Mis-classified, Binary, Endogenous Regressors: Identification and Inference*

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

[1]Department of Economics, University of Pennsylvania
[2]NBER

This Version: August 24, 2017, 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, and find that these 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. We then turn our attention to inference. We show that both our model, and related models from the literature that assume regressor exogeneity, suffer from weak identification when the effect of interest is small. To address this difficulty, we exploit the inequality restrictions that emerge from our derivation of the sharp identified set under mean independence only. These restrictions remain informative irrespective of the strength of identification. Combining these with the moment equalities that emerge from our identification result, we propose a robust inference procedure using tools from the moment inequality literature. Our method performs well in simulations, both for the exogenous and endogenous regressor case.

Keywords: Instrumental variables, Measurement error, Endogeneity, Weak identification, Moment inequalities

JEL Codes: C10, C18, C25, C26

*We thank Daron Acemoglu, Manuel Arellano, Kristy Buzard, Xu Cheng, Bernardo da Silveira, Bo Honoré, Sophocles Mavroeidis, Yuya Takahashi, and seminar participants at Cambridge, CEMFI, Chicago Booth, Manchester, Northwestern, Oxford, Penn State, Princeton, UCL, 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

Additively separable model

$$y = h(T^*, \mathbf{x}) + \varepsilon$$

where ε is a mean-zero error term, T^* is an endogenous binary regressor of interest and $\mathbf{x}$ is a vector of exogenous controls. Since T^* is binary, we can re-write this as linear in T^* conditional on $\mathbf{x}$

$$\begin{aligned} y &= c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon \\ \beta(\mathbf{x}) &= h(1, \mathbf{x}) - h(0, \mathbf{x}) \\ c(\mathbf{x}) &= h(0, \mathbf{x}) \end{aligned}$$

Goal is to use an instrumental variable z to identify $\beta(\mathbf{x})$ when we observe not T^* but a mis-measured binary surrogate T.

How are we different from Ura?

1. We maintain nondifferential measurement error assumption throughout; Ura's main purpose is to relax it.
2. We focus on an additively separable model; Ura explicitly studies at LATE setting
3. We obtain point and partial identification results; Ura present only partial identification results

Old Introduction: Many treatments of interest in applied work are binary. To take a particularly prominent example, consider treatment status in a randomized controlled trial. Even if the randomization is pristine, which yields a valid binary instrument (the offer of treatment), subjects may select into treatment based on unobservables, and given the many real-world complications that arise in the field, measurement error may be an important concern. This paper studies the use of a discrete instrumental variable to identify the causal effect of an endogenous, mis-measured, binary treatment in a model with additively separable errors. Specifically, we consider the following model

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

where $T^* \in \{0, 1\}$ is a mis-measured, endogenous treatment, $\mathbf{x}$ is a vector of exogenous controls, and ε is a mean-zero error. Since T^* is potentially endogenous, $\mathbb{E}[\varepsilon|T^*, \mathbf{x}]$ may not be zero. Our goal is to non-parametrically estimate the average treatment effect (ATE) function

$$\tau(\mathbf{x}) = h(1, \mathbf{x}) - h(0, \mathbf{x}). \tag{2}$$

using a single discrete instrumental variable $z \in \{z_k\}_{k=1}^K$. We assume throughout that z is a relevant instrument for T^*, in other words

$$\mathbb{P}(T^* = 1|z_j, \mathbf{x}) \neq \mathbb{P}(T^* = 1|z_k, \mathbf{x}), \quad \forall k \neq j. \tag{3}$$

While the structural relationship involves T^*, we observe only a noisy measure T, polluted by non-differential measurement error. In particular, we assume that

$$\mathbb{P}(T = 1|T^* = 0, z, \mathbf{x}) = \alpha_0(\mathbf{x}) \tag{4}$$

$$\mathbb{P}(T = 0|T^* = 1, z, \mathbf{x}) = \alpha_1(\mathbf{x}) \tag{5}$$

where the mis-classification error rates $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ can depend on $\mathbf{x}$ but not z, and additionally that, conditional on true treatment status, observed treatment status provides no additional information about the error term. In other words, we assume that

$$\mathbb{E}[\varepsilon|T^*, T, z, \mathbf{x}] = \mathbb{E}[\varepsilon|T^*, z, \mathbf{x}]. \tag{6}$$

Although a relevant case for applied work, the setting we consider here has received little attention in the literature. The only existing result for the case of an endogenous treatment appears in an important paper by Mahajan (2006), who is primarily concerned with the case of an exogenous treatment. As we show below, Mahajan's identification result for the endogenous treatment case is incorrect. As far as we are aware, this leaves the problem considered in this paper completely unsolved.

