# Corrupt Binary Confounders*

Pierre Nguimkeu§

Georgia State University

February 2026

**Abstract**

Researchers facing mismeasured confounders must often choose between omitting the variable (risking omitted variable bias) or including a mismeasured proxy (risking measurement error bias). While conventional econometric wisdom suggests that including a proxy reduces bias, the recent non-classical measurement error literature demonstrates that including a bad proxy may exacerbate bias relative to exclusion. We show that this "fear of the proxy" is unfounded when the confounder is binary. We prove that while binary misclassification is inherently non-classical, including a corrupt binary proxy strictly reduces the bias in the coefficient of interest compared to omitting the confounder entirely. To correct the bias in the naive linear model, we propose a sequential two-step estimator and a joint GMM estimator, which are consistent and asymptotically normal. These corrections do not require external instruments and accommodate both cases of known or unknown misclassification probabilities. Our Monte Carlo simulations quantify the bias of including a corrupt binary confounder and demonstrate the finite-sample performance of our correction methods. We apply the proposed methods to estimate returns to education among household business owners in Cameroon.

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

*I thank Ismael Mourifié and Desiré Kédagni for helpful comments.

§ Department of Economics, Georgia State University, 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 measured with error, 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. Recent work by Pei, Pischke & Schwandt (2019), reinforcing earlier results 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), Wolpin (2013), Bound et al. (2001), Bollinger (2003), 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 develop two practical, fully consistent estimators that correct the residual bias left by the naive OLS model when misclassification probabilities are unknown to the researcher: (i)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, (ii) and an efficient joint GMM Estimator that combines them. We establish the consistency and asymptotic normality of our proposed estimators and show that they are fully flexible to also incorporate the unusual case when 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 estimators, confirming the dominance of our correction methods. 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 et al. (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 demonstrate scenarios where the correlation structure of the continuous error term can destabilize OLS and cause the naive estimator to be 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 (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 (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 which 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 and consistent estimation of the main coefficient of interest, without requiring external instruments or repeated measures.

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 variables added to the model, including continuous and/or discrete regressors. To simplify the exposition, we ignore them in our discussion, given that they can be "partialled out" through orthogonal projection without changing our reasoning. The error term $\varepsilon_i$ is assumed to have mean zero and is not correlated with $s_i$ and The main parameter of interest for the researcher is $\beta$.

We assume that the researcher can only observe $y_i$, the dependent variable, $s_i$ the main regressor of interest, and $x_i$ a corrupt 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, as is common in the 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

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

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

the probability of reporting $x_i = 0$ 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 $z_i$. This condition is without loss of generality because if $p + q > 1$, the reasoning still hold by using $1 - x_i$ as the proxy for $z_i$ rather than $x_i$. If $p + q = 1$, the proxy is pure noise and is not of interest. This would mean $\text{Cov}(x_i, z_i) = 0$ and is the only situation we rule out. We focus our discussion on "population' relationships, hence all the sample statistics are discussed in terms of their asymptotic limits.

If the researcher decides to ignore the confounder $z_i$ and drops it from the model, i.e. they estimate the following (omitted variable) model:

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

where $\varepsilon_i^o = \varepsilon_i + \gamma z_i$, then the estimation of $\beta$ using Equation (4) would lead to an omitted variable bias. But if the researcher has a candidate surrogate $x_i$ for $z_i$, then they may choose to run an operational (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 $\beta$ using Equation (5) would lead to a corrupt proxy bias.

In the next section, we discuss the coefficient $\beta$ from estimating these two models, and we compare the biases of including the corrupt confounder as opposed to dropping it. To simplify the exposition of this comparison, we first treat $s_i$ as a scalar, but our correction estimators are discussed for the general case where $s_i$ could be a vector of variables. We assume throughout that the measurement error is exogenous to the outcome equation, that is, $\text{Cov}(\varepsilon, u) = 0$.

