Identifying the Effect of a Mis-classified, Binary, Endogenous Regressor*

Francis J. DiTraglia[1] and Camilo García-Jimeno[2,3]

[1]Department of Economics, University of Pennsylvania
[2]Institute for Quantitative Theory and Methods, Emory University
[3]NBER

This Version: July 30, 2018, First Version: October 31, 2015

Abstract

This paper studies identification and inference for the effect of a mis-classified, binary, endogenous regressor when a discrete-valued instrumental variable is available. We begin by showing that the only existing point identification result for this model is incorrect. We go on to derive the sharp identified set under mean independence assumptions for the instrument and measurement error. The resulting bounds are novel and informative, but fail to point identify the effect of interest. This motivates us to consider alternative and slightly stronger assumptions: we show that adding second and third moment independence assumptions suffices to identify the model.

Keywords: Instrumental variables, Measurement error, Endogeneity

JEL Codes: C10, C25, C26

*We thank Daron Acemoglu, Manuel Arellano, Kristy Buzard, Xu Cheng, Bernardo da Silveira, Bo Honoré, Arthur Lewbel, Chuck Manski, Sophocles Mavroeidis, Francesca Molinari, Yuya Takahashi, the associate editor, two anonymous referees, and seminar participants at Cambridge, CEMFI, Chicago Booth, Manchester, Northwestern, Oxford, Penn State, Princeton, UCL, the 2016 Greater New York Area Econometrics Colloquium, Camp Econometrics IX, and the 2017 North American Summer Meeting of the Econometric Society for valuable comments and suggestions. This document supersedes an earlier version entitled "On Mis-measured Binary Regressors: New Results and Some Comments on the Literature."

1 Introduction

Measurement error and endogeneity are pervasive features of economic data. Conveniently, a valid instrumental variable corrects for both problems when the measurement error is classical, i.e. uncorrelated with the true value of the regressor. Many regressors of interest in applied work, however, are binary and thus cannot be subject to classical measurement error.[1] When faced with non-classical measurement error, the instrumental variables estimator can be severely biased. In this paper, we study an additively separable model of the form

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon \tag{1}$$

where ε is a mean-zero error term, T^* is a binary, potentially endogenous regressor of interest, and $\mathbf{x}$ is a vector of exogenous controls.[2] We ask whether, and if so under what conditions, a discrete instrumental variable z suffices to non-parametrically identify the causal effect $\beta(\mathbf{x})$ of T^*, when we observe not T^* but a mis-classified binary surrogate T.

We proceed under the assumption of non-differential measurement error. This condition has been widely used in the existing literature and imposes that T provides no additional information beyond that contained in $(T^*, \mathbf{x})$. Even in this fairly standard setting, identification remains an open question: we begin by showing that the only existing identification result for this model is incorrect. We then go on to derive the sharp identified set under the standard first-moment assumptions from the related literature. We show that regardless of the number of values that z takes on, the model is not point identified. This motivates us to consider alternative, and slightly stronger assumptions. We show that, given a binary instrument, the addition of a second moment independence assumption suffices to identify a model with one-sided mis-classification. Adding a second moment restriction on the measurement error along with a third moment independence assumption for the instrument suffices to identify the model in general. This result likewise requires only a binary z.

Our work relates to a large literature that considers departures from classical measurement error, by allowing the measurement error to be related to the true value of the unobserved regressor. Chen et al. (2005) obtain identification in a general class of moment condition models with mis-measured data by relying on the existence of an auxiliary dataset from which they can estimate the measurement error process. In contrast, Hu and Shennach (2008) and Song (2015) rely on an instrumental variable and an additional conditional location assumption on the measurement error distribution. More recently, Hu et al. (2015) use a continuous instrument to identify the ratio of partial effects of two continuous regressors, one measured with error, in a linear single index model. Unfortunately, these approaches cannot be applied to the case of a mis-measured binary regressor.

A number of papers have studied models with an exogenous binary regressor subject to non-differential measurement error. One group of papers asks what can be learned without recourse to an instrumental variable. An early contribution by Aigner (1973) characterizes the asymptotic bias of OLS in this setting, and proposes a correction using outside infor-

[1] The only way to mis-classify a true one is downwards, as a zero, while the only way to mis-classify a true zero is upwards, as a one. This creates negative dependence between the truth and measurement error.

[2] Because T^* is binary, there is no loss of generality from writing the model in this form rather than the more familiar $y = h(T^*, \mathbf{x}) + \varepsilon$. Simply define $\beta(\mathbf{x}) = h(1, \mathbf{x}) - h(0, \mathbf{x})$ and $c(\mathbf{x}) = h(0, \mathbf{x})$.

mation on the mis-classification process. Related work by Bollinger (1996) provides partial identification bounds. More recently, Chen et al. (2008a) use higher moment assumptions to obtain identification in a linear model, and Chen et al. (2008b) extend these results to the non-parametric setting. van Hasselt and Bollinger (2012) and Bollinger and van Hasselt (2015) provide additional partial identification results. For results on the partial identification of discrete probability distributions under mis-classification, see Molinari (2008).

Continuing under the assumption of exogeneity and non-differential measurement error, another group of papers relies on the availability of either an instrumental variable or a second measure of T^*. Black et al. (2000) and Kane et al. (1999) consider a linear model and show that when *two* alternative measures T_1 and T_2 of T^* are available, a non-linear GMM estimator can be used to recover the effect of interest. Subsequently, Frazis and Loewenstein (2003) note that an instrumental variable can take the place of one of the measures. Mahajan (2006) extends the results of Black et al. (2000) and Kane et al. (1999) to a more general setting using a binary instrument in place of one of the treatment measures, establishing non-parametric identification of the conditional mean function. When T^* is in fact exogenous, this coincides with the causal effect. Hu (2008) derives related results when the mis-classified discrete regressor may take on more than two values. Lewbel (2007a) provides an identification result for the same model as Mahajan (2006) under different assumptions. In particular, his "instrument-like variable" need not satisfy the usual exclusion restriction so long as it does not interact with T^* and takes on three or more values.

Much less is known about the case in which a binary, or discrete, regressor is not only mis-classified but endogenous. The first paper to provide a formal result for this case is Mahajan (2006). He extends his main result to the case of an endogenous treatment, providing an explicit proof of identification under the usual IV assumption in a model with additively separable errors. As we show below, however, this result is false.[3] Several more recent papers also consider the case of a mis-classified, endogenous, binary regressor. Kreider et al. (2012), partially identify the effects of food stamps on health outcomes of children under weak measurement error assumptions by relying on auxiliary data. Similarly, Battistin et al. (2014) study the returns to schooling in a setting with multiple mis-reported measures of educational qualifications. Unlike these two papers, our approach does not depend on the availability of auxiliary data. In a different vein, Shiu (2016) uses an exclusion restriction for the participation equation and an additional valid instrument to identify the effect of a discrete, mis-classified endogenous regressor in a semi-parametric selection model. Similarly, Nguimkeu et al. (2016) use exclusion restrictions for both the participation equation and measurement error equation to identify a parametric model with endogenous participation and one-sided endogenous mis-reporting. Unlike those of the preceding two papers, our results rely neither on parametric assumptions nor additional exclusion restrictions. Other than Mahajan (2006), the paper most closely related to our own is that of Ura (Forthcoming), who derives partial identification results for a local average treatment effect without the non-differential assumption. In contrast, we study an additively separable model under non-differential measurement error and derive both partial and point identification results.

The remainder of the paper is organized as follows. Section 2.1 describes our model and assumptions, Section 2.2 relates our results to existing work, and Sections 2.3–2.4 present

[3] Appendix B provides a detailed explanation of the error in Mahajan's proof.

our identification results. Section 3 provides a brief discussion of how to carry out inference using our identification results, and Section 4 concludes. Proofs appear in Appendix A, and we give a detailed explanation of the error in Mahajan (2006) in Appendix B. Appendix C explains how our partial identification bounds from Section 2.3 can be interpreted in a local average treatment effects (LATE) setting.

2 Identification

2.1 Baseline Assumptions

As defined in the preceding section, our model is $y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$, where ε is a mean-zero error term, and the parameter of interest is $\beta(\mathbf{x})$ – the effect of an unobserved, binary, endogenous regressor T^*. Suppose we observe a valid and relevant binary instrument z. In the discussion following Corollary 2.2 below, we explain how these results generalize to the case of an arbitrary discrete-valued instrument. We assume that the model and instrument satisfy the following conditions:

Assumption 2.1.

(i) $y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$ *where* $T^* \in \{0, 1\}$ *and* $\mathbb{E}[\varepsilon] = 0$*;*

(ii) $z \in \{0, 1\}$*, where* $0 < \mathbb{P}(z = 1|\mathbf{x}) < 1$*, and* $\mathbb{P}(T^* = 1|\mathbf{x}, z = 1) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 0)$*;*

(iii) $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$.

Assumption 2.1(i) is a restatement of the additively separable model from Equation 1, which includes as a special case the linear model $y = c + \beta T^* + \mathbf{x}'\boldsymbol{\gamma} + \varepsilon$ that is pervasive in empirical economics. Assumptions 2.1(ii) and (iii) are the textbook instrumental variable relevance and validity conditions, respectively. Under Assumption 2.1, the Wald estimator

$$\left[\mathbb{E}\left(y|z = 1, \mathbf{x}\right) - \mathbb{E}\left(y|z = 0, \mathbf{x}\right)\right] / \left[\mathbb{E}\left(T^*|z = 1, \mathbf{x}\right) - \mathbb{E}\left(T^*|z = 0, \mathbf{x}\right)\right]$$

identifies $\beta(\mathbf{x})$. Unfortunately this estimator is infeasible, as we observe not T^* but a mis-classified binary surrogate T.[4] To make further progress, we must impose conditions on the process that generates T. Accordingly, define the following mis-classification probabilities:

$$\begin{aligned} \alpha_0(\mathbf{x}, z) &= \mathbb{P}\left(T = 1|T^* = 0, \mathbf{x}, z\right) & \alpha_0(\mathbf{x}) &= \mathbb{P}\left(T = 1|T^* = 0, \mathbf{x}\right) \\ \alpha_1(\mathbf{x}, z) &= \mathbb{P}\left(T = 0|T^* = 1, \mathbf{x}, z\right) & \alpha_1(\mathbf{x}) &= \mathbb{P}\left(T = 0|T^* = 1, \mathbf{x}\right). \end{aligned}$$

Assumption 2.2.

(i) $\alpha_0(\mathbf{x}, z) = \alpha_0(\mathbf{x})$*,* $\alpha_1(\mathbf{x}, z) = \alpha_1(\mathbf{x})$

(ii) $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$

(iii) $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, z, T^*]$

[4] Although it involves T^*, Assumption 2.1(ii) is testable: see the discussion following Lemma 2.1.