We begin by showing that the proof in Appendix A.2 of Mahajan (2006) leads to a contradiction. Throughout his paper, Mahajan (2006) maintains an assumption (Assumption 4) which he calls the "Dependency Condition." This assumption requires that the instrumental variable be relevant, namely that it generates variation in true treatment status. When extending his result for an exogenous treatment to the more general case of an endogenous one, however, he must impose an additional condition on the model (Equation 11), which turns out to violate the Dependency Condition. Since one cannot impose the condition in Equation 11 of Mahajan (2006), we go on to study the prospects for identification in this model more broadly. We consider two possibilities. First, since Mahajan's identification results require only a binary instrument, we borrow an idea from Lewbel (2007) and explore whether expanding the support of the instrument yields identification based on moment equations similar to those used by Mahajan (2006). While allowing the instrument to take on additional values does increase the number of available moment conditions, we show that these moments cannot point identify the treatment effect, regardless of how many (finite) values the instrument takes on.

We then consider a new source of identifying information that arises from imposing stronger assumptions on the instrumental variable. If the instrument is not merely mean independent but in fact *statistically independent* of the regression error term, as in a randomized controlled trial or a true natural experiment, additional moment conditions become available. We show that adding a conditional second moment independence assumption on the instrument identifies the *difference* of mis-classification rates $\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})$. Because these rates must equal each other when there is no mis-classification error, our result can be used to test a necessary condition for the absence of measurement error. It can also be used to construct simple and informative partial identification bounds for the treatment effect. When one of the mis-classification rates is known, this identifies the treatment effect. More generally, however, this is not the case. We go on to show that a conditional third moment independence assumption on the instrument point identifies both $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ and hence the ATE function $\tau(\mathbf{x})$. Both our point identification and partial identification results require only a binary instrument, and lead to simple, closed-form method of moments estimators.

This project is still in progress. The present draft focuses on establishing identification in the simplest possible way: by directly solving a set of equations implied by conditional moment restrictions on ε. Additional results regarding efficient estimation and sharp bounds for α_0, α_1 under weaker conditions on the instrumental variable are currently in progress. For some additional discussion of these results, see our conclusion below in Section 4.

The remainder of this paper is organized as follows. In section 2 we discuss the literature in relation to the problem considered here. Section 3 introduces notation and assumptions, and presents our main results. Section 4 concludes. All proofs appear in the Appendix.

Old Literature Review: Measurement error is a pervasive feature of economic data, motivating a long tradition of measurement error modelling in econometrics. The textbook case considers a continuous regressor (treatment) subject to classical measurement error in a linear model. In this setting, the measurement error is assumed to be unrelated to the true, unobserved, value of the treatment of interest. Regardless of whether this unobserved treatment is exogenous or endogenous, a single valid instrument suffices to identify its effect. When an instrument is unavailable, Lewbel (1997) shows that higher moment assumptions can be used to construct one, provided that the mis-measured treatment is exogenous. When it is endogenous, Lewbel (2012) uses a heteroskedasticity assumption to obtain identification.

Departures from the linear, classical measurement error setting pose serious identification challenges. One strand of the literature considers relaxing the assumption of linearity while maintaining that of classical measurement error. Schennach (2004), for example, uses repeated measures of each mis-measured treatment to obtain identification, while Schennach (2007) uses an instrumental variable. Both papers consider the case of exogenous treatments.[1] More recently, Song et al. (2015) rely on a repeated measure of the mis-measured treatment and the existence of a set of additional regressors, conditional upon which the treatment of interest is unrelated to the unobservables, to obtain identification. Another strand of the literature considers relaxing the assumption of classical measurement error, by allowing the measurement error to be related to the true value of the unobserved treatment. 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.