## 2.1 Bias from Omission and Bias from Inclusion

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

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

where $\text{Var}(\cdot)$ and $\text{Cov}(\cdot, \cdot)$ denote the population variance and covariance, respectively, and $\delta = \text{Var}(s_i)^{-1}\text{Cov}(s_i, z_i)$ is the population slope coefficient in the regression of $z_i$ on $s_i$. The size of the omitted variable bias is given by $\text{Bias}(\hat{\beta}^o) = \gamma\delta$ and

depends on both the coefficient $\gamma$ in the true model (1) as well as the coefficient $\delta$ in the (infeasible) regression of the omitted confounder $z_i$ on the main regressor of interest $s_i$.

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

$$\text{plim}\,\hat{\beta}^c = \beta + \gamma \text{Var}(s|x)^{-1}\text{Cov}(s, z|x). \tag{7}$$

where $s|x$ and $z|x$ are the variables $s$ and $z$ after removing the components linearly explained by $x$. That is, $s|x$ and $z|x$ are the residuals of the linear regressions of $s$ and $z$ on $x$, respectively. This conditional variance can be expressed in terms of unconditional variance as follows.

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

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

$$\text{Cov}(s, z|x) = \text{Cov}(s, z) - \frac{\text{Cov}(s, x)\text{Cov}(x, z)}{\text{Var}(x)}.$$

The expression of the bias in the estimation of $\beta$ 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 $z$ is binary and the measurement error in $x$ is a misclassification in $z$ (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 $\text{Cov}(s, z|x)$ in terms of the observables $x$ and $s$ and the misclassification probabilities. As shown by previous authors (e.g., Aigner 1973, Frazis & Loewenstein 2003) we can write:

$\text{Cov}(s, z) = k_1\text{Cov}(s, x)$ and $\text{Cov}(x, z) = k_1 k_2 \text{Var}(x)$, 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, $\text{Cov}(s, z|x) = (1 - k_2)\text{Cov}(s, z)$, and it follows that

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

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 $0 \leq q < 1 - p_x$, 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 $k_2 = (1-p-q)^2\dfrac{\text{Var}(z)}{\text{Var}(x)}$, 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 $\lambda = \dfrac{1-k_2}{1-R^2_{s,x}}$ 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 summarize it in the following proposition.

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

This results comes from the fact that, as shown in the appendix, we must always have $\lambda \leq 1$ 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 in general be higher than the 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 $\text{Bias}(\hat{\beta}^c) = \text{Bias}(\hat{\beta}^o)$.

# 3 Correction Estimators

We present solutions for the potentially costly inclusion of a corrupt confounder in the regression model. The proposed solutions either take misclassification probabilities as given (e.g. through existing validation studies), or estimate them through a misclassification model assuming knowledge of the distribution and/or predictors for the true confounder. For the latter, we discuss a two-step estimator that estimates misclassification probabilities as a first step, and a GMM estimator that jointly estimates misclassification probabilities along with model parameters. For all these procedures, the covariances between the misclassification error and the model variables obtained earlier in terms of sample statistics are what we use to correct for the bias within the naive least squares normal equations.

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:

$$\begin{bmatrix} \text{Var}(s_i) & k_1 \text{Cov}(s_i, x_i) \\ \text{Cov}(x_i, s_i) & k_1 k_2 \text{Var}(x_i) \end{bmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{bmatrix} \text{Cov}(s_i, y_i) \\ \text{Cov}(x_i, y_i) \end{bmatrix} \tag{10}$$

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 $\beta$ and $\gamma$. 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) simplify to the normal equations of the OLS estimators $\hat{\beta}^c$ and $\hat{\gamma}^c$ in the operational model given by Equation (5).