Assumption 2.2, or a variant thereof, is standard in the theoretical literature on mis-classification (Black et al., 2000; Frazis and Loewenstein, 2003; Hu, 2008; Lewbel, 2007a; Mahajan, 2006) and in empirical studies that allow for measurement error in a binary or discrete variable (Battistin et al., 2014; Feng and Hu, 2013; Kane et al., 1999). Assumption 2.2 (i) states that the mis-classification probabilities do not depend on z. Assumption 2.2 (ii) restricts the extent of mis-classification and is equivalent to requiring that T and T^* be positively correlated. Assumption 2.2 (iii) is often referred to as "non-differential measurement error." Intuitively, it maintains that T provides no additional information about ε, and hence y, given knowledge of $(T^*, z, \mathbf{x})$. While Assumption 2.2(ii) is quite mild, Assumptions 2.2 (i) and (iii) are more restrictive, as discussed by Bound et al. (2001). To take a specific example, suppose that y is log wage and T^* is an indicator for college completion. If T is a potentially erroneous measure of college completion taken from a university's administrative records, then the assumption of non-differential measurement error is quite plausible. If, on the other hand, T is a self-report of college completion and there are "returns to lying" about college completion, i.e. employers only imperfectly observe worker ability, this assumption is less plausible.[5] Note, however, that our assumptions on the mis-classification process are *conditional* on $\mathbf{x}$: we place no restrictions on the relationship between observed covariates and the mis-classification errors. In contrast, Bound et al. (2001) considers *unconditional* versions of our Assumption 2.2. Instrument validity – Assumption 2.1 (iii) – is more plausible after conditioning on a rich set of exogenous controls, and the same is true of our mis-classification assumptions. For more discussion of settings in which the assumption of non-differential measurement error is warranted, see Carroll et al. (2006).

2.2 Point Identification Results from the Literature

Existing results from the literature – see for example Frazis and Loewenstein (2003) and Mahajan (2006) – establish that $\beta(\mathbf{x})$ is point identified if Assumptions 2.1–2.2 are augmented to include the following condition:

Assumption 2.3 (Joint Exogeneity)**.** $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*] = 0$.

Assumption 2.3 strengthens the mean independence condition from Assumption 2.1 (iii) to hold *jointly* for T^* and z. By iterated expectations, this implies that T^* is exogenous, i.e. $\mathbb{E}[\varepsilon|\mathbf{x}, T^*] = 0$. If T^* is endogenous, Assumption 2.3 clearly fails. Mahajan (2006) argues, however, that the following restriction, along with our Assumptions 2.1–2.2, suffices to identify $\beta(\mathbf{x})$ when T^* may be endogenous:

Assumption 2.4 (Mahajan (2006) Equation 11)**.** $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, T^*]$.

Assumption 2.4 does not require $\mathbb{E}[\varepsilon|\mathbf{x}, T^*]$ to be zero, but maintains that it does not vary with z. We show in Appendix B, however, that under Assumptions 2.1–2.2, Assumption 2.4 can only hold if T^* is exogenous. If z is a valid instrument and T^* is endogenous, then Assumption 2.4 implies that there is no first-stage relationship between z and T^*. As such, identification in the case where T^* is endogenous is an open question.

[5] See Hu and Lewbel (2012) for a proposal to estimate the "returns to lying" in this context.

2.3 Partial Identification

In this section we derive the sharp identified set under Assumptions 2.1–2.2 and show that $\beta(\mathbf{x})$ is not point identified. For a discussion of how our partial identification results can be interpreted in a local average treatment effects (LATE) setting, see Appendix C.

To simplify the notation, define the following shorthand for the unobserved and observed first stage probabilities

$$p_k^*(\mathbf{x}) = \mathbb{P}(T^* = 1|\mathbf{x}, z = k), \quad p_k(\mathbf{x}) = \mathbb{P}(T = 1|\mathbf{x}, z = k). \tag{2}$$

We first state two lemmas that that will be used repeatedly below.

Lemma 2.1. *Under Assumption 2.2 (i),*

$$\begin{aligned}[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\, p_k^*(\mathbf{x}) &= p_k(\mathbf{x}) - \alpha_0(\mathbf{x}) \\ [1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\, [1 - p_k^*(\mathbf{x})] &= 1 - p_k(\mathbf{x}) - \alpha_1(\mathbf{x})\end{aligned}$$

where the first-stage probabilities $p_k^*(\mathbf{x})$ and $p_k(\mathbf{x})$ are as defined in Equation 2.

Lemma 2.2. *Under Assumptions 2.1 and 2.2 (i)–(ii),*

$$\beta(\mathbf{x})\, Cov(z, T|\mathbf{x}) = [1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\, Cov(y, z|\mathbf{x})$$

Lemma 2.1 relates the observed first-stage probabilities $p_k(\mathbf{x})$ to their unobserved counterparts $p_k^*(\mathbf{x})$ in terms of the mis-classification probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. By Assumption 2.2 (ii), $1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x}) > 0$ so that Lemma 2.1 bounds $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ in terms of the observed first-stage probabilities. Moreover, by taking differences evaluated at $k = 1$ and $k = 0$, this Lemma shows that $p_0^*(\mathbf{x}) = p_1^*(\mathbf{x})$ if and only if $p_0(\mathbf{x}) = p_1(\mathbf{x})$. In other words, Assumption 2.1 (ii) is testable under Assumption 2.2 (ii). Lemma 2.2 relates the instrumental variables (IV) estimand, $\text{Cov}(y, z|\mathbf{x})/\text{Cov}(z, T|\mathbf{x})$, to the mis-classification probabilities. Since $1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x}) > 0$, IV is biased *upwards* in the presence of mis-classification. Together these lemmas bound the causal effect of interest: $\beta(\mathbf{x})$ lies between the reduced form and IV estimators. Without Assumption 2.2 (iii), non-differential measurement error, these bounds are sharp.

Theorem 2.1. *Under Assumptions 2.1 and 2.2 (i)–(ii), $\alpha_0(\mathbf{x}) \leq p_k(\mathbf{x}) \leq 1 - \alpha_1(\mathbf{x})$ for $k = 0, 1$ and*

$$\mathbb{E}[y|\mathbf{x}, z = k] = c(\mathbf{x}) + \beta(\mathbf{x}) \left[\frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})} \right]. \tag{3}$$

Provided that $p_0(\mathbf{x}) \neq p_1(\mathbf{x})$, these expressions characterize the sharp identified set for $c(\mathbf{x})$, $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$, and $\alpha_1(\mathbf{x})$.

Corollary 2.1. *Under the conditions of Theorem 2.1, the sharp identified set for $\beta(\mathbf{x})$ is the closed interval between the reduced form estimand $Cov(y, z|\mathbf{x})/\, Var(z|\mathbf{x})$ and the IV estimand $Cov(y, z|\mathbf{x})/\, Cov(z, T|\mathbf{x})$.*

Corollary 2.1 follows by taking differences of the expression for $\mathbb{E}[y|\mathbf{x}, z = k]$ across $k = 1$ and $k = 0$, and substituting the maximum and minimum value for $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x})$

consistent with the observed first-stage probabilities.[6] Note that the only role of the condition $p_0(\mathbf{x}) \neq p_1(\mathbf{x})$ in the preceding two results is to ensure that it is possible to satisfy Assumption 2.1 (ii). Frazis and Loewenstein (2003) point out that the IV estimand provides an upper bound for $\beta(\mathbf{x})$, and Lemmas 2.1–2.2 are well-known in the literature (see e.g. Frazis and Loewenstein, 2003; Mahajan, 2006). Nevertheless, we are unaware of any published result that explicitly states both bounds from Corollary 2.1 or proves that they are sharp under Assumptions 2.1 and 2.2 (i)–(ii).

Neither Theorem 2.1 nor Corollary 2.1 imposes Assumption 2.2 (iii) – non-differential measurement error. While this assumption plays an important role in existing identification results for an exogenous T^* (see Section 2.2), its identifying power under endogeneity has not been addressed in the literature.[7] We now show that this assumption in general yields further restrictions on probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$, but fails to point identify $\beta(\mathbf{x})$. To simplify the proof of sharpness, we assume that y is continuously distributed, which is natural in an additively separable model. Without this assumption, the bounds that we derive are still valid, but may not be sharp. Nevertheless, the reasoning from our proof can be generalized to cases in which y does not have a continuous support set.

Theorem 2.2. *Suppose that the conditional distribution of y given $(\mathbf{x}, T, z)$ is continuous. Further suppose that the conditions of Theorem 2.1 and Assumption 2.2 (iii) hold. For any k such that $\mathbb{E}[y|\mathbf{x}, T=0, z=k] \neq \mathbb{E}[y|\mathbf{x}, T=1, z=k]$, let $\mathcal{A}_k$ denote the set of pairs $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big)$ such that $\alpha_0(\mathbf{x}) < p_k(\mathbf{x}) < 1 - \alpha_1(\mathbf{x})$ and*

$$\underline{\mu}_{tk}\Big(\underline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big),\, \mathbf{x}\Big) \leq \mu_k\big(\alpha_0(\mathbf{x}), \mathbf{x}\big) \leq \overline{\mu}_{tk}\Big(\overline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big),\, \mathbf{x}\Big)$$

for all $t = 0, 1$ where

$$\underline{\mu}_{tk}\big(q, \mathbf{x}\big) = \mathbb{E}\left[y \mid y \leq q, \mathbf{x}, T = t, z = k\right], \qquad \overline{\mu}_{tk}\big(q, \mathbf{x}\big) = \mathbb{E}\left[y \mid y > q, \mathbf{x}, T = t, z = k\right]$$

$$\mu_k\big(\alpha_0(\mathbf{x}), \mathbf{x}\big) = \frac{p_k(\mathbf{x})\mathbb{E}[y|\mathbf{x}, z=k, T=1] - \alpha_0(\mathbf{x})\mathbb{E}[y|\mathbf{x}, z=k]}{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}$$

and we define

$$\underline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = F_{tk}^{-1}\Big(r_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) \,\Big|\, \mathbf{x}\Big)$$
$$\overline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = F_{tk}^{-1}\Big(1 - r_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) \,\Big|\, \mathbf{x}\Big)$$

[6] If *a priori* restrictions on α_0 and α_1 are available, e.g. $\alpha_0 = 0$, $\alpha_1 = 0$, or $\alpha_0 = \alpha_1$, these bounds can be improved. For more discussion, see Corollary 2.2 of DiTraglia and García-Jimeno (2017).