Many treatments of interest in economics, however, are binary, and in this case classical measurement error is impossible. Because a true 1 can only be mis-measured as a 0 and a true 0 can only be mis-measured as a 1, the measurement error must be *negatively* correlated with the true treatment status (Aigner, 1973; Bollinger, 1996). For this reason, even in a textbook linear model, the instrumental variables estimator can only remove the effect of endogeneity, not that of measurement error (Frazis and Loewenstein, 2003). Measurement error in a discrete variable is usually called mis-classification.[2] The simplest form of mis-classification is so-called *non-differential* measurement error. In this case, conditional on true treatment status, and possibly a set of exogenous covariates, the measurement error is assumed to be unrelated to all other variables in the system.

A number of papers have studied this problem without the use of instrumental variables under the assumption that the mis-measured binary treatment is exogenous. The first to address this problem was Aigner (1973), who characterized the asymptotic bias of the OLS estimator in this setting, and proposed a technique for correcting it using outside information

[1] For comprehensive reviews of the challenges of addressing measurement error in non-linear models, see Chen et al. (2011) and Schennach (2013).

[2] For general results on the partial identification of discrete probability distributions using mis-classified observations, see Molinari (2008).

on the mis-classification process. Another early contribution by Bollinger (1996) provides partial identification bounds. More recently, Chen et al. (2008a) use higher moment assumptions to obtain identification in a linear regression 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.

Continuing under the assumption of an exogenous treatment, a number of other papers in the literature have considered the identifying power of an instrumental variable, or something like one. Black et al. (2000) and Kane et al. (1999) more-or-less simultaneously pointed out that when *two* alternative measures of treatment are available, both subject to non-differential measurement error, a non-linear GMM estimator can be used to recover the treatment effect. In essence, one measure serves as an instrument for the other although the estimator is quite different from IV.[3] Subsequently, Frazis and Loewenstein (2003) correctly note that an instrumental variable can take the place of one of the measures of treatment in a linear model with an exogenous treatment, allowing one to implement a variant of the GMM estimator proposed by Black et al. (2000) and Kane et al. (1999). However, as we will show below, the assumptions required to obtain this result are stronger than Frazis and Loewenstein (2003) appear to realize: the usual IV assumption that the instrument is mean independent of the regression error is insufficient for identification.

Mahajan (2006) extends the results of Black et al. (2000) and Kane et al. (1999) to a more general nonparametric regression setting using a binary instrument in place of one of the treatment measures. Although unaware of Frazis and Loewenstein (2003), Mahajan (2006) makes the correct assumption over the instrument and treatment to guarantee identification of the conditional mean function. When the treatment is in fact exogenous, this coincides with the treatment effect. Hu (2008) derives related results when the mis-classified discrete regressor may take on more than two values. Lewbel (2007) provides an identification result for the same model as Mahajan (2006) under different assumptions. In particular, the variable that plays the role of the "instrument" need not satisfy the exclusion restriction provided that it does not interact with the treatment and takes on at least three distinct values.

Much less is known about the case in which a binary, or discrete, treatment is not only mis-measured but endogenous. Frazis and Loewenstein (2003) briefly discuss the prospects for identification in this setting. Although they do not provide a formal proof they argue, in the context of their parametric linear model, that the treatment effect is unlikely to be

[3] Ignoring covariates, the observable moments in this case are the joint probability distribution of the two binary treatment measures and the conditional means of the outcome variable given the two measures. Although the system is highly non-linear, it can be manipulated to yield an explicit solution for the treatment effect provided that the true treatment is exogenous.

identified unless one is willing to impose strong and somewhat unnatural conditions.[4] 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, Mahajan's proof is incorrect.

The results we derive here most closely relate to the setting considered in Mahajan (2006) in that we study non-parametric identification of the effect of a binary, endogenous treatment, using a discrete instrument. Unlike Mahajan (2006) we consider and indeed show the necessity of using higher-moment information to identify the causal effect of interest. Unlike Kreider et al. (2012), who partially identify the effects of food stamps on health outcomes of children under weak measurement error assumptions, we do not rely on auxiliary data. Unlike Shiu (2015), who considers a sample selection model with a discrete, mis-measured, endogenous regressor, we do not rely on a parametric assumption about the form of the first-stage. Finally unlike Ura (2015), who studies local average treatment effects under very general forms of mis-classification but presents only partial identification results, we point identify an average treatment effect under non-differential measurement error. Moreover, unlike the identification strategies from the existing literature described above, we do not rely upon continuity of the instrument, a large support condition, or restrictions on the relationship between the true, unmeasured treatment and its observed surrogate, subject to the condition that the measurement error process is non-differential.