## 3.1 Estimation with Known Misclassification Rates

Given knowledge of $(p, q)$, the factors $k_1$ and $k_2$ are known and Equation (10) fully identifies $\gamma$ and $\beta$. We can therefore define consistent estimators for $\gamma$ and $\beta$ 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* $\beta$ *and* $\gamma$ *i.e.,* $\operatorname{plim} \hat{\beta} = \beta$ *and* $\operatorname{plim} \hat{\gamma} = \gamma$.
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., $p = q = 0$, the proposed 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, the proposed consistent estimator can be obtained from the naive OLS estimator in probability limits as follows:

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

This means that when misclassification probabilities are known, consistent estimators for $\gamma$ and $\beta$ 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 $\alpha$, its consistent estimation can readily obtain from the consistent estimation of $\beta$ and $\gamma$ as follows:

**Corollary 1.** *Under the conditions of Theorem 2, define the estimator of $\alpha$ 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 $v_\alpha$ 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 data, if any exist (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* $\gamma$ *and* $\beta$*, 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 or nonlinear least squares 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

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

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 $\alpha$ 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 $v_\alpha^*$ is its asymptotic variance.*

*Proof.* Appendix. □

While the asymptotic variance and its sample-based analogue of the Two-Step estimator of $(\alpha, \beta', \gamma)$ are given in the appendix, many researchers sometimes prefer using bootstrapping for these types of two-step 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 $(\beta, \gamma)$, we potentially sacrifice asymptotic efficiency. A joint GMM approach would stack the MLE scores and the structural moments into a single vector. By solving this system simultaneously with an optimal weighting matrix, 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.

To achieve the semi-parametric efficiency bound, we consider the joint estimation of $\eta = (\beta', \gamma, p, q)'$. We define the stacked moment vector $g_i(\eta)$:

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

where $\nabla_\phi \ell_i(\phi)$ is the score of the misclassification log-likelihood 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 the weighting matrix $\mathbf{W}_n$ is chosen such that $\mathbf{W}_n^{-1} \xrightarrow{p} \Omega = E[g_i(\eta_0) g_i(\eta_0)']$. We then have the following result.

**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 analogue 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 for the Two-Step estimator above, bootstrapping can also be employed to obtain standard errors for the GMM estimator of $(\alpha, \beta', \gamma, p, q)$ 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 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 explanatory variable of interest, is $s_i \sim \chi^2_{(1)}$, the true confounder is $z_i = \mathbf{1}[-1 + \pi s_i + v_i > 0]$, with $\pi \in \{-0.9, 0.9\}$, where $v_i \sim \mathcal{N}(0, 1)$. The parameter $\pi$ controls for the level of 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

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

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 consider a set of different combinations values of misclassification probabilities $(p, q)$, that is, $(0, 0.15)$; $(0, 0.30)$; $(0.15, 0)$; $(0.15, 0.15)$; $(0.15, 0.30)$; $(0.30, 0.15)$; $(0.30, 0.30)$, which allows us to evaluate the bias and performance of the naive 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, $(\beta, \gamma)$, are obtained for the infeasible OLS estimator using the correct confounder, the pessimistic OLS estimator that omits the confounder, the naive OLS estimator that includes the corrupt binary proxy, and the proposed correction estimators (the Two-step estimator that estimates misclassification probabilities with the Hausman et al. (1998) approach in a first step and the GMM estimator that jointly estimates all parameters).[1] The estimated probabilities of misclassification assume the correct specification of the distribution of the correct confounder.[2]

[1] We do not report the estimator that takes misclassification probabilities as given since it is a special case of the Two-step estimator with $\hat{p} = p$ and $\hat{q} = q$, and we are only showing bias.

[2] We 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.

