrightarrow{d} N(0, \Omega)$, where $\Omega = E[g_i(\eta_0)g_i(\eta_0)']$. By the Uniform Law of Large Numbers, $G_n \xrightarrow{p} G = E[\nabla_\eta g_i(\eta_0)]$. Thus, the asymptotic variance is $\mathbf{V} = (G'\Omega^{-1}G)^{-1}$. This achieves the semi-parametric efficiency bound for the joint estimation problem. The efficient variance of $\hat{\theta}$ is then obtained as a sub-matrix of $\mathbf{V}$.

iii. Sample-Based Variance Estimator

To conduct inference in finite samples, we replace the population moments in the asymptotic variance formula with their sample analogs. Let $\hat{\eta}$ be the estimated parameter vector. The sample-based robust variance estimator for $\hat{\eta}$ is given by:

$$\hat{\mathbf{V}}_n = (\hat{G}_n'\hat{\Omega}_n^{-1}\hat{G}_n)^{-1}$$

where the components are defined as follows:

- **Jacobian Matrix ($\hat{G}_n$):** The matrix of sample derivatives of the moment conditions with respect to the parameters:

$$\hat{G}_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_\eta g_i(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} 0 & \nabla^2_{\phi\phi}\ell_i(\hat{\phi}) \\ \nabla_\theta \psi_i(\hat{\theta}, \hat{\phi}) & \nabla_\phi \psi_i(\hat{\theta}, \hat{\phi}) \end{pmatrix}$$

- **Covariance of Moments ($\hat{\Omega}_n$):** The heteroskedasticity-robust estimator of the moment covariance:

$$\hat{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\eta})g_i(\hat{\eta})'$$

where $g_i(\hat{\eta})$ is the vector of stacked moments conditions (scores and structural moments) for observation i evaluated at $\hat{\eta}$.

Standard errors for the coefficients are obtained by taking the square root of the diagonal elements of $\frac{1}{n}\hat{\mathbf{V}}_n$. Note that information flows between the misclassification model and the regression model is captured by the off-diagonal elements of $\hat{\Omega}_n$, providing more precision for all estimates.

□

B Proofs of Corollaries 1 and 2

Proof. Here we prove the consistency and asymptotic normality of the intercept for all cases. The proof uses the same argument for all estimators discussed earlier, except for minor considerations that we explain below. Given the structural model $Y_i = \alpha + S_i'\beta + \gamma Z_i + \varepsilon_i$, the estimator for the intercept is defined as:

$$\hat{\alpha} = \bar{Y} - \bar{S}'\hat{\beta} - \hat{\gamma}\frac{\bar{X} - \hat{p}}{1 - \hat{p} - \hat{q}}$$

i. Consistency of intercept

The estimator $\hat{\alpha}$ can be defined using the continuous function $g(\cdot)$ such that:

$$\hat{\alpha} = g(\bar{Y}, \bar{S}, \bar{X}, \hat{\beta}, \hat{\gamma}, \hat{p}, \hat{q}) = \bar{Y} - \bar{S}'\hat{\beta} - \hat{\gamma}\left(\frac{\bar{X} - \hat{p}}{1 - \hat{p} - \hat{q}}\right)$$

We know, by the consistency of the slope and misclassification estimators that:

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\gamma} \xrightarrow{p} \gamma, \quad \hat{p} \xrightarrow{p} p, \quad \hat{q} \xrightarrow{p} q$$

Since (Y_i, S_i, X_i) are i.i.d. and satisfy the conditions for the Weak Law of Large Numbers, then by Slutsky's Theorem and the Continuous Mapping Theorem:

$$\operatorname{plim}\hat{\alpha} = E[Y] - E[S]'\beta - \gamma\left(\frac{E[X] - p}{1 - p - q}\right) = E[Y] - E[S]'\beta - \gamma E[Z] = \alpha \quad (15)$$

Thus, $\hat{\alpha}$ is a consistent estimator of the true intercept α. This result also holds when (p, q) is known, as in this case we can simply see it as $\hat{p} = p$ and $\hat{p} = p$.

ii. Asymptotic Normality of the intercept

To derive the asymptotic distribution of the intercept, we treat α as a component of the joint parameter vector $\eta = (\alpha, \beta', \gamma, p, q)'$. We follow the Delta Method approach by defining the intercept as a function of the slope parameters, nuisance parameters, and sample moments.