2 Identification Results

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

[4] For example, one could consider using the results of Hausman et al. (1998), who study regressions with a mis-classified, discrete *outcome* variable, as a first-stage in an IV setting. In principle, this approach would fully identify the mis-classification error process. Using these results, however, requires either an explicit, nonlinear, parametric model for the first stage, or an identification at infinity argument.

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

Assumptions 2.1(ii) and (iii) are the standard instrument relevance and mean independence assumptions.[5] If T^* were observed, Assumption 2.1 would suffice to identify $\beta(\mathbf{x})$. Unfortunately we observe not T^* but a mis-classified binary surrogate T. Define the following mis-classification probabilities:

$$\alpha_0(\mathbf{x}, z) = \mathbb{P}\left(T = 1|T^* = 0, \mathbf{x}, z\right), \quad \alpha_1(\mathbf{x}, z) = \mathbb{P}\left(T = 0|T^* = 1, \mathbf{x}, z\right). \tag{7}$$

Following the existing literature for the case of an exogenous regressor (Black et al., 2000; Frazis and Loewenstein, 2003; Kane et al., 1999; Lewbel, 2007; Mahajan, 2006), we impose the following conditions on the mis-classification process.

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

Assumption 2.2 (i) states that the mis-classification probabilities do not depend on z. As we maintain this assumption throughout, we drop the dependence of α_0 and α_1 on z and write $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. 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 $(T^*, z, \mathbf{x})$.

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)

[5]Point out that even though Assumption 2.1 (ii) refers to the unobserved T^*, under Assumptions 2.2 (i) and (ii) we have $(p_k^* - p_\ell^*)(1 - \alpha_0 - \alpha_1) = p_k - p_\ell$ so it suffices for an *observed* first-stage to exist.

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.

2.3 Partial Identification

In this section we derive the sharp identified set for under Assumptions 2.1–2.2 and show that $\beta(\mathbf{x})$ is not point identified. To simplify the notation, define the following shorthand for the observed and unobserved first stage probabilities

$$p_k^*(\mathbf{x}) = \mathbb{P}(T^* = 1|\mathbf{x}, z = k) \tag{8}$$
$$p_k(\mathbf{x}) = \mathbb{P}(T = 1|\mathbf{x}, z = k). \tag{9}$$

We first state two lemmas that have appeared in various guises throughout the literature. These will be used repeatedly below.

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

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

where the first-stage probabilities $p_k^*(\mathbf{x})$ and $p_k(\mathbf{x})$ are as defined in Equations 8–9.

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 provides non-trivial bounds for $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ in terms of the observed first-stage probabilities. Lemma 2.2 relates the instrumental variables (IV) estimand, $\text{Cov}(y, z|\mathbf{x})/\text{Cov}(z, T)$, 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. Combining the two lemmas yields a well-known bound, namely that $\beta(\mathbf{x})$ lies between the reduced form and IV estimators. Our first result shows that *without* Assumption 2.2 (non-differential measurement error) these bounds are sharp.

Theorem 2.1. *Under Assumptions 2.1 and 2.2 (i)–(ii), the sharp identified set is characterized by*

$$\mathbb{E}[y|\mathbf{x}, z=k] = c(\mathbf{x}) + \beta(\mathbf{x})\left[\frac{p_k(\mathbf{x})-\alpha_0(\mathbf{x})}{1-\alpha_0(\mathbf{x})-\alpha_1(\mathbf{x})}\right] \tag{10}$$

and $\alpha_0(\mathbf{x}) \leq p_k(\mathbf{x}) \leq 1-\alpha_1(\mathbf{x})$ *for* $k=0,1$ *where* $p_k(\mathbf{x})$ *is defined in Equation 9.*

Corollary 2.1. *Under the conditions of Theorem 2.1 the sharp identified set for* $\beta(\mathbf{x})$ *is the closed interval between* $\Delta^y(\mathbf{x})$ *and* $\Delta^y(\mathbf{x})/\Delta^T(\mathbf{x})$ *where* $\Delta^y(\mathbf{x}) \equiv \mathbb{E}(y|\mathbf{x}, z=1) - \mathbb{E}(y|\mathbf{x}, z=0)$ *and* $\Delta^T(\mathbf{x}) = p_1(\mathbf{x}) - p_0(\mathbf{x})$*, with* $p_k(\mathbf{x})$ *as defined in Equation 9 for* $k=0,1$*.*

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. When the mis-classification probabilities are known *a priori* to satisfy additional restrictions, these bounds can be tightened.[6] The following corollary collects results for two common cases: one-sided misclassification (either $\alpha_0(\mathbf{x})$ or $\alpha_1(\mathbf{x})$ equals zero), and symmetric mis-classification ($\alpha_0(\mathbf{x}) = \alpha_1(\mathbf{x})$).