[7] The only exception is the incorrect result of Mahajan (2006) described in Section 2.2 and Appendix B.

where $F_{tk}^{-1}(\cdot|\mathbf{x})$ *is the conditional quantile function of* y *given* $(\mathbf{x}, T = t, z = k)$,

$$r_{0k}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = \frac{\alpha_1(\mathbf{x})}{1 - p_k(\mathbf{x})}\left[\frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}\right]$$
$$r_{1k}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = \frac{1 - \alpha_1(\mathbf{x})}{p_k(\mathbf{x})}\left[\frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}\right]$$

and $p_k(\mathbf{x})$ *is defined in Equation 2. The sharp identified set for* $c(\mathbf{x})$, $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$ *and* $\alpha_1(\mathbf{x})$ *is characterized by Equation 3 and* $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big) \in \mathcal{A}^*$ *where*

(i) $\mathcal{A}^* \equiv \mathcal{A}_0 \cap \mathcal{A}_1$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *for all* $k = 0, 1$*;*

(ii) $\mathcal{A}^* \equiv \mathcal{A}_k$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *and* $\mathbb{E}[y|\mathbf{x}, T = 0, z = \ell] = \mathbb{E}[y|\mathbf{x}, T = 1, z = \ell]$*;*

(iii) $\mathcal{A}^* \equiv \big\{\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big) \colon \alpha_0(\mathbf{x}) \leq p_k(\mathbf{x}) \leq 1 - \alpha_1(\mathbf{x})$ *for all* $k\big\}$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] = \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *for all* $k = 0, 1$*.*

Imposing Assumption 2.2 (iii) strictly improves upon the identified set from Theorem 2.1 unless $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] = \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ for all k. Even if $\beta(\mathbf{x}) = 0$, the difference of these observable means is generically nonzero.[8] The intuition for Theorem 2.2 is as follows. For simplicity, suppress dependence on $\mathbf{x}$. Now, fix $(T = t, z = k)$ and (α_0, α_1). The observed distribution of y given $(T = t, z = k)$, call it F_{tk}, is a mixture of two unobserved distributions: the distribution of y given $(T = k, z = k, T^* = 1)$, call it F_{tk}^1, and the distribution of y given $(T = t, z = k, T^* = 0)$, call it F_{tk}^0. The mixing probabilities are r_{tk} and $1 - r_{tk}$ from the statement of Theorem 2.2 and are fully determined by (α_0, α_1) and p_k. Assumptions 2.1 (i) and 2.2 (ii) imply that the unobserved means $\mathbb{E}[y|T^*, T, z]$ are fully determined by (α_0, α_1) given the observed means $\mathbb{E}[y|T, z]$. The question is whether it is possible, given the observed distribution F_{tk}, to construct F_{tk}^1 and F_{tk}^0 with the required values for $\mathbb{E}[y|T^*, T, z]$ such that $F_{tk} = r_{tk}F_{tk}^1 + (1 - r_{tk})F_{tk}^0$ for all combinations (t, k). If not, then (α_0, α_1) does not belong to the identified set. Our proof provides necessary and sufficient conditions for such a mixture to exist at a given point (α_0, α_1). We can then appeal to the reasoning from Theorem 2.1 to complete the argument. By ruling out values for α_0 and α_1, Theorem 2.2 restricts β via Lemma 2.2. While these restrictions can be very informative, they do not yield point identification.

Corollary 2.2. *Under Assumptions 2.1 and 2.2 the identified set for* $\beta(\mathbf{x})$ *contains both the IV estimand* $Cov(y, z|\mathbf{x})/Cov(z, T|\mathbf{x})$ *and the true coefficient* $\beta(\mathbf{x})$*.*

Corollary 2.2 follows by Lemma 2.2 because $\alpha_0(\mathbf{x}) = \alpha_1(\mathbf{x}) = 0$ always belongs to the sharp identified set from Theorem 2.2. Non-differential measurement error cannot exclude the possibility that there is no mis-classification because in this case it is trivial to construct the required mixtures. Although we focus throughout this paper on the case of a binary instrument, one might wonder whether point identification can be achieved by increasing

[8] Suppress dependence on $\mathbf{x}$ for simplicity. There are only two settings in which $\mathbb{E}[y|T = 0, z = k] = \mathbb{E}[y|T = 1, z = k]$. The first is if the true value of either α_0 or α_1 lies at the upper boundary of the identified set from Theorem 2.1. The second is if $\beta = \mathbb{E}[\varepsilon|T^* = 0, z = k] - \mathbb{E}[\varepsilon|T^* = 0, z = k]$.

the support of z, perhaps along the lines of Lewbel (2007a). The answer turns out to be no. Suppose that we were to modify Assumptions 2.1 and 2.2 to hold for all values of z in some discrete support set. By Lemma 2.2, a binary instrument identifies $\beta(\mathbf{x})$ up to knowledge of the mis-classification probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. It follows that *any* pair of values (k, ℓ) in the support set of z identifies the same object. Accordingly, to identify $\beta(\mathbf{x})$ it is necessary and sufficient to identify the mis-classification probabilities. A binary instrument fails to identify these probabilities because we can never exclude the possibility of zero mis-classification. The same is true of a discrete K-valued instrument. Increasing the support of z does, however, shrink the identified set by increasing the number of restrictions available: in this case Theorems 2.1–2.2 continue to apply replacing "$k = 0, 1$" with "for all k."

2.4 Point Identification

The results of the preceding section establish that $\beta(\mathbf{x})$ is not point identified under Assumptions 2.1 and 2.2. In light of this, there are two possible ways to proceed: either one can report partial identification bounds based on our characterization of the sharp identified set from Theorem 2.2, or one can attempt to impose stronger assumptions to obtain point identification. In this section we consider the second possibility. We begin by defining the following functions of the model parameters:

$$\theta_1(\mathbf{x}) = \beta(\mathbf{x}) \left[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]^{-1} \tag{4}$$

$$\theta_2(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^2 \left[1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right] \tag{5}$$

$$\theta_3(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^3 \left[\left\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right\}^2 + 6\alpha_0(\mathbf{x}) \left\{1 - \alpha_1(\mathbf{x})\right\}\right] \tag{6}$$

Now consider the following additional assumption:

Assumption 2.5. $\mathbb{E}[\varepsilon^2|\mathbf{x}, z] = \mathbb{E}[\varepsilon^2|\mathbf{x}]$

Assumption 2.5 is a *second moment* version of the standard mean exclusion restriction for the instrument z – Assumption 2.1 (iii). It requires that the conditional variance of the error term given the covariates $\mathbf{x}$ does not depend on z, but does *not* require homoskedasticity with respect to $\mathbf{x}, T^*$ or T. Assumption 2.5 allows us to derive the following lemma:

Lemma 2.3. *Under Assumptions 2.1, 2.2 and 2.5,*

$$Cov(y^2, z|\mathbf{x}) = 2Cov(yT, z|\mathbf{x})\theta_1(\mathbf{x}) - Cov(T, z|\mathbf{x})\theta_2(\mathbf{x})$$

where $\theta_1(\mathbf{x})$ and $\theta_2(\mathbf{x})$ are defined in Equations 4–5.

Lemma 2.2 identifies $\theta_1(\mathbf{x})$. Since $\text{Cov}(z, T|\mathbf{x}) \neq 0$ by Assumption 2.1 (ii), we can solve for $\theta_2(\mathbf{x})$ in terms of observables only, using Lemma 2.3. Given knowledge of $\theta_1(\mathbf{x})$, we can solve Equation 5 for the difference of mis-classification rates so long as $\beta(\mathbf{x}) \neq 0$.

Corollary 2.3. *Under Assumptions 2.1–2.2 and 2.5, $\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})$ is identified so long as $\beta(\mathbf{x}) \neq 0$.*

Corollary 2.3 identifies the difference of mis-classification error rates. Hence, under one-sided mis-classification, $\alpha_0(\mathbf{x}) = 0$ or $\alpha_1(\mathbf{x}) = 0$, augmenting our baseline Assumptions 2.1–2.2 with Assumption 2.5 suffices to identify $\beta(\mathbf{x})$. Notice that $\beta(\mathbf{x}) = 0$ if and only if $\theta_1(\mathbf{x}) = 0$. Thus, $\beta(\mathbf{x})$ is still identified in the case where Corollary 2.3 fails to apply.

Assumption 2.5 does not suffice to identify $\beta(\mathbf{x})$ without *a priori* restrictions on the mis-classification error rates. To achieve identification in the general case, we impose the following additional conditions:

Assumption 2.6.

(i) $\mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*]$

(ii) $\mathbb{E}[\varepsilon^3|\mathbf{x}, z] = \mathbb{E}[\varepsilon^3|\mathbf{x}]$

Assumption 2.6 (i) is a second moment version of the non-differential measurement error assumption, Assumption 2.2 (iii). It requires that, given knowledge of $(\mathbf{x}, T^*, z)$, T provides no additional information about the variance of the error term. Note that Assumption 2.6 (i) does not require homoskedasticity of ε with respect to $\mathbf{x}$ or T^*. Assumption 2.6 (ii) is a third moment version of Assumption 2.5. It requires that the conditional third moment of the error term given $\mathbf{x}$ does not depend on z. This condition neither requires nor excludes skewness in the error term conditional on covariates: it merely states that the skewness is unaffected by the instrument. While Assumptions 2.5 and 2.6 may appear somewhat unusual, they are implied by the more intuitive independence conditions $\varepsilon \perp\!\!\!\perp z|\mathbf{x}$ and $\varepsilon \perp\!\!\!\perp T|(\mathbf{x}, T^*, z)$. Although $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$ and $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, z, T^*]$ are technically weaker than assuming full independence, we would be somewhat dubious of any supposed "natural experiment" that purportedly satisfied mean exclusion but not independence. Indeed, as discussed by Imbens and Rubin (1997), an instrument satisfying mean exclusion but not independence could become invalid if the outcome variable were transformed, for example by taking logs. As it is not uncommon for applied papers to report results in both logs and levels (e.g. Angrist, 1990), our view is that researchers *implicitly* assume more than mean exclusion in typical applications of instrumental variables. Analogous reasoning applies to the non-differential measurement error assumption.

Assumption 2.6 allows us to derive the following Lemma which, combined with Lemma 2.3, leads to point identification:

Lemma 2.4. *Under Assumptions 2.1–2.2 and 2.5–2.6,*

$$Cov(y^3, z|\mathbf{x}) = 3Cov(y^2T, z|\mathbf{x})\theta_1(\mathbf{x}) - 3Cov(yT, z|\mathbf{x})\theta_2(\mathbf{x}) + Cov(T, z|\mathbf{x})\theta_3(\mathbf{x})$$

where $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$ are defined in Equations 4–5.

Theorem 2.3. *Under Assumptions 2.1–2.2 and 2.5–2.6, $\beta(\mathbf{x})$ is identified. If $\beta(\mathbf{x}) \neq 0$, then $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are likewise identified.*

Lemmas 2.2–2.4 yield a linear system of three equations in $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$. Under Assumption 2.1 (ii), the system has a unique solution so $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$ are identified.

The proof of Theorem 2.3 shows that, so long as $\beta(\mathbf{x}) \neq 0$, Equations 4–6 can be solved for $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. In particular, using steps from the proof of Theorem 2.3

$$\beta(\mathbf{x}) = \text{sign}\big[\theta_1(\mathbf{x})\big]\sqrt{3\big[\theta_2(\mathbf{x})/\theta_1(\mathbf{x})\big]^2 - 2\big[\theta_3(\mathbf{x})/\theta_1(\mathbf{x})\big]}.$$

If we relax Assumption 2.2 (ii) and assume $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) \neq 1$ only, $\beta(\mathbf{x})$ is only identified up to sign: in this case the sign of $\theta_1(\mathbf{x})$ need not equal that of $\beta(\mathbf{x})$.