Let $\eta = (\beta', \gamma, p, q)'$ denote the vector of slope and misclassification parameters, and let $\mu = (E[Y], E[S]', E[X])'$ be the vector of population moments. The estimator for the intercept can be expressed as a differentiable function $g(\eta, \mu)$:

$$\hat{\alpha} = g(\hat{\eta}, \bar{\mu}) = \bar{Y} - \bar{S}'\hat{\beta} - \hat{\gamma}\left(\frac{\bar{X} - \hat{p}}{1 - \hat{p} - \hat{q}}\right) \tag{16}$$

Given the consistency and asymptotic normality of the Two-step and GMM estimators, we have:

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} N(0, \mathbf{V}_\eta) \tag{17}$$

$$\sqrt{n}(\bar{\mu} - \mu) \xrightarrow{d} N(0, \mathbf{V}_\mu) \tag{18}$$

By the Delta Method, the intercept $\hat{\alpha}$ also follows an asymptotic normal distribution:

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \nabla g' \mathbf{\Sigma} \nabla g) \tag{19}$$

where ∇g is the gradient of the intercept function $g(\cdot)$ with respect to all parameters and moments evaluated at their true values:

$$\nabla g = \left[\frac{\partial g}{\partial \eta'}, \frac{\partial g}{\partial \mu'}\right]' \tag{20}$$

and $\mathbf{\Sigma}$ represents the joint asymptotic covariance matrix of $(\hat{\eta}, \bar{\mu})$. When (p, q) is known, $\theta = (\beta', \gamma)$ is used in place of $\eta = (\beta', \gamma, p, q)'$ and the same reasoning applies. □

Data availability statement

The empirical example in this study is based on the 2005 Survey of Employment and the Informal Sector (EESI) for Cameroon. These data are publicly available on request to the Cameroon National Institute of Statistics (https://ins-cameroun.cm).

References

Abrevaya, J. & Hausman, J. (1999), 'Semiparametric estimation with mismeasured dependent variables: An application to panel data on employment spells', *Annales D'Economie et de Statistique* **55**(56), 243–75.

Adom, I. M. (2025), 'A structural analysis of firms' transition from informality to formality in nigeria', *Journal of Development Economics* p. 103684.

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

Alderman, H., Behrman, J. R., Lavy, V. & Menon, R. (2001), 'Child health and school enrollment: A longitudinal analysis', *Journal of Human resources* pp. 185–205.

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Battistin, E., Nadai, M. & Sianesi, B. (2014), 'Misreported schooling, multiple measures and returns to educational qualifications', *Journal of Econometrics* **181**(2), 136–150.

Benjamin, N. C. & Mbaye, A. A. (2012), 'The informal sector, productivity, and enforcement in w est a frica: A firm-level analysis', *Review of Development Economics* **16**(4), 664–680.

Black, D. A., Berger, M. C. & Scott, F. A. (2000), 'Bounding parameter estimates with nonclassical measurement error', *Journal of the American Statistical Association* **95**(451), 739–748.

Black, D., Sanders, S. & Taylor, L. (2003), 'Measurement of higher education in the census and current population survey', *Journal of the American Statistical Association* **98**(463), 545–554.

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

Bollinger, C. R. (2001), 'Measurement error and the union wage differential.', *Southern Economic Journal* **68**(1), 60–76.

Bollinger, C. R. (2003), 'Measurement error in human capital and the black-white wage differential.', *Review of Economics and Statistics* **85**(3), 578–585.

Bollinger, C. R. & David, M. H. (1997), 'Modeling discrete choice with response error: Food stamp participation', *Journal of the American Statistical Association* **92**(439), 827–835.

Bollinger, C. R. & Minier, J. (2014), 'On the robustness of coefficient estimates to the inclusion of proxy variables.', *Journal of Econometric Methods* **4**(1), 101–122.

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

Bound, J., Brown, C. & Mathiowetz, N. (2001), Measurement error in survey data, *in* 'Handbook of econometrics', Vol. 5, Elsevier, pp. 3705–3843.