## 4.2 Simulation Results

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 Infeasible OLS estimator (Column 6), which utilizes the true confounder $z_i$, as the ideal benchmark, and the OLS-2 (Omitted) estimator (Column 7) 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, the OLS-1 (Omitted) estimator consistently exhibits severe omitted variable biases in the estimate of the coefficient of interest, $\beta$, whose true value is set at -0.2. When $\pi = -0.9$, the $\hat{\beta}$ estimate from OLS-1 is near -0.322 and when $\pi = 0.9$, 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 the OLS-2 (Proxy) estimator (Column 8) is demonstrably closer to the true value than the OLS-1 (Omitted) estimate. It is only at the highest corruption case (e.g., $\pi = -0.9$, $p = 30\%$, $q = 30\%$), that the $\hat{\beta}$ estimate from OLS-1, $-0.323$, is close to the Proxy estimate $-0.316$. Although still biased, the inclusion of the corrupt proxy always achieves partial bias reduction. This pattern empirically validates our main theoretical result: $|\text{Bias}(\hat{\beta}_{\text{Include}})| < |\text{Bias}(\hat{\beta}_{\text{Omit}})|$. 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 estimator (OLS-2) is better than the omission estimator (OLS-1), it still systematically suffers from residual asymptotic bias. When $\pi = -0.9$, this bias increases with higher levels of corruption, and when $\pi = 0.9$, the bias is so severe that the estimator completely flips sign. This persistent bias is substantial and could lead to dramatic policy prescription, especially with sign-switching estimates. The coefficient of the mismeasured variable, $\gamma$, also exhibits severe attenuation bias. While the true value is $\gamma = 4.0$, the OLS-2 proxy estimator frequently collapses toward zero or is substantially understated (e.g., it takes the low values 0.502 or 1.201 in high corruption scenarios). This distortion of the confounder's coefficient is what contributes to drive the residual bias in the OLS-2 (Proxy) estimation of $\hat{\beta}$. These results clearly underscore that, while inclusion is better than omission, there are still substantial inconsistencies of the OLS-2 naive estimator, highlighting the necessity of our proposed corrections.

## Table 1: Simulations Results

| $\pi$ | $p$ | $q$ | Para | True Values | OLS-0 (Infeasible) | OLS-1 (Omitted) | OLS-2 (Proxy) | Correction Est. Two-Step | Correction Est. GMM |
|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| -0.9 | 0% | 0% | $\beta$ | -0.2 | -0.200 | -0.322 | -0.200 | -0.200 | -0.200 |
| | | | $\gamma$ | 4.0 | 3.996 | 0.000 | 3.996 | 3.996 | 3.996 |
| | | 15% | $\beta$ | -0.2 | -0.199 | -0.321 | -0.219 | -0.199 | -0.199 |
| | | | $\gamma$ | 4.0 | 3.998 | 0.000 | 3.932 | 3.998 | 3.999 |
| | | 30% | $\beta$ | -0.2 | -0.200 | -0.323 | -0.240 | -0.200 | -0.200 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 3.868 | 4.000 | 4.000 |
| | 15% | 0% | $\beta$ | -0.2 | -0.200 | -0.322 | -0.285 | -0.194 | -0.199 |
| | | | $\gamma$ | 4.0 | 3.998 | 0.000 | 1.412 | 4.139 | 4.005 |
| | | 15% | $\beta$ | -0.2 | -0.200 | -0.322 | -0.296 | -0.193 | -0.199 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 1.206 | 4.216 | 4.018 |
| | | 30% | $\beta$ | -0.2 | -0.200 | -0.322 | -0.305 | -0.182 | -0.198 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 0.993 | 4.528 | 4.029 |
| | 30% | 0% | $\beta$ | -0.2 | -0.199 | -0.322 | -0.303 | -0.170 | -0.199 |
| | | | $\gamma$ | 4.0 | 4.004 | 0.000 | 0.858 | 4.887 | 4.024 |
| | | 15% | $\beta$ | -0.2 | -0.200 | -0.322 | -0.311 | -0.179 | -0.198 |
| | | | $\gamma$ | 4.0 | 4.000 | 0.000 | 0.685 | 4.619 | 4.061 |
| | | 30% | $\beta$ | -0.2 | -0.200 | -0.323 | -0.316 | -0.182 | -0.190 |
| | | | $\gamma$ | 4.0 | 4.003 | 0.000 | 0.502 | 4.504 | 4.094 |
| 0.9 | 0% | 0% | $\beta$ | -0.2 | -0.201 | 0.500 | -0.201 | -0.201 | -0.201 |
| | | | $\gamma$ | 4.0 | 4.000 | 0.000 | 4.000 | 4.000 | 4.000 |
| | | 15% | $\beta$ | -0.2 | -0.200 | 0.502 | -0.001 | -0.203 | -0.202 |
| | | | $\gamma$ | 4.0 | 4.000 | 0.000 | 3.368 | 4.009 | 4.005 |
| | | 30% | $\beta$ | -0.2 | -0.199 | 0.501 | 0.145 | -0.202 | -0.201 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 2.907 | 4.012 | 4.009 |
| | 15% | 0% | $\beta$ | -0.2 | -0.201 | 0.503 | 0.058 | -0.203 | -0.201 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 2.976 | 4.008 | 4.001 |
| | | 15% | $\beta$ | -0.2 | -0.200 | 0.502 | 0.215 | -0.205 | -0.201 |
| | | | $\gamma$ | 4.0 | 3.999 | 0.000 | 2.337 | 4.016 | 4.003 |
| | | 30% | $\beta$ | -0.2 | -0.200 | 0.502 | 0.325 | -0.209 | -0.204 |
| | | | $\gamma$ | 4.0 | 4.002 | 0.000 | 1.839 | 4.037 | 4.017 |
| | 30% | 0% | $\beta$ | -0.2 | -0.201 | 0.503 | 0.211 | -0.205 | -0.202 |
| | | | $\gamma$ | 4.0 | 4.004 | 0.000 | 2.375 | 4.020 | 4.006 |
| | | 15% | $\beta$ | -0.2 | -0.199 | 0.503 | 0.337 | -0.210 | -0.203 |
| | | | $\gamma$ | 4.0 | 4.001 | 0.000 | 1.716 | 4.041 | 4.013 |
| | | 30% | $\beta$ | -0.2 | -0.200 | 0.503 | 0.419 | -0.223 | -0.209 |
| | | | $\gamma$ | 4.0 | 4.000 | 0.000 | 1.201 | 4.087 | 4.030 |