3 Estimation and Inference

We now briefly outline how the identification results from Section 2 can be used to estimate and carry out statistical inference for the parameters of interest: $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \beta(\mathbf{x})\big)$. Lemmas 2.2–2.4 yield a system of linear moment equations in the reduced form parameters $\boldsymbol{\theta}'(\mathbf{x}) = \big(\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \theta_3(\mathbf{x})\big)$. Defining a vector of intercepts $\boldsymbol{\kappa}'(\mathbf{x}) = \big(\kappa_1(\mathbf{x}), \kappa_2(\mathbf{x}), \kappa_3(\mathbf{x})\big)$, and a vector of observables $\mathbf{w}' = (T, y, yT, y^2, y^2T, y^3)$, we can write this system as

$$\mathbb{E}\left[\left\{\boldsymbol{\Psi}\big(\boldsymbol{\theta}(\mathbf{x})\big)\mathbf{w}_i - \boldsymbol{\kappa}(\mathbf{x})\right\} \otimes \begin{pmatrix} 1 \\ z \end{pmatrix} \middle| \mathbf{x} = \boldsymbol{x}\right] = \mathbf{0} \tag{7}$$

$$\boldsymbol{\Psi}\big(\boldsymbol{\theta}(\mathbf{x})\big) \equiv \begin{bmatrix} -\theta_1(\mathbf{x}) & 1 & 0 & 0 & 0 & 0 \\ \theta_2(\mathbf{x}) & 0 & -2\theta_1(\mathbf{x}) & 1 & 0 & 0 \\ -\theta_3(\mathbf{x}) & 0 & 3\theta_2(\mathbf{x}) & 0 & -3\theta_1(\mathbf{x}) & 1 \end{bmatrix}. \tag{8}$$

Using Equations 4–6, we can re-write $\boldsymbol{\Psi}$ as a function of $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \beta(\mathbf{x})\big)$, leaving us with a just-identified, non-parametric conditional moment problem. Because the conditioning variables in Equation 7 are the same as the arguments of the unknown functions $(\alpha_0, \alpha_1, \beta)$, this problem fits within the framework of Lewbel (2007b), permitting straightforward estimation and inference via a local GMM procedure. If $\beta(\mathbf{x})$ is close to zero, however, this procedure can perform poorly; in this case the moment conditions from Equations 7, are only weakly informative about $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. An earlier version of this paper (DiTraglia and García-Jimeno, 2017) discusses this problem in more detail and provides a solution based on generalized moment selection (Andrews and Soares, 2010) that combines the moment inequalities implied by our partial identification results from Section 2.3 with the moment equalities from Equation 7.

4 Conclusion

This paper has studied identification and inference for a mis-classified, binary, endogenous regressor in an additively separable model using a discrete instrumental variable. We have shown that the only existing identification result for this model is incorrect, and gone on to derive the sharp identified set under standard first-moment assumptions from the literature. Strengthening these assumptions to hold for second and third moments, we have established point identification for the effect of interest. An interesting extension of the results presented above would be to consider the case of discrete regressors that take on more than two values.

A Proofs

All of the results in this paper hold $\mathbf{x}$ *fixed*. This allows us to completely ignore the presence of covariates in the proofs that follow. Accordingly we work in terms of scalars $\alpha_0, \alpha_1, \beta, p_k$, etc. rather than functions $\alpha(\mathbf{x})$, $\alpha_1(\mathbf{x})$, $\beta(\mathbf{x}), p_k(\mathbf{x})$. The former should be understood as the value of the latter evaluated at some particular $\mathbf{x}$.

A.1 Partial Identification Results

Proof of Lemma 2.1. Follows from a simple calculation using the law of total probability. □

Proof of Lemma 2.2. Immediate since $\text{Cov}(z, T) = (1 - \alpha_0 - \alpha_1)\text{Cov}(z, T^*)$ by Lemma 2.1. □

Proof of Theorem 2.1. To show that $\alpha_0 \leq p_k \leq 1 - \alpha_1$, substitute $p_k^* = 0$ and $p_k^* = 1$, respectively, into Lemma 2.1 and rearrange. To show that $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$, take conditional expectations of Equation 1 and apply Assumption 2.1 (iii) and Lemma 2.1.

To prove sharpness we need to show that for any $(c, \beta, \alpha_0, \alpha_1)$ that satisfy $\alpha_0 \leq p_k \leq 1 - \alpha_1$ and $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ we can construct a valid joint distribution for (y, T, T^*, z) that is compatible with the observed distribution of (y, T, z), provided that $p_1 \neq p_0$. To establish this result, we factorize the joint distribution of (y, T, T^*, z) into the product of a conditional $y|(T, T^*, z)$ and marginal (T, T^*, z). The argument proceeds in two steps. Our first step relies on the fact that Assumptions 2.1 (i) and (iii) do not constrain the distribution of (T, T^*, z) while 2.1 (ii) and 2.2 (i)–(ii) constrain *only* the distribution of (T, T^*, z). Under these latter three assumptions, we show how to construct a valid joint distribution for (T, T^*, z) that is compatible with the observed distribution of (T, z) for any (α_0, α_1) satisfying $\alpha_0 \leq p_k \leq 1 - \alpha_1$. Our second step shows how to construct a valid conditional distribution for y given (T, T^*, z) under Assumptions 2.1 (i) and (iii) that is compatible with the observed conditional distribution of y given (T, z) for any $(c, \beta, \alpha_0, \alpha_1)$ satisfying $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)(1 - \alpha_0 - \alpha_1)$. Combining the two steps gives the required joint distribution for (y, T^*, T, z).

For the first step, we need to construct a valid joint probability mass function $p(T^*, T, z)$ with support set $\{0, 1\} \times \{0, 1\} \times \{0, 1\}$. By Assumption 2.2 (i), $p(T|T^*, z) = p(T|T^*)$ and hence

$$p(T^*, T, z) = p(T|T^*)p(T^*|z)p(z).$$

Since $p(z)$ is observed, to construct a valid joint probability mass function $p(T^*, T, z)$ it suffices to construct valid *conditional* probability mass functions $p(T|T^*)$ and $p(T^*|z)$. Since $\alpha_0 \leq p_k \leq 1-\alpha_1$, both α_0 and α_1 are guaranteed to lie between zero and one. This gives a valid construction of $p(T|T^*)$. Moreover the corresponding values of p_k^* implied by Lemma 2.1 are also guaranteed to lie between zero and one. This gives a valid construction of $p(T^*|z)$ that satisfies Assumption 2.1 (ii), since $p_1 \neq p_0$ by assumption and $(p_1 - p_0) = (p_1^* - p_0^*)(1 - \alpha_0 - \alpha_1)$ by Lemma 2.1. Because our construction relies on Lemma 2.1, which is simply an application of the law of total probability, the resulting distribution $p(T, T^*, z)$ is automatically compatible with $p(T, z) = p(T|z)p(z)$.

For the second step, we need to construct a valid conditional distribution for y given (T, T^*, z). To begin we define the following notation:

$$\begin{aligned} r_{tk} &\equiv \mathbb{P}(T^* = 1|T = t, z = k) & F_t(\tau) &\equiv \mathbb{P}(y \leq \tau|z = k) \\ F_{tk}(\tau) &\equiv \mathbb{P}(y \leq \tau|T = t, z = k) & F_{tk}^{t^*}(\tau) &\equiv \mathbb{P}(y \leq \tau|T^* = t^*, T = t, z = k) \\ G_k(\tau) &\equiv \mathbb{P}(\varepsilon \leq \tau|z = k) & G_{tk}^{t^*}(\tau) &\equiv \mathbb{P}(\varepsilon \leq \tau|T^* = t^*, T = t, z = k). \end{aligned}$$

Assumption 2.1 (i) imposes a relationship between $G_{tk}^{t^*}$ and $F_{tk}^{t^*}$ for each t^*, namely

$$G_{tk}^0(\tau) = F_{tk}^0(\tau + c), \quad G_{tk}^1(\tau) = F_{tk}^1(\tau + c + \beta) \tag{A.1}$$

and thus we see that

$$\begin{aligned} G_k(\tau) = {} & r_{1k} p_k F_{1k}^1(\tau + c + \beta) + r_{0k}(1 - p_k) F_{0k}^1(\tau + c + \beta) \\ & + (1 - r_{1k}) p_k F_{1k}^0(\tau + c) + (1 - r_{0k})(1 - p_k) F_{0k}^0(\tau + c) \end{aligned} \tag{A.2}$$

applying the law of total probability and Bayes' rule. Moreover,

$$F_{tk}(\tau) = r_{tk} F_{tk}^1(\tau) + (1 - r_{tk}) F_{tk}^0(\tau) \tag{A.3}$$

for all $t, k \in \{0, 1\}$, and by Bayes' rule,

$$r_{1k} = (1 - \alpha_1) p_k^* / p_k, \quad r_{0k} = \alpha_1 p_k^* / (1 - p_k). \tag{A.4}$$

There are four cases, corresponding to different possibilities for the r_{tk}. The first case violates one of our model assumptions. For each of the remaining cases, we show that it is possible to construct the required distributions F_{tk}^0, F_{tk}^1 under Assumptions 2.1 (i) and (iii) for any $(c, \beta, \alpha_0, \alpha_1)$ such that $\mathbb{E}(y|z = k) = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$.

Case I: $r_{1k} = 0, r_{0k} \neq 0$ By Equation A.4 this requires $\alpha_1 = 1$, violating Assumption 2.2 (ii).

Case II: $r_{0k} = r_{1k} = 0$ By Equation A.4, this requires $p_k^* = 0$ which in turn requires $p_k = \alpha_0$. By Equation A.3 we have $F_{tk}^0 = F_{tk}$, while F_{tk}^1 is unrestricted. Substituting into A.2,

$$G_k(\tau) = p_k F_{1k}(\tau + c) + (1 - p_k) F_{0k}(\tau + c) = F_k(\tau + c)$$

Now, since $F_k(\tau + c)$ is the conditional CDF of $y - c$ given that $z = k$, and G_k is the conditional CDF of ε given $z = k$, we see that Assumption 2.1 (i) is satisfied if and only if $\mathbb{E}(y|z = k) = c$, which is equal to $c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ since $p_k - \alpha_0 = 0$.