Corollary 2.2. *Under the conditions of Theorem 2.1, restrictions on the misclassification probabilities shrink the sharp identified set for* $\beta(\mathbf{x})$ *to the closed interval between* $B\Delta^y(\mathbf{x})/\Delta^T(\mathbf{x})$ *and* $\Delta^y(\mathbf{x})/\Delta^T(\mathbf{x})$ *where*

(i) $\alpha_0(\mathbf{x}) = 0$ *implies* $B = \max_k \mathbb{P}(T=1|\mathbf{x}, z=k)$

(ii) $\alpha_1(\mathbf{x}) = 0$ *implies* $B = 1 - \min_k \mathbb{P}(T=1|\mathbf{x}, z=k)$

(iii) $\alpha_0(\mathbf{x}) = \alpha_1(\mathbf{x})$ *implies* $B = 1 - 2\min\left\{\min_k \mathbb{P}(T=1|\mathbf{x}, z=k), 1-\max_k \mathbb{P}(T=1|\mathbf{x}, z=k)\right\}$

for $k=0,1$*, where* $\Delta^T(\mathbf{x})$ *and* $\Delta^y(\mathbf{x})$ *are as defined in Corollary 2.1.*

Theorem 2.1 and Corollaries 2.1–2.2 do not impose Assumption 2.2 (iii) – non-differential measurement error. We now show that this assumption yields further restrictions on the mis-classification probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. While these restrictions are more complicated to describe than those from Theorem 2.1, they are straightforward to implement in practice and can be extremely informative, as we will show in our simulation exercises below. To the best

[6] Frazis and Loewenstein (2003) consider a model in which α_0 and α_1 do not depend on the exogenous covariates $\mathbf{x}$. In this case $\alpha_0 \leq \mathbb{P}(T=1|\mathbf{x}, z) \leq 1-\alpha_1$ and they suggest minimizing the bounds over $\mathbf{x}$.

of our knowledge, the sharp bounds that we derive by adding Assumption 2.2 (iii) are new to the literature. Our result uses two additional conditions to simplify the proof of sharpness. First, we assume that y is continuously distributed. This is natural in an additively separable model and holds in our simulation examples below. 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. We also impose $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ for any k. This holds generically and is not essential to the proof: it merely simplifies the description of the identified set.

Theorem 2.2. *Suppose that the conditional distribution of y given $(\mathbf{x}, T, z)$ is continuous for any values of the conditioning variables and $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ for all k. Then, under Assumptions 2.1 and 2.2, the sharp identified set is characterized by Equation 10 from Theorem 2.1 along with $\alpha_0(\mathbf{x}) < p_k(\mathbf{x}) < 1 - \alpha_1(\mathbf{x})$ for $k = 0, 1$ and*

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

for all pairs (t, k) where

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

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

and we define

$$\begin{aligned}
\underline{q}_{tk}(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}) &= F_{tk}^{-1}\left(r_{tk}(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}) \,\middle|\, \mathbf{x}\right) \\
\overline{q}_{tk}(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}) &= F_{tk}^{-1}\left(1 - r_{tk}(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}) \,\middle|\, \mathbf{x}\right)
\end{aligned}$$

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

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

and $p_k(\mathbf{x})$ is defined in Equation 9.

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 $(T = 1, z = k, T^* = 1)$, call it F^1_{tk}, and the distribution of y given $(T = t, z = k, T^* = 0)$, call it F^0_{tk}. 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^1_{tk} and F^0_{tk} with the required values for $\mathbb{E}[y|T^*, T, z]$ such that $F_{tk} = r_{tk}F^1_{tk} + (1 - r_{tk})F^0_{tk}$ 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 in practice, they do not yield point identification.