Brown, B. W. & Newey, W. K. (2002), 'Generalized method of moments, efficient bootstrapping, and improved inference', *Journal of Business & Economic Statistics* **20**(4), 507–517.

Card, D. (1996), 'The effect of unions on the structure of wages: A longitudinal analysis', *Econometrica* **64**(4), 957–979.

Cavanagh, C. & Sherman, R. P. (1998), 'Rank estimators for monotonic index models', *Journal of Econometrics* **84**, 351–381.

Chen, X., Hong, H. & Nekipelov, D. (2011), 'Nonlinear models of measurement errors', *Journal of Economic Literature* **49**(4), 901–937.

Courtemanche, C., Denteh, A. & Tchernis, R. (2019), 'Estimating the associations between snap and food insecurity, obesity, and food purchases with imperfect administrative measures of participation', *Southern Economic Journal* .

De Luca, G., Magnus, J. R. & Peracchi, F. (2018), 'Balanced variable addition in linear models', *Journal of Economic Surveys* **32**(4), 1183–1200.

Deaton, A. (2019), *The analysis of household surveys: A microeconometric approach to development policy*, World Bank Publications.

DiTraglia, F. J. & García-Jimeno, C. (2019), 'Identifying the effect of a misclassified, binary, endogenous regressor', *Journal of Econometrics* **209**, 376–390.

Dovonon, P. & Gonçalves, S. (2017), 'Bootstrapping the gmm overidentification test under first-order underidentification', *Journal of Econometrics* **201**(1), 43–71.

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

Freeman, R. B. (1984), 'Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables', *Journal of Labor Economics* **1**(2), 1–26.

Frost, P. A. (1979), 'Proxy variables and specification bias', *The review of economics and Statistics* pp. 323–325.

Gonçalves, S., Hounyo, U., Patton, A. J. & Sheppard, K. (2023), 'Bootstrapping two-stage quasi-maximum likelihood estimators of time series models', *Journal of Business & Economic Statistics* **41**(3), 683–694.

Greenland, S. (1980), 'The effect of misclassification in the presence of covariates', *American journal of epidemiology* **112**(4), 564–569.

Gregory, C. A. & van Hasselt, M. (2025), 'Bayesian learning in the presence of misreporting and endogeneity: The case of snap participation and diet quality', *Journal of Human Resources* .

Griliches, Z. (1986), 'Economic data issues', *Handbook of econometrics* **3**, 1465–1514.

Hahn, J. (1996), 'A note on bootstrapping generalized method of moments estimators', *Econometric Theory* **12**(1), 187–197.

Haider, S. J. & Stephens Jr, M. (2025), 'Correcting for misclassified binary regressors using instrumental variables', *Journal of Business & Economic Statistics* **43**(3), 592–602.

Han, A. (1987), 'Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator', *Journal of Econometrics* **35**(2-3), 303–316.

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

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

Hu, Y. & Shennach, S. (2008), 'Instrumental variable treatment of nonclassical measurement error models', *Econometrica* **76**, 195–216.

INS-Cameroon (2005), *Enquête sur L'emploi et le secteur informel au Cameroun en 2005*, Rapport principal, Yaoundé, Cameroon.

Kane, T. J. & Rouse, C. E. (1995), 'Labor-market returns to two-and four-year college', *The American Economic Review* **85**(3), 600–614.

Kane, T. J., Rouse, C. E. & Staiger, D. (1999), Estimating returns to schooling when schooling is misreported, Working Paper 7235, National Bureau of Economic Research.

Kasahara, H. & Shimotsu, K. (2022), 'Identification of regression models with a misclassified and endogenous binary regressor', *Econometric Theory* **38**(6), 1117–1139.

Kede Ndouna, F. & Tsafack Nanfosso, R. (2023), 'Effet de l'inclusion financière sur la formalisation des petites et moyennes entreprises au cameroun', *Journal of Small Business & Entrepreneurship* **35**(1), 56–85.

Kreider, B. (2010), 'Regression coefficient identification decay in the presence of infrequent classification errors', *The Review of Economics and Statistics* **92**(4), 1017–1023.

Kreider, B. & Pepper, J. (2007), 'Disability and employment: Reevaluating the evidence in light of reporting errors', *Journal of the American Statistical Association* **102**(478), 432–441.