Case III: $r_{1k} \neq 0, r_{0k} = 0$ By Equation A.4 this requires $\alpha_1 = 0$ and $p_k^* \neq 0$. By Equation A.3 we have $F_{0k}^0 = F_{0k}$ and since $r_{1k} \neq 1$, we can solve to obtain

$$F_{1k}^1(\tau) = \frac{1}{r_{1k}} \left[F_{1k}(\tau) - (1 - r_{1k}) F_{1k}^0(\tau) \right]$$

Substituting into Equation A.2, we obtain

$$\begin{aligned} G_k(\tau) = {} & \left[(1 - p_k) F_{0k}(\tau + c) + p_k F_{1k}(\tau + c + \beta) \right] \\ & + p_k (1 - r_{1k}) \left[F_{1k}^0(\tau + c) - F_{1k}^0(\tau + c + \beta) \right] \end{aligned}$$

Now, $F_{0k}(\tau + c)$ is the conditional CDF of $(y - c)$ given $(T = 0, z = k)$ while $F_{1k}(\tau + c + \beta)$ is the conditional CDF of $(y - c - \beta)$ given $(T = 1, z = k)$. Similarly, $F_{1k}^0(\tau + c)$ is the conditional CDF of ε given $(T^* = 0, T = 1, z = k)$ while $F_{1k}^0(\tau + c + \beta)$ is the conditional CDF of $(\varepsilon - \beta)$ given $(T^* = 0, T = 1, z = k)$. Since $G_k(\tau)$ is the conditional CDF of ε given $z = k$, we see that

Assumption 2.1 (iii) is satisfied if and only if

$$\begin{aligned}0 = (1-p_k)\mathbb{E}(y-c|T=0,z=k) + p_k\mathbb{E}(y-c-\beta|T=1,z=k) \\ + p_k(1-r_{1k})\left[\mathbb{E}(\varepsilon|T^*=0,T=1,z=k) - \mathbb{E}(\varepsilon-\beta|T^*=0,T=1,z=k)\right]\end{aligned}$$

Rearranging, this is equivalent to

$$\mathbb{E}(y|z=k) = c + (1-\alpha_1)\beta\left(\frac{p_k-\alpha_0}{1-\alpha_0-\alpha_1}\right) = c + \beta\left(\frac{p_k-\alpha_0}{1-\alpha_0-\alpha_1}\right)$$

since $\alpha_1 = 0$ in this case. As explained above, $F^0_{0k} = F_{0k}$ in the present case while F^1_{0k} is undefined. We are free to choose any distributions for F^0_{1k} and F^1_{1k} that satisfy Equation A.3, for example $F^0_{1k} = F^1_{1k} = F_{1k}$.

Case IV: $r_{1k} \neq 0, r_{0k} \neq 0$ In this case, we can solve Equation A.3 to obtain

$$F^1_{tk}(\tau) = \frac{1}{r_{tk}}\left[F_{tk}(\tau) - (1-r_{tk})F^0_{tk}(\tau)\right]$$

Substituting this into Equation A.2, we have

$$\begin{aligned}G_k(\tau) = F_k(\tau+c+\beta) + p_k(1-r_{1k})\left[F^0_{1k}(\tau+c) - F^0_{1k}(\tau+c+\beta)\right] \\ + (1-p_k)(1-r_{0k})\left[F^0_{0k}(\tau+c) - F^0_{0k}(\tau+c+\beta)\right]\end{aligned}$$

using the fact that $F_k(\tau) = p_kF_{1k}(\tau) + (1-p_k)F_{0k}(\tau)$. Now, $F_k(\tau+c+\beta)$ is the conditional CDF of $(y-c-\beta)$ given $z=k$, while $F^0_{tk}(\tau+c)$ is the conditional CDF of ε given $(T=t,z=k)$ and $F^0_{tk}(\tau+c+\beta)$ is the conditional CDF of $(\varepsilon-\beta)$ given $(T=t,z=k)$. Since $G_k(\tau)$ is the conditional CDF of ε given $z=k$, we see that Assumption 2.1 (iii) is satisfied if and only if

$$\begin{aligned}0 &= \mathbb{E}[y-c-\beta|z=k] + p_k(1-r_{1k})\left[\mathbb{E}(\varepsilon|T^*=0,T=1,z=k) - \mathbb{E}(\varepsilon-\beta|T^*=0,T=1,z=k)\right] \\ &\quad + (1-p_k)(1-r_{0k})\left[\mathbb{E}(\varepsilon|T^*=0,T=0,z=k) - \mathbb{E}(\varepsilon-\beta|T^*=0,T=0,z=k)\right] \\ 0 &= \mathbb{E}[y-c-\beta|z=k] + \beta\left[p_k(1-r_{1k}) + (1-p_k)(1-r_{0k})\right]\end{aligned}$$

But since $[p_k(1-r_{1k}) + (1-p_k)(1-r_{0k})] = (1-p^*_k)$ and $p^*_k = (p_k-\alpha_0)/(1-\alpha_0-\alpha_1)$, this becomes

$$\mathbb{E}[y|z=k] = c + \beta\left[(p_k-\alpha_0)(1-\alpha_0-\alpha_1)\right].$$

Thus, in this case we are free to choose *any* distributions for F^0_{tk} and F^1_{tk} that satisfy Equation A.3. For example we could take $F^0_{tk} = F^1_{tk} = F_{tk}$. □

Proof of Corollary 2.1. The result follows by substituting the largest and smallest possible values for $\alpha_0+\alpha_1$ and taking the difference of the expressions for $\mathbb{E}[y|z=k]$. □

Proof of Theorem 2.2. The only difference between the conditions of Theorem 2.1 and those of 2.2 is that the latter imposes Assumption 2.2 (iii) while the former does not. Accordingly, the present argument builds on the proof of Theorem 2.1 and relies on the notation defined within it. Under Assumption 2.1 (i), Assumption 2.2 (iii) is equivalent to $\mathbb{E}[y|T,T^*,z] = \mathbb{E}[y|T^*,z]$. Hence, non-differential measurement error constrains only the conditional distribution of y given (T,T^*,z). For this reason, we need only revisit the second step of the proof of Theorem 2.1. Consider a point $(c,\beta,\alpha_0,\alpha_1)$ that satisfies Equation 3 and $\alpha_0 \leq p_k \leq 1-\alpha_1$ for all k. Since this point lies in the

identified set from Theorem 2.1, it suffices to determine whether there exist valid conditional CDFs F^0_{tk}, F^1_{tk} such that $F_{tk} = (1 - r_{tk})F^0_{tk} + r_{tk}F^1_{tk}$ for all t, k and $\mathbb{E}[y|T, T^*, z] = \mathbb{E}[y|T^*, z]$.

Let $\mu^{t^*}_{tk} \equiv \mathbb{E}[y|T = t, z = k, T^* = t^*]$, $\mu_{tk} \equiv \mathbb{E}[y|T = t, z = k]$, and $\mu^{t^*}_k \equiv \mathbb{E}[y|z = k, T^* = t^*]$. By Assumption 2.2 (iii) $\mu^{t^*}_{tk} = \mu^{t^*}_k$ for $t^* = 0, 1$. Hence, by iterated expectations,

$$\mu_{0k} = (1 - r_{0k})\mu^0_k + r_{0k}\mu^1_k$$
$$\mu_{1k} = (1 - r_{1k})\mu^0_k + r_{1k}\mu^1_k.$$

Now, (μ_{0k}, μ_{1k}) are observed while r_{0k} and r_{1k} depend only on the observed first-stage probability p_k and the mis-classification probabilities (α_0, α_1). Thus, at a given point $(c, \beta, \alpha_0, \alpha_1)$ in the identified set from Theorem 2.1 the preceding equations form a linear system in μ^0_k and μ^1_k. After some algebra, we find that the determinant is

$$r_{1k} - r_{0k} = \left[\frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}\right]\left[\frac{1 - p_k - \alpha_1}{p_k(1 - p_k)}\right].$$

Suppose first that $r_{0k} = r_{1k} = r$ so the determinant condition fails. This occurs if and only if $\alpha_0 = p_k$ or $\alpha_1 = 1 - p_k$. If $\mu_{0k} \neq \mu_{1k}$, the system is inconsistent: no solution for (μ^0_k, μ^1_k) exists. Hence $\alpha_0 = p_k$ and $\alpha_1 = 1 - p_k$ are excluded from the identified set under non-differential measurement error so long as $\mu_{0k} \neq \mu_{1k}$. If instead $\mu_{0k} = \mu_{1k} = \mu$, the system is consistent but rank deficient: any pair (μ^0_k, μ^1_k) such that $\mu = (1 - r)\mu^0_k + r\mu^1_k$ is a solution and hence satisfies the assumption of non-differential measurement error. One such solution is $\mu^1_k = \mu^0_k = \mu$ so we are free to set $F^0_{0k} = F^1_{0k} = F_{0k}$ and $F^0_{1k} = F^1_{1k} = F_{1k}$. Hence, if $\mu_{0k} = \mu_{1k}$ then $\alpha_0 = p_k$ lies within the sharp identified set if $p_k < p_\ell$ and $\alpha_1 = 1 - p_k$ lies in the sharp identified set if $p_\ell < p_k$.

Now suppose that $r_{0k} \neq r_{1k}$, which occurs if and only if $\alpha_0 \neq p_k$ and $\alpha_1 \neq 1 - p_k$. In this case the system has a unique solution, namely

$$\mu^0_k = \frac{r_{1k}\mu_{0k} - r_{0k}\mu_{1k}}{r_{1k} - r_{0k}} = \frac{(1 - p_k)\mathbb{E}(y|T = 0, z = k) - \alpha_1\mathbb{E}(y|z = k)}{1 - p_k - \alpha_1}$$
$$\mu^1_k = \frac{(\mu_{1k} - \mu_{0k}) + (r_{1k}\mu_{0k} - r_{0k}\mu_{1k})}{r_{1k} - r_{0k}} = \frac{p_k\mathbb{E}(y|T = 1, z = k) - \alpha_0\mathbb{E}(y|z = k)}{p_k - \alpha_0}.$$

Since $\mu^0_k = \mu^0_{0k} = \mu^0_{1k}$ and $\mu^1_k = \mu^1_{0k} = \mu^1_{1k}$ under non-differential measurement error, the mis-classification probabilities (α_0, α_1) combined with the observable moments completely determine the means of F^0_{tk} and F^1_{tk} whenever the determinant condition holds. If $\mu_{0k} = \mu_{1k}$ then $\mu^0_k = \mu^1_k$ so we are free to set $F^0_{0k} = F^1_{0k} = F_{0k}$ and $F^0_{1k} = F^1_{1k} = F_{1k}$. Combining this with the reasoning from the preceding paragraph, we see that Assumption 2.2 (iii) imposes *no additional restrictions* for any k such that $\mu_{0k} = \mu_{1k}$. Accordingly, for the remainder of the proof we consider only the case in which $\mu_{0k} \neq \mu_{1k}$. Given (α_0, α_1), r_{tk}, μ^0_k, and μ^1_k are fixed. The question is whether, for a given pair (α_0, α_1) and observed CDFs F_{tk}, we can construct valid CDFs F^0_{tk}, F^1_{tk} such that

$$\int_{\mathbb{R}} \tau F^0_{tk}(d\tau) = \mu^0_k, \quad \int_{\mathbb{R}} \tau F^1_{tk}(d\tau) = \mu^1_k, \quad F_{tk}(\tau) = r_{tk}F^1_{tk}(\tau) + (1 - r_{tk})F^0_{tk}(\tau).$$

For a given pair (t, k), there are two cases: $0 < r_{tk} < 1$ and $r_{tk} \in \{0, 1\}$.

Case I: $r_{tk} \in \{0, 1\}$ If $r_{tk} = 1$ then $\mu^1_k = \mu_{tk}$ so we can set $F^1_{tk} = F_{tk}$. In this case F^0_{tk} is unrestricted. Analogously, if $r_{tk} = 0$, $\mu^0_k = \mu_{tk}$ so we can set $F^0_{tk} = F_{tk}$ with F^1_{tk} unrestricted.