*Notes.* These are simulations results with 2000 replications and 1000 observations, where the misclassification probabililities are $(p,\ q) \in \{(0,0);\ (0,.15);\ (0,.3);\ (.15,0);\ (.15,.15);\ (.15,.3);\ (.3,0);\ (.3,.15);\ (.3,.3)\}$ and correlation between $z_i$ and $s_i$ is captured by $\pi \in \{-0.9;\ 0.9\}$. Column 5 shows the true coefficients. Column 6 reports the infeasible OLS estimation based on the true unobserved confounder. Column 7 reports OLS when counfounder is omitted. Column 8 reports OLS when the corrupt proxy is used. Column 9 and 10 report the correction estimators when $(p, q)$ is unknown and estimated using Hausman et al. (1998) as a first step (Two-Step estimator) and jointly estimated with the structural parameters (GMM estimator), respectively.

The performance of our proposed correction procedures is uniformly excellent, achieving the goal of eliminating asymptotic bias in the structural model's coefficients $(\beta, \gamma)$. When the misclassification probabilities $(p, q)$ are estimated using the Hausman et al. (1998) method as a first step, our Two-step correction estimator of $\beta$ (Column 9) performs nearly identically to the Infeasible OLS estimator (Column 6). Estimates for $\beta$ are consistently within tight tolerance of the true value (e.g., between -0.223 and -0.179), irrespective of the levels of corruption including extreme misclassification cases. Similarly, the estimation of $\gamma$ is accurately recovered within tight bounds of the true value (ranging between 3.996 and 4.506). This confirms the correctness of our derived theoretical bias formulas.