Kreider, B., Pepper, J. V., Gundersen, C. & Jolliffe, D. (2012), 'Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported', *Journal of the American Statistical Association* **107**(499), 958–975.

Lamarche, C. (2025), 'Quantile regression with a one-sided misclassified binary regressor', *The Annals of Applied Statistics* **19**(3), 2539–2554.

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

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

Manski, C. (1985), 'Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator', *Journal of Econometrics* **27**, 313–334.

McCallum, B. T. (1972), 'Relative asymptotic bias from errors of omission and measurement.', *Econometrica* **40**(4), 757–58.

Meyer, B. D. & Mittag, N. (2017), 'Misclassification in binary choice models', *Journal of Econometrics* **200**(2), 295–311.

Meyer, B. D., Mittag, N. & Goerge, R. M. (2022), 'Errors in survey reporting and imputation and their effects on estimates of food stamp program participation', *Journal of Human Resources* **57**(5), 1605–1644.

Newey, W. K. & McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of econometrics* **4**, 2111–2245.

Nguimkeu, P. (2014), 'A structural econometric analysis of the informal sector heterogeneity', *Journal of Development Economics* **107**, 175–191.

Nguimkeu, P. (2022), 'A structural model of informality with constrained entrepreneurship', *Economic Development and Cultural Change* **70**(3), 941–980.

Nguimkeu, P., Denteh, A. & Tchernis, R. (2019), 'On the estimation of treatment effects with endogenous misreporting', *Journal of Econometrics* **208**(2), 487–506.

Nguimkeu, P. & Okou, C. (2021), 'Leveraging digital technologies to boost productivity in the informal sector in Sub-Saharan Africa', *Review of Policy Research* **38**(6), 707–731.

Nguimkeu, P., Rosenman, R. & Tennekoon, V. (2021), 'Regression with a misclassified binary regressor: Correcting for the hidden bias'.

Ogburn, E. L. & VanderWeele, T. J. (2012), 'On the nondifferential misclassification of a binary confounder', *Epidemiology* **23**(3), 433.

Ogburn, E. L. & Vanderweele, T. J. (2013), 'Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders', *Biometrika* **100**(1), 241–248.

Pei, Z., Pischke, J.-S. & Schwandt, H. (2019), 'Poorly measured confounders are more useful on the left than on the right', *Journal of Business & Economic Statistics* **37**(2), 205–216.

Poterba, J. M. & Summers, L. H. (1995), 'Unemployment benefits and labor market transitions: A multinomial logit model with errors in classification', *Review of Economics and Statistics* **77**(2), 207–216.

Savoca, E. (2000), 'Measurement errors in binary regressors: an application to measuring the effects of specific psychiatric diseases on earnings', *Health Services and Outcomes Research Methodology* **1**(2), 149–164.

Schennach, S. M. (2016), 'Recent advances in the measurement error literature', *Annual review of economics* **8**(1), 341–377.

Sjoelander, A., Peña, J. M. & Gabriel, E. E. (2022), 'Bias results for nondifferential mismeasurement of a binary confounder', *Statistics & Probability Letters* **186**, 109474.

Tommasi, D. & Zhang, L. (2024), 'Bounding program benefits when participation is misreported', *Journal of Econometrics* **238**(1), 105556.

Ura, T. (2018), 'Heterogeneous treatment effects with mismeasured endogenous treatment', *Quantitative Economics* **9**(3), 1335–1370.

van Hasselt, M. & Bollinger, C. (2012), 'Binary misclassification and identification in regression models', *Economics Letters* **115**, 81–84.

Wickens, M. R. (1972), 'A note on the use of proxy variables', *Econometrica: Journal of the Econometric Society* pp. 759–761.

Wolpin, K. (2013), *Empirical Methods for the Study of Labour Force Dynamics*, Routledge.

Wossen, T., Abdoulaye, T., Alene, A., Nguimkeu, P., Feleke, S., Rabbi, I. Y., Haile, M. G. & Manyong, V. (2019), 'Estimating the productivity impacts of technology adoption in the presence of misclassification', *American Journal Agricultural Economics* **101**(1), 1–16.

Yanagi, T. (2019), 'Inference on local average treatment effects for misclassified treatment', *Econometric Reviews* **38**(8), 938–960.