Case II: $0 < r_{tk} < 1$ Define the function $\mu_{tk}(\xi) = \mathbb{E}[y|y \in I_{tk}(\xi), T = t, z = k]$ and the closed interval $I_{tk}(\xi) = \left[F_{tk}^{-1}(1-\xi-r_{tk}), F_{tk}^{-1}(1-\xi)\right]$ where $0 \leq \xi \leq 1 - r_{tk}$. The function μ_{tk} is decreasing in ξ, attaining its maximum $\overline{\mu}_{tk}$ at $\xi = 0$ and its minimum $\underline{\mu}_{tk}$ at $\xi = 1 - r_{tk}$.

Suppose first that μ_k^1 does *not* lie in the interval $[\underline{\mu}_{tk}, \overline{\mu}_{tk}]$. We show that it is impossible to construct valid CDFs F_{tk}^0 and F_{tk}^1 that satisfy $F_{tk}(\tau) = r_{tk}F_{tk}^1(\tau) + (1 - r_{tk})F_{tk}^0(\tau)$. Since $r_{tk} \neq 1$, we can solve the expression for F_{tk} to yield $F_{tk}^0(\tau) = \left[F_{tk}(\tau) - r_{tk}F_{tk}^1(\tau)\right]/(1 - r_{tk})$. Hence, since $r_{tk} \neq 0$, the requirement that $0 \leq F_{tk}^0(\tau) \leq 1$ implies

$$\frac{F_{tk}(\tau) - (1 - r_{tk})}{r_{tk}} \leq F_{tk}^1(\tau) \leq \frac{F_{tk}(\tau)}{r_{tk}} \tag{A.5}$$

Now define $\underline{F}_{tk}^1(\tau) = \min\{1, F_{tk}(\tau)/r_{tk}\}$ and $\overline{F}_{tk}^1(\tau) = \max\{0, F_{tk}(\tau)/r_{tk} - (1 - r_{tk})/r_{tk}\}$. By combining Equation A.5 with $0 \leq F_{tk}^1(\tau) \leq 1$, we obtain $\overline{F}_{tk}^1(\tau) \leq F_{tk}^1(\tau) \leq \underline{F}_{tk}^1(\tau)$. Thus, $\overline{F}_{tk}^1$ first-order stochastically dominates F_{tk}^1 which first-order stochastically dominates $\underline{F}_{tk}^1$. Hence,

$$\int \tau \underline{F}_{tk}^1(d\tau) \leq \int \tau F_{tk}^1(d\tau) \leq \int \tau \overline{F}_{tk}^1(d\tau).$$

But notice that

$$\underline{\mu}_{tk} = \int \tau \underline{F}_{tk}^1(d\tau), \quad \mu_k^1 = \int \tau F_{tk}^1(d\tau), \quad \overline{\mu}_{tk} = \int \tau \overline{F}_{tk}^1(d\tau)$$

so we have $\underline{\mu}_{tk} \leq \mu_k^1 \leq \overline{\mu}_{tk}$ which contradicts $\mu_k^1 \notin [\underline{\mu}_{tk}, \overline{\mu}_{tk}]$.

Now suppose that $\mu_k^1 \in \left[\underline{\mu}_{tk}, \overline{\mu}_{tk}\right]$. We show how to construct densities f_{tk}^1 and f_{tk}^0 that yield CDFs F_{tk}^0 F_{tk}^1 satisfying the requirements described above. Since the conditional distribution of y given (T, z) is continuous, μ_{tk} is continuous on its domain and takes on all values in $\left[\underline{\mu}_{tk}, \overline{\mu}_{tk}\right]$ by the intermediate value theorem. Thus, there exists a ξ^* such that $\mu_{tk}(\xi^*) = \mu_k^1$. Let $f_{tk}(\tau) = dF_{tk}(\tau)/d\tau$ which is non-negative by the assumption that y is continuously distributed. Now, define

$$f_{tk}^1(\tau) = \frac{f_{tk}(\tau) \times \mathbf{1}\{\tau \in I_{tk}(\xi^*)\}}{r_{tk}}, \quad f_{tk}^0(\tau) = \frac{f_{tk}(\tau) \times \mathbf{1}\{\tau \in I_{tk}(\xi^*)\}}{1 - r_{tk}}.$$

Clearly $f_{tk}^1 \geq 0$ and $f_{tk}^0 \geq 0$. Integrating,

$$\int_{\mathbb{R}} f_{tk}^1(\tau)\, d\tau = \frac{1}{r_{tk}} \int_{I_{tk}(\xi^*)} f_{tk}(\tau)\, d\tau = 1, \quad \int_{\mathbb{R}} f_{tk}^0(\tau)\, d\tau = \frac{1}{1 - r_{tk}} \int_{I_{tk}^C(\xi^*)} f_{tk}(\tau)\, d\tau = 1$$

where I_{tk}^C is the complement of I_{tk}. By construction

$$r_{tk} \int_A f_{tk}^1(\tau)\, d\tau + (1 - r_{tk}) \int_A f_{tk}^0(\tau)\, d\tau = \int_A f_{tk}(\tau)\, d\tau$$

for any set A. Finally,

$$\int_{\mathbb{R}} \tau f_{tk}^1(\tau)\, d\tau = \frac{1}{r_{tk}} \int_{I_{tk}(\xi^*)} \tau f_{tk}(\tau)\, d\tau = \mu_{tk}(\xi^*) = \mu_k^1.$$

□

A.2 Point Identification Results

In the proofs of Lemma 2.3, Lemma 2.4, and Theorem 2.3, we employ the shorthand $\pi \equiv \text{Cov}(T, z)$, $\eta_j \equiv \text{Cov}(y^j, z)$, and $\tau_j \equiv \text{Cov}(Ty^j, z)$ for $j = 1, 2, 3$. Hence Lemma 2.2 becomes $\eta_1 = \pi\theta_1$, while Lemma 2.3 becomes $\eta_2 = 2\tau_1\theta_1 - \pi\theta_2$, and Lemma 2.4 becomes $\eta_3 = 3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3$.

Proof of Lemma 2.3. By Assumption 2.1 (i) and the basic properties of covariance,

$$\begin{aligned}\eta_2 &= \beta^2\text{Cov}(T^*, z) + 2\beta\left[c\,\text{Cov}(T^*, z) + \text{Cov}(T^*\varepsilon, z)\right] + 2c\,\text{Cov}(\varepsilon, z) + \text{Cov}(\varepsilon^2, z)\\ \tau_1 &= c\pi + \text{Cov}(T\varepsilon, z) + \beta\text{Cov}(TT^*, z)\end{aligned}$$

using the fact that T^* is binary. Now, by Assumptions 2.1 (iii) and 2.5 we have $\text{Cov}(\varepsilon, z) = \text{Cov}(\varepsilon^2, z) = 0$. And, using Assumptions 2.2 (i) and (ii), one can show that $\text{Cov}(TT^*, z) = (1 - \alpha_1)\text{Cov}(T^*, z)$ and $\text{Cov}(T^*, z) = \pi/(1 - \alpha_0 - \alpha_1)$. Hence,

$$\begin{aligned}\eta_2 &= \theta_1\left(\beta + 2c\right)\pi + 2\beta\text{Cov}(T^*\varepsilon, z)\\ 2\tau_1\theta_1 - \pi\theta_2 &= \left[2\theta_1 c + 2\theta_1^2(1 - \alpha_1) - \theta_2\right]\pi + 2\theta_1\text{Cov}(T\varepsilon, z)\end{aligned}$$

but since $\theta_2 = \theta_1^2\left[(1 - \alpha_1) + \alpha_0\right]$, we see that $\left[2\theta_1^2(1 - \alpha_1) - \theta_2\right] = \theta_1\beta$. Thus, it suffices to show that $\beta\text{Cov}(T^*\varepsilon, z) = \theta_1\text{Cov}(T\varepsilon, z)$. This equality is trivially satisfied when $\beta = 0$, so suppose that $\beta \neq 0$. In this case it suffices to show that $(1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon, z) = \text{Cov}(T\varepsilon, z)$. Define $m^*_{tk} = \mathbb{E}\left[\varepsilon | T^* = t, z = k\right]$ and $p^*_k = \mathbb{P}(T^* = 1 | z = k)$. Then, by iterated expectations, Bayes' rule, and Assumption 2.2 (iii)

$$\begin{aligned}\text{Cov}(T^*\varepsilon, z) &= q(1 - q)\left(p^*_1 m^*_{11} - p^*_0 m^*_{10}\right)\\ \text{Cov}(T\varepsilon, z) &= q(1 - q)\left\{(1 - \alpha_1)\left[p^*_1 m^*_{11} - p^*_0 m^*_{10}\right] + \alpha_0\left[(1 - p^*_1)m^*_{01} - (1 - p^*_0)m^*_{00}\right]\right\}\end{aligned}$$

But by Assumption 2.1 (iii), $\mathbb{E}[\varepsilon | z = k] = m^*_{1k}p^*_k + m^*_{0k}(1 - p^*_k) = 0$ and thus we obtain $m^*_{0k}(1 - p^*_k) = -m^*_{1k}p^*_k$. Therefore $(1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon, z) = \text{Cov}(T\varepsilon, z)$ as required. □

Proof of Lemma 2.4. Since T^* is binary, if follows from the basic properties of covariance that,

$$\begin{aligned}\eta_3 &= \text{Cov}\left[(c + \varepsilon)^3, z\right] + 3\beta\text{Cov}[(c + \varepsilon)^2 T^*, z] + 3\beta^2\text{Cov}[(c + \varepsilon)T^*, z] + \beta^3\text{Cov}(T^*, z)\\ \tau_2 &= \text{Cov}\left[(c + \varepsilon)^2 T, z\right] + 2\beta\text{Cov}\left[(c + \varepsilon)TT^*, z\right] + \beta^2\text{Cov}(TT^*, z)\end{aligned}$$

By Assumptions 2.1 (iii), 2.5, and 2.6 (ii) , $\text{Cov}\left[(c + \varepsilon)^3, z\right] = 0$. Expanding,

$$\begin{aligned}\eta_3 &= 3\beta\text{Cov}(T^*\varepsilon^2, z) + \left(3\beta^2 + 6c\beta\right)\text{Cov}(T^*\varepsilon, z) + \left(\beta^3 + 3c\beta^2 + 3c^2\beta\right)\text{Cov}(T^*, z)\\ \tau_2 &= c^2\text{Cov}(T, z) + \beta(\beta + 2c)\text{Cov}(TT^*, z) + \text{Cov}(T\varepsilon^2, z) + 2c\text{Cov}(T\varepsilon, z) + 2\beta\,\text{Cov}(TT^*\varepsilon, z)\end{aligned}$$