Similarly, the GMM correction estimator which estimates the parameters of the structural model $(\beta, \gamma)$ jointly with misclassfication probabilities using the Hausman et al. (1998) score and the structural moment conditions, demonstrates minimal loss of accuracy compared to the Two-step estimator. The estimates for $\beta$ remain extremely accurate (e.g., ranging from -0.209 to -0.190) irrespective of the level of corruption. This result confirms the practical feasibility of our method, showing that the model is point identified and consistently estimable even when misclassification rates are unknown to the researcher, as it would be common in practice. Notice, in particular, that this correction estimator maintains its high level of accuracy even in the most challenging scenarios (e.g., high correlation $\pi = 0.9$ combined with high misclassification $p = 30\%, q = 30\%$). This stability underscores the robustness and reliability of our approach.

In conclusion, these simulations support the two central tenets of our paper: (i) the naive approach using a corrupt binary proxy is a necessary first step that correctly minimizes bias compared to omitting the confounder entirely; and (ii) the proposed correction estimators are robust and effective solutions for achieving full consistency when some useful information about the misclassified confounder (correct misclassification rates or correct distribution function) can be assumed.

# 5 Returns to Education in Cameroon

This section illustrates the usefulness of the proposed methods in a real data example of returns to education and gender earnings gaps among household businesses in Cameroon.

## 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 section 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 under-reporting 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 2: 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** | | | |
| Formal (=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.

Table 2 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. While these differences are striking, they rely on self-reported formality status, which may be mismeasured. In settings with weak enforcement and ambiguous regulatory thresholds, businesses may under-report 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 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 3. Across all specifications, education is positively associated with income from household businesses, and is significant for all estimators except for the GMM. However, the magnitude of returns varies sharply across estimators. When formality is omitted, the estimated return to education is large (12.6), likely reflecting omitted variable bias, as education is correlated with formality. When self-reported formality is naively included, the return to education falls by almost half (6.4). Once we correct for misclassification in formality, the return to education declines further (4.6 for Two-step and 2.3 for GMM) and is only marginally significant for Two-step and insignificant for GMM. This pattern indicates that part of what is often interpreted as a high return to education among household businesses actually reflects selection into (true) formality, rather than productivity gains from schooling alone.

The age profile of income 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. Female-owned household businesses earn substantially less than male-owned businesses in all models. The raw gender gap is extremely large when formality is omitted (-52 000 CFA). Controlling for self-reported formality reduces the gap to -39 000 CFA. After correcting for misclassification, the gap remains large (–35 000 CFA for Two-Step and -31 000 CFA for GMM) and highly significant. This suggests that formalization alone cannot explain gender income disparities in household businesses. Even after accounting for true formality status, women face sizeable income penalties, pointing to deeper structural constraints such as differential access to capital, networks, or market opportunities. The estimated coefficient on formality is very large and statistically significant in both the proxy and corrected models, but increases further after correction (from 225 to 252 and 320). This confirms, as found in the literature, that formal firms are more productive and earn substantially more than informal firms (Nguimkeu 2022, Nguimkeu & Okou 2021, Adom 2025).

In general, the proposed correction estimators may increase standard errors, reflecting the uncertainty of the misclassification and preventing the false precision of biased naive estimates. Panel B provides direct evidence that self-reported formality is noisy: The estimated false negative rate ($q \approx 13.5\%$) is sizable and

statistically significant, implying that a non-trivial share of truly formal businesses report themselves as informal. The false positive rate is close to zero and insignificant, suggesting little over-reporting of formality. This asymmetric misclassification is consistent with institutional realities in Cameroon, where entrepreneurs may under-report formality due to weak enforcement, fear of taxation, or ambiguity about legal status.

**Table 3: Returns to Education and Gender Earnings Gaps in Cameroon**

| | Omitted OLS | Proxy OLS | Two-Step | GMM |
|---|---|---|---|---|
| **Panel A: Income Equation (in 1000 CFA)** | | | | |
| Education | 12.572*** | 6.399*** | 4.559* | 2.248 |
| | (1.680) | (1.287) | (2.508) | (47.05) |
| Age | 9.119*** | 6.172*** | 5.293*** | 4.868* |
| | (1.736) | (1.296) | (1.526) | (2.918) |
| $Age^2$ | -0.088*** | -0.066*** | -0.060*** | -0.060 |
| | (0.021) | (0.016) | (0.018) | (1.910) |
| Female | -52.146*** | -38.667*** | -34.647*** | -31.094*** |
| | (9.336) | (7.361) | (8.607) | (0.023) |
| Formal business | — | 225.483*** | 251.945*** | 319.945*** |
| | — | (30.617) | (71.028) | (9.483) |
| Intercept | -184.16*** | -87.292*** | -53.790 | -32.352 |
| | (39.787) | (28.426) | (40.491) | (93.749) |
| **Panel B: Misclassification Parameters (Formality Status)** | | | | |
| False positive rate $\hat{p}$ | | | 0.0046 | 0.0062 |
| | | | (0.0104) | (0.0109) |
| False negative rate $\hat{q}$ | | | 0.1348*** | 0.1343** |
| | | | (0.0436) | (0.0525) |

*Notes*: Dependent variable is income from household businesses in 1000 CFA. 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). The first stage and score functions use all the model covariates in addition to household initial wealth as predictors for true formality status, which is insignificant in the income equation. Bootstrap standard errors in parentheses. $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

# 6 Conclusion

This paper addresses a fundamental dilemma in empirical research: the choice between omitting a mismeasured confounder or including 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 the uncertainty from the first-stage misclassification probabilities into the final standard errors. Notably, these corrections do not require external instruments for the structural equation, making them highly applicable in data-constrained environments. Our Monte Carlo simulations confirm that these methods exhibit strong finite-sample performance, successfully eliminating the attenuation and contamination biases that plague naive linear models. The empirical application to household businesses in Cameroon highlights the importance of these corrections. We find that failing to account for misclassification in business formality leads to a significant overestimation of both the returns to education and the gender earnings gap, while simultaneously underestimating the gains from formalization. By applying our proposed correction, we reveal a more nuanced economic landscape: returns to education and gender earnings gaps are less pronounced among household business owners than naive models suggest. In contrast, the potential benefits of transitioning businesses from the informal to the formal sector are substantially larger.

In summary, this study provides both a theoretical safety net for the use of mismeasured binary proxies and a practical toolkit for their correction in regression analysis. For practitioners this means rather than reverting to the known hazards of omitted variable bias out of fear of misclassification error in binary counfounder, 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

In this section we prove all our theoretical results.

## 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 $\theta$ that satisfies $E[\psi_i(\theta)] = 0$. Substituting the true model into the moment conditions:

$$E[\psi_i(\theta)] = \begin{pmatrix} E[s_i(s_i\beta + \gamma z_i + \varepsilon_i - s_i\beta - k_1\gamma x_i)] \\ E[x_i(s_i\beta + \gamma z_i + \varepsilon_i - s_i\beta - k_1 k_2 \gamma x_i)] \end{pmatrix} = \begin{pmatrix} \gamma(E[s_i z_i] - k_1 E[s_i x_i]) + E[s_i\varepsilon_i] \\ \gamma(E[x_i z_i] - k_1 k_2 E[x_i^2]) + E[x_i\varepsilon_i] \end{pmatrix}$$

Given $\text{Cov}(s, z) = k_1\text{Cov}(s, x)$, $\text{Cov}(x, z) = k_1 k_2 \text{Var}(x)$, and $E[\varepsilon | s, z] = 0$, all terms vanish at the true $\theta$. 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:

$$A = E\left[\frac{\partial \psi_i}{\partial \theta'}\right] = -\begin{bmatrix} E[s_i s_i'] & k_1 E[s_i x_i] \\ E[x_i s_i'] & k_1 k_2 E[x_i^2] \end{bmatrix}$$

The meat matrix $B$ for the heteroskedastic case is:

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

where $\xi_{1,i} = \varepsilon_i + \gamma(z_i - k_1 x_i)$ and $\xi_{2,i} = \varepsilon_i + \gamma(z_i - k_1 k_2 x_i)$.