Now, define $s^*_{tk} = \mathbb{E}[\varepsilon^2 | T^* = t, z = k]$ and $p^*_k = \mathbb{P}(T^* = 1 | z = k)$. By iterated expectations, Bayes' rule, and Assumption 2.6 (i),

$$\begin{aligned}\text{Cov}(T^*\varepsilon^2, z) &= q(1 - q)(p^*_1 s^*_{11} - p^*_0 s^*_{10})\\ \text{Cov}(T\varepsilon^2, z) &= q(1 - q)\left\{(1 - \alpha_1)\left[p^*_1 s^*_{11} - p^*_0 s^*_{10}\right] + \alpha_0\left[(1 - p^*_1)s^*_{01} - (1 - p^*_0)s^*_{00}\right]\right\}\end{aligned}$$

By Assumption 2.5, $\mathbb{E}[\varepsilon^2 | z = 1] = \mathbb{E}[\varepsilon^2 | z = 0]$ and thus, by iterated expectations we have $p^*_1 s^*_{11} - p^*_0 s^*_{10} = -\left[(1 - p^*_1)s^*_{01} - (1 - p^*_0)s^*_{00}\right]$ which implies

$$\text{Cov}(T\varepsilon^2, z) = (1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon^2, z). \tag{A.6}$$

Similarly by iterated expectations and Assumptions 2.2 (i)–(ii)

$$\text{Cov}(TT^*\varepsilon, z) = q(1-q)(1-\alpha_1)(p_1^* m_{1k}^* - p_0^* m_{10}^*) = (1-\alpha_1)\text{Cov}(T^*\varepsilon, z) \tag{A.7}$$

where m_{tk}^* is defined as in the proof of Lemma 2.3. As shown in the proof of Lemma 2.3,

$$\text{Cov}(TT^*, z) = (1-\alpha_1)\text{Cov}(T^*, z), \quad \text{Cov}(T^*, z) = \frac{\pi}{1-\alpha_0-\alpha_1}, \quad \text{Cov}(T^*\varepsilon, z) = \frac{\text{Cov}(T\varepsilon, z)}{1-\alpha_0-\alpha_1}$$

and combining these equalities with Equations A.6 and A.7, it follows that

$$\begin{aligned}
\tau_2 &= 2\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right]\text{Cov}(T^*\varepsilon, z) + \left[(1-\alpha_1)(c+\beta)^2 - c^2\alpha_0\right]\text{Cov}(T^*, z) \\
&\quad + (1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon^2, z) \\
\tau_1 &= (1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon, z) + \left[(1-\alpha_1)(c+\beta) - c\alpha_0\right]\text{Cov}(T^*, z)
\end{aligned}$$

using $\tau_1 = c\pi + \text{Cov}(T\varepsilon, z) + \beta\text{Cov}(TT^*, z)$ as shown in the proof of Lemma 2.3. Thus,

$$3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3 = K_1\text{Cov}(T^*\varepsilon^2, z) + K_2\text{Cov}(T^*\varepsilon, z) + K_3\text{Cov}(T^*, z)$$

where $K_1 \equiv 3\theta_1(1-\alpha_0-\alpha_1) = 3\beta$ and

$$\begin{aligned}
K_2 &\equiv 6\theta_1\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] - 3\theta_2(1-\alpha_0-\alpha_1) \\
K_3 &\equiv 3\theta_1\left[(1-\alpha_1)(c+\beta)^2 - c^2\alpha_0\right] - 3\theta_2\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] + \theta_3(1-\alpha_0-\alpha_1)
\end{aligned}$$

Substituting the definitions of θ_1, θ_2, and θ_3 from Equations 4–6, tedious but straightforward algebra shows that $K_2 = 3\beta^2 + 6c\beta$ and $K_3 = \beta^3 + 3c\beta^2 + 3c^2\beta$. Therefore the coefficients of η_3 equal those of $3\tau_2 - 3\tau_1\theta_2 + \pi\theta_3$ and the result follows. □

Proof of Theorem 2.3. Collecting the results of Lemmas 2.2–2.4, we have

$$\eta_1 = \pi\theta_1, \quad \eta_2 = 2\tau_1\theta_1 - \pi\theta_2, \quad \eta_3 = 3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3$$

which is a linear system in $\theta_1, \theta_2, \theta_3$ with determinant $-\pi^3$. Since $\pi \neq 0$ by assumption 2.1 (ii), θ_1, θ_2 and θ_3 are identified. Now, so long as $\beta \neq 0$, we can rearrange Equations 5 and 6 to obtain

$$A = \theta_2/\theta_1^2 = 1 + (\alpha_0 - \alpha_1) \tag{A.8}$$

$$B = \theta_3/\theta_1^3 = (1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1) \tag{A.9}$$

Equation A.8 gives $(1-\alpha_1) = A - \alpha_0$. Hence $(1-\alpha_0-\alpha_1) = A - 2\alpha_0$ and $\alpha_0(1-\alpha_1) = \alpha_0(A-\alpha_0)$. Substituting into Equation A.9 and simplifying, $(A^2 - B) + 2A\alpha_0 - 2\alpha_0^2 = 0$. Substituting for α_0 analogously yields a quadratic in $(1-\alpha_1)$ with *identical* coefficients. It follows that one root of $(A^2 - B) + 2Ar - 2r^2 = 0$ is α_0 and the other is $1-\alpha_1$. Solving,

$$r = \frac{A}{2} \pm \sqrt{3A^2 - 2B} = \frac{1}{\theta_1^2}\left(\frac{\theta_2}{2} \pm \sqrt{3\theta_2^2 - 2\theta_1\theta_3}\right). \tag{A.10}$$

Substituting Equations 5 and 6, simple algebra shows that $3\theta_2^2 - 2\theta_1\theta_3 = \theta_1^2(1-\alpha_0-\alpha_1)^2$. This quantity is strictly greater than zero since $\theta_1 \neq 0$ and $\alpha_0 + \alpha_1 \neq 1$. It follows that both roots of the quadratic are real. Moreover, $3\theta_2^2/\theta_1^4 - 2\theta_3/\theta_1^3$ identifies $(1-\alpha_0-\alpha_1)^2$. Substituting into Equation 4, it follows that β is identified up to sign. If $\alpha_0 + \alpha_1 < 1$ then $\text{sign}(\beta) = \text{sign}(\theta_1)$ so that both the

sign and magnitude of β are identified. If $\alpha_0 + \alpha_1 < 1$ then $1 - \alpha_1 > \alpha_0$ so $(1 - \alpha_1)$ is the larger root of $(A^2 - B) + 2Ar - 2r^2 = 0$ and α_0 is the smaller root. □

B Comment on Mahajan (2006) A.2

Expanding on our discussion from Section 2.2 above, we now show that Mahajan's identification argument for an endogenous regressor in an additively separable model (A.2) is incorrect. Unless otherwise indicated, all notation used below is as defined in Section 2.

The first step of Mahajan (2006) A.2 argues (correctly) that under Assumptions 2.1 and 2.2 (i)–(ii), knowledge of $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ is sufficient to identify $\beta(\mathbf{x})$. This step is equivalent to our Lemma 2.2 above. The second step appeals to Mahajan (2006) Theorem 1 to argue that $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are indeed point identified. To understand the logic of this second step, we first re-state Mahajan (2006) Theorem 1 in our notation. As in Section 2 above, T^* denotes an unobserved binary random variable, z is a instrument, T an observed binary surrogate for T^*, y an outcome of interest, and $\mathbf{x}$ a vector covariates.

Assumption B.1 (Mahajan (2006) Theorem 1)**.** *Define $g(T^*, \mathbf{x}) \equiv \mathbb{E}[y|\mathbf{x}, T^*]$ and $v \equiv y - g(T^*, \mathbf{x})$. Suppose that knowledge of $(y, T^*, \mathbf{x})$ is sufficient to identify g and that:*

(i) $\mathbb{P}(T^* = 1|\mathbf{x}, z = 0) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 1)$.

(ii) T *is conditionally independent of* z *given* $(\mathbf{x}, T^*)$.

(iii) $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$

(iv) $\mathbb{E}[v|\mathbf{x}, z, T^*, T] = 0$

(v) $g(1, \mathbf{x}) \neq g(0, \mathbf{x})$

Theorem B.1 (Mahajan (2006) Theorem 1)**.** *Under Assumption B.1, $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are point identified, as is $g(T^*, \mathbf{x})$.*

Assumption B.1 (i) is equivalent to our Assumption 2.1 (ii), while Assumptions B.1 (ii)–(iii) are equivalent to our Assumptions 2.2 (i)–(ii). Assumption B.1 (v) serves the same purpose as $\beta(\mathbf{x}) \neq 0$ in our Theorem 2.3: unless T^* affects y, we cannot identify the mis-classification probabilities. The key difference between Theorem B.1 and the setting we consider in Section 2 comes from Assumption B.1 (iv). This is essentially a stronger version of our Assumptions 2.1 (iii) and 2.2 (iii) but applies to the *projection error* v, defined in Assumption B.1 rather than the structural error ε, defined in Assumption 2.1 (i). Accordingly, Theorem B.1 identifies the conditional mean function g rather than the causal effect $\beta(\mathbf{x})$.

Although the meaning of the error term changes when we move from a structural to a reduced form model, the meaning of the mis-classification error rates does not: $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are simply conditional probabilities for T given $(T^*, \mathbf{x})$. Step 2 of Mahajan (2006) A.2 relies on this insight. The idea is to find a way to satisfy Assumption B.1 (iv) simultaneously with Assumptions 2.1 (iii) and 2.2 (iii), while allowing T^* to be endogenous. If this can be achieved, $\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})$ will be identified via Theorem B.1, and identification of $\beta(\mathbf{x})$ will follow from step 1 of A.2 (our Lemma 2.2). To this end, Mahajan (2006) invokes the condition

$$\mathbb{E}(y|\mathbf{x}, z, T^*, T) = \mathbb{E}(y|\mathbf{x}, T^*). \tag{B.1}$$

Because Mahajan (2006) A.2 assumes an additively separable model – our Assumption 2.1 (i) – we see that

$$\mathbb{E}(y|\mathbf{x}, z, T^*, T) = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T)$$

so Equation B.1 is equivalent to $\mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T) = \mathbb{E}(\varepsilon|\mathbf{x}, T^*)$. Note that this allows T^* to be endogenous, as it does not require $\mathbb{E}(\varepsilon|\mathbf{x}, T^*) = 0$. Now, applying Equation B.1 to the definition of v from Assumption B.1, we have

$$\mathbb{E}(v|\mathbf{x}, z, T^*, T) = \mathbb{E}\left[y - \mathbb{E}(y|\mathbf{x}, T^*) \mid \mathbf{x}, z, T^*, T\right] = 0$$

which satisfies Assumption B.1 (iv) as required. Based on this reasoning, Mahajan (2006) claims that Equation B.1 along with Assumptions B.1 (iv), 2.1, and 2.2 (i)–(ii) suffice to identify the effect $\beta(\mathbf{x})$ of an endogenous T^*, so long as $g(1, \mathbf{x}) \neq g(0, \mathbf{x})$. As we now show, however, these Assumptions are contradictory unless T^* is exogenous.