### iii. Sample-Based Robust Estimator

Let $\hat{\xi}_{1,i} = y_i - s_i'\hat{\beta} - k_1\hat{\gamma}x_i$ and $\hat{\xi}_{2,i} = y_i - s_i'\hat{\beta} - k_1 k_2\hat{\gamma}x_i$. The empirical sandwich estimator for the variance is:

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

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 $\hat{\phi} = (\hat{p}, \hat{q})'$ 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 $\phi$, 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 $\theta_0$. Thus, $\hat{\theta} \xrightarrow{p} \theta_0$.

## ii. Asymptotic Normality

Let $\nabla_\phi \ell_i(\phi)$ be the score of the first-step MLE. 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 $(\theta_0,\phi_0)$ 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 $\hat{\xi}_{2,i} = y_i - s_i'\hat{\beta} - k_1k_2\hat{\gamma}x_i$ the residuals. The empirical sandwich components are:

- **Jacobian Matrix:** $\hat{A} = -\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} s_is_i' & \hat{k}_1 s_i x_i \\ x_i s_i' & \hat{k}_1\hat{k}_2 x_i^2\end{bmatrix}$
- **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 One-Step GMM estimator $\hat{\eta}$, which estimates the structural parameters $\theta = (\beta', \gamma)'$ and the misclassification parameters $\phi = (p, q)'$ simultaneously. Let $\eta = (\theta', \phi')' = (\beta', \gamma, p, q)'$ 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 $\phi$.
- $\psi_i(\phi, \theta)$ 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 $\phi$ 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 $\phi$, the structural parameters $\theta$ 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 $\eta_0$ (due to identification) and $\hat{\eta}$ minimizes $Q_n(\eta)$, it follows that $\hat{\eta} \xrightarrow{p} \eta_0$.

### ii. Asymptotic Normality

By the first-order condition (FOC), $\nabla_\eta Q_n(\hat{\eta}) = 0$, i.e.,

$$G_n(\hat{\eta})' \mathbf{W}_n \bar{g}_n(\hat{\eta}) = 0$$

where $G_n(\eta) = \frac{1}{n}\sum \nabla_\eta g_i(\eta)$ and $\bar{g}_n(\eta) = \frac{1}{n}\sum g_i(\eta)$.

We expand the first-order condition around the true value $\eta_0$, and by the Mean Value Theorem, we can write

$$\bar{g}_n(\hat{\eta}) \approx \bar{g}_n(\eta_0) + G_n(\eta_0)(\hat{\eta} - \eta_0)$$

Substituting this into the FOC:

$$G_n' \mathbf{W}_n [\bar{g}_n(\eta_0) + G_n(\hat{\eta} - \eta_0)] \approx 0.$$

Solving for $\hat{\eta}$:

$$\sqrt{n}(\hat{\eta} - \eta_0) \approx -(G_n' \mathbf{W}_n G_n)^{-1} G_n' \mathbf{W}_n \sqrt{n}\bar{g}_n(\eta_0)$$

By the Central Limit Theorem (CLT), $\sqrt{n}\bar{g}_n(\eta_0) \xrightarrow{d} N(0, \Omega)$, where $\Omega = E[g_i(\eta_0)g_i(\eta_0)']$. By the Law of Large Numbers (LLN), $G_n \xrightarrow{p} G = E[\nabla_\eta g_i(\eta_0)]$ and $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$.

Thus, the asymptotic variance is:

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

**Asymptotic Efficiency:** If the optimal weighting matrix $\mathbf{W} = \Omega^{-1}$ is used (feasible via a two-step procedure), the variance simplifies to:

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

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 analogues. 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 $\alpha$. 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 $\alpha$ 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 $\nabla g$ 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. □

# Declaration of AI use

During the preparation of this manuscript, we used Gemini (Google) 1.5 Flash in order to refine the literature review and verify simulation code in R and MATLAB. After using this tool, we reviewed and edited the content as needed, and we take full responsibility for the content of this publication.

# 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.

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.

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.

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.

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. & 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).

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.

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'.

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.

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.