By Equation B.1 and Assumption 2.1 (i), $\mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T) = \mathbb{E}(\varepsilon|\mathbf{x}, T^*)$ and thus by iterated expectations, we obtain

$$\mathbb{E}(\varepsilon|\mathbf{x}, T^*, z) = \mathbb{E}_{T|\mathbf{x},T^*,z}\left[\mathbb{E}(\varepsilon|\mathbf{x}, T^*, T, z)\right] = \mathbb{E}_{T|\mathbf{x},T^*,z}\left[\mathbb{E}(\varepsilon|\mathbf{x}, T^*)\right] = \mathbb{E}(\varepsilon|\mathbf{x}, T^*). \tag{B.2}$$

Now, let $m^*_{tk}(\mathbf{x}) = \mathbb{E}(\varepsilon|\mathbf{x}, T^* = t, z = k)$. Using this notation, Equation B.2 is equivalent to $m^*_{t0}(\mathbf{x}) = m^*_{t1}(\mathbf{x})$ for $t = 0, 1$. Combining iterated expectations with Assumption 2.1 (iii),

$$\mathbb{E}(\varepsilon|\mathbf{x}, z = k) = [1 - p^*_k(\mathbf{x})]m^*_{0k}(\mathbf{x}) + p^*_k(\mathbf{x})m^*_{1k}(\mathbf{x}) = 0 \tag{B.3}$$

for $k = 0, 1$ where $p^*_k(\mathbf{x}) \equiv \mathbb{P}(T^* = 1|\mathbf{x}, z = k)$. But substituting $m^*_{t0}(\mathbf{x}) = m^*_{t1}(\mathbf{x})$ into Equation B.3 for $k = 0, 1$, we obtain

$$\begin{aligned}[1 - p^*_0(\mathbf{x})]m^*_{00}(\mathbf{x}) + p^*_0(\mathbf{x})m^*_{10}(\mathbf{x}) = 0\\ [1 - p^*_1(\mathbf{x})]m^*_{00}(\mathbf{x}) + p^*_1(\mathbf{x})m^*_{10}(\mathbf{x}) = 0\end{aligned}$$

The preceding two equalities are convex combinations of m^*_{00} and m^*_{10}. The only way that both can equal zero simultaneously is if either $p^*_0(\mathbf{x}) = p^*_1(\mathbf{x})$, contradicting Assumption 2.1 (ii), or if $m^*_{tk}(\mathbf{x}) = 0$ for all (t, k), which implies that T^* is exogenous. Hence Mahajan (2006) A.2 fails: given the assumption that z is a valid instrument for ε, Equation B.1 implies that either there is no first-stage relationship between z and T^* or that T^* is exogenous. The root of the problem with A.2 is the attempt to use *one* instrument to satisfy both the assumptions of Theorem B.1 and Lemma 2.2. If one had access to a second instrument w, or equivalently a second mis-measured surrogate for T^*, that satisfied Assumptions B.1, one could use w to recover $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ via Theorem B.1 and z to recover the IV estimand $\beta(\mathbf{x})/[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]$ via Lemma 2.2.

C Unobserved Heterogeneity

While allowing for arbitrary observed heterogeneity through the covariates $\mathbf{x}$, all of the results presented above assume an additively separable model – Assumption 2.1 (i). In this section we briefly discuss how our partial identification results can be interpreted in a local average treatment effects (LATE) setting. For simplicity, we suppress explicit conditioning on the covariates $\mathbf{x}$ throughout.

In lieu of Assumption 2.1 (i), consider a non-separable model of the form $y = h(T^*, z, \varepsilon)$. Let $T^*(z)$ denote an individual's potential treatment and $Y(t^*, z)$ denote her potential outcome, where $t^*, z \in \{0, 1\}$. Using this notation we can write $Y(t^*, z) = h(t^*, z, \varepsilon)$. Let $J \in \{a, c, d, n\}$ index the

four LATE principal strata: a = always-taker, c = complier, d = defier, and n = never-taker. If $J = a$, then $T^*(z) = 1$; if $J = c$, then $T^*(z) = z$; if $J = d$, then $T^*(z) = 1 - z$; and if $J = n$, then $T^*(z) = 0$. In a LATE model, Assumption 2.1 (iii) is replaced by the standard LATE assumptions:

Assumption C.1 (Unconfounded Type)**.** $\mathbb{P}(J = j|z = 1) = \mathbb{P}(J = j|z = 0)$ *for all* $j \in \{a, c, d, n\}$.

Assumption C.2 (Mean Exclusion Restriction)**.** *For all* $t^* \in \{0, 1\}$ *and* $j \in \{a, c, d, n\}$,

$$\mathbb{E}\left[Y(t^*, 0)|T^* = t^*, z = 1\right] = \mathbb{E}\left[Y(t^*, 1)|T^* = t^*, z = 1\right] = \mathbb{E}[Y(t^*)|J = j].$$

Assumption C.3 (Monotonicity)**.** $\mathbb{P}\big(T^*(1) \geq T^*(0)\big) = 1$

As is well known, Assumption 2.1 (iii) combined with the preceding three conditions implies that the instrumental variables estimand based on T^* identifies the average treatment effect among compliers:

$$\frac{\mathbb{E}[y|z = 1] - \mathbb{E}[y|z = 0]}{p_1^* - p_0^*} = \mathbb{E}[Y(1) - Y(0)|J = c].$$

The numerator of the preceding expression is observed, but under mis-classification the denominator is not. Notice, however, that Assumptions 2.2 (i)–(ii) only concern the joint distribution of T given (T^*, z). As such, they have the same meaning in a LATE model as in an additively separable model. Imposing these conditions, Lemma 2.1 continues to hold in a LATE model. It follows that $p_1 - p_0 = (1 - \alpha_0 - \alpha_1)(p_1^* - p_0^*)$ so that

$$\frac{\mathbb{E}[y|z = 1] - \mathbb{E}[y|z = 0]}{p_1 - p_0} = \frac{\mathbb{E}[Y(1) - Y(0)|J = c]}{1 - \alpha_0 - \alpha_1}.$$

Moreover, $\alpha_0 \leq p_k \leq 1 - \alpha_1$ for all k. Thus, the bound from Corollary 2.1 remains valid in a LATE model: $\mathbb{E}[Y(1) - Y(0)|J = c]$ must lie between the IV and reduced form estimands.

Unlike Assumptions 2.2 (i)–(ii), Assumption 2.2 (iii), non-differential measurement error, is explicitly stated in terms of the unobservable error term in an additively separable model. Our derivation of the additional restrictions on (α_0, α_1) implied by non-differential measurement error in the proof of Theorem 2.2, however, does not use Assumption 2.2 (iii) directly. Rather, it uses a condition that is *equivalent* to it in an additively separable model, namely $\mathbb{E}[Y|T^*, T, z] = \mathbb{E}[Y|T^*, z]$. Hence, as long as this equality holds, regardless of whether one is in an additively separable model or a LATE model, the bounds on (α_0, α_1) from Theorem 2.2 remain valid. Since $Y = (1 - T^*)Y(0) + T^*Y(1)$, the appropriate modification of Assumption 2.2 (iii) is as follows.

Assumption C.4 (Non-differential Measurement Error)**.**

$$\mathbb{E}[Y(0)|T^*, T, z] = \mathbb{E}[Y(0)|T^*, z] \quad \textit{and} \quad \mathbb{E}[Y(1)|T^*, T, z] = \mathbb{E}[Y(1)|T^*, z]$$

To summarize, if one wishes to re-interpret our parameter β as a local average treatment effect, the partial identification bounds from Theorems 2.1 and 2.2 above remain valid. Assumption 2.1 (i) is replaced by $Y = h(T^*, z, \varepsilon)$, Assumption 2.1 (iii) is replaced by Assumptions C.1–C.3, and Assumption 2.2 (iii) is replaced by Assumption C.4. In a LATE model, however, our proofs of sharpness no longer apply, as they do not consider the testable implications of the LATE assumptions themselves. For partial identification results that consider these implications but do not impose non-differential measurement error, see Ura (Forthcoming). For discussion of the testable implications of a LATE model, see Kitagawa (2015).

References

Aigner, D. J., 1973. Regression with a binary independent variable subject to errors of observation. Journal of Econometrics 1, 49–60.

Andrews, D. W., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78 (1), 119–157.

Angrist, J. D., 1990. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. The American Economic Review, 313–336.

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

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, 387–399.

Bollinger, C. R., van Hasselt, M., 2015. Bayesian moment-based inference in a regression models with misclassification error, working Paper.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Handbook of econometrics. Vol. 5. Elsevier, pp. 3705–3843.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Stefanski, L. A., 2006. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC.

Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. The Review of Economic Studies 72 (2), 343–366.

Chen, X., Hu, Y., Lewbel, A., 2008a. Nonparametric identification of regression models containing a misclassified dichotomous regressor with instruments. Economics Letters 100, 381–384.

Chen, X., Hu, Y., Lewbel, A., 2008b. A note on the closed-form identification of regression models with a mismeasured binary regressor. Statistics & Probability Letters 78 (12), 1473–1479.

DiTraglia, F. J., García-Jimeno, C., 2017. Mis-classified, binary, endogenous regressors: Identification and inference. Tech. rep., NBER working paper #23814.

Feng, S., Hu, Y., 2013. Misclassification errors and the underestimation of the us unemployment rate. American Economic Review 103 (2), 1054–70.

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

Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. Journal of Econometrics 144 (1), 27–61.

Hu, Y., Lewbel, A., 2012. Returns to lying? identifying the effects of misreporting when the truth is unobserved. Frontiers of Economics in China 7 (2), 163–192.

Hu, Y., Shennach, S. M., January 2008. Instrumental variable treatment of nonclassical measurement error models. Econometrica 76 (1), 195–216.

Hu, Y., Shiu, J.-L., Woutersen, T., 2015. Identification and estimation of single-index models with measurement error and endogeneity. The Econometrics Journal 18 (3), 347–362.

Imbens, G. W., Rubin, D. B., 1997. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies 64 (4), 555–574.

Kane, T. J., Rouse, C. E., Staiger, D., July 1999. Estimating returns to schooling when schooling is misreported. Tech. rep., National Bureau of Economic Research, NBER Working Paper 7235.

Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.

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.

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

Lewbel, A., 2007b. A local generalized method of moments estimator. Economics Letters 94, 124–128.

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

Molinari, F., 2008. Partial identification of probability distributions with misclassified data. Journal of Econometrics 144 (1), 81–117.

Nguimkeu, P., Denteh, A., Tchernis, R., 2016. On the estimation of treatment effects with endogenous misreporting. Working Paper.

Shiu, J.-L., 2016. Identification and estimation of endogenous selection models in the presence of misclassification errors. Economic Modelling 52 (Part B), 507–518.

Song, S., 2015. Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors. Journal of Econometrics 185 (1), 95–109.

Ura, T., Forthcoming. Heterogeneous treatment effects with mismeasured endogenous treatment. Quantitative Economics.

van Hasselt, M., Bollinger, C. R., 2012. Binary misclassification and identification in regression models. Economics Letters 115, 81–84.