Corollary 2.3. *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 2.3 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 the support of z. 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. If z takes on more than two values, our results in Theorems 2.1–2.2 continue to apply if "$k = 0, 1$" is replaced by "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:

$$\theta_1(\mathbf{x}) = \beta(\mathbf{x})\left[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]^{-1} \tag{11}$$

$$\theta_2(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^2\left[1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right] \tag{12}$$

$$\theta_3(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^3\left[\left\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right\}^2 + 6\alpha_0(\mathbf{x})\left\{1 - \alpha_1(\mathbf{x})\right\}\right] \tag{13}$$

along with the observable quantities

$$\pi(\mathbf{x}) = \mathrm{Cov}(T, z|\mathbf{x}), \quad \eta_j(\mathbf{x}) = \mathrm{Cov}(y^j, z|\mathbf{x}), \quad \tau_j(\mathbf{x}) = \mathrm{Cov}(Ty^j, z|\mathbf{x}) \tag{14}$$

for $j = 1, 2, 3$. Using this notation, Lemma 2.2 can be written as $\eta_1(\mathbf{x}) = \pi(\mathbf{x})\theta_1(\mathbf{x})$. 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. Notice that this assumption 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,* $\eta_2(\mathbf{x}) = 2\tau_1(\mathbf{x})\theta_1(\mathbf{x}) - \pi(\mathbf{x})\theta_2(\mathbf{x})$*, where* $\pi(\mathbf{x}), \tau_1(\mathbf{x}), \eta_1(\mathbf{x})$ *and* $\eta_2(\mathbf{x})$ *are defined in Equation 14, and* $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ *in Equations 11–12.*

Lemma 2.2 identifies $\theta_1(\mathbf{x})$. Since $\pi(\mathbf{x}) \neq 0$ 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 12 for the difference of mis-classification rates so long as $\beta(\mathbf{x}) \neq 0$.

Corollary 2.4. *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 2.4 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 2.4 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 unfamiliar, we consider them to be fairly natural in the context of an additively separable model in which one has already assumed that $\mathbb{E}[\varepsilon|z] = 0$ and $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\mathbf{x}, z, T^*]$ – Assumptions 2.1 (iii) and 2.2 (iii) from above.[7] For example, if an applied researcher reports results both for an outcome in logs and levels, she has implicitly assumed *independence* rather than first moment exclusion. Assumptions 2.1 (iii), 2.5 and 2.6 (ii) are of course implied by $\varepsilon \perp z|\mathbf{x}$ while Assumptions 2.2 (iii) and 2.6 (i) are implied by $\varepsilon \perp T|(\mathbf{x}, T^*, z)$. Of course, achieving identification via Assumptions 2.5–2.6 will involve using information beyond first moments and as such places higher demands on the data. 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,*

$$\eta_3(\mathbf{x}) = 3\tau_2(\mathbf{x})\theta_1(\mathbf{x}) - 3\tau_1(\mathbf{x})\theta_2(\mathbf{x}) + \pi(\mathbf{x})\theta_3(\mathbf{x})$$

where $\pi(\mathbf{x}), \tau_1(\mathbf{x}), \eta_1(\mathbf{x}), \eta_2(\mathbf{x})$, *and* $\eta_3(\mathbf{x})$ *are defined in Equation 14, and* $\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \theta_3(\mathbf{x})$ *are defined in Equations 11–12.*

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.*

[7] If one wishes to weaken our Assumption 2.1 (i) to allow for some form of unobserved heterogeneity, our higher moment assumptions may impose additional restrictions. We discuss this issue further in Section 2.5

Lemmas 2.2–2.4 yield a linear system of three equations in three unknowns. 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 11–13 can be solved for $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. 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.

2.5 Unobserved Heterogeneity

Although our results allow arbitrary observed heterogeneity, additive separability places restrictions on unobserved heterogeneity. Although it is not the main focus of our paper, unlike Ura, briefly comment on how these results can be interpreted, say, in a LATE context. First, the bounds stuff all goes through (???) provided one is willing to make the LATE assumptions. Higher moment restrictions do impose restrictions. Can say what happens with the second moment assumption since we already derived this: it's a restriction on the variance of the potential outcome distributions, conditional on $\mathbf{x}$.

3 Identification-Robust Inference

4 Simulation Study

5 Coverage and Width of Confidence Intervals

5.1 Endogenous Regressor

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	90	90	90	91	90	91	90	90
	0.1	91	93	94	94	94	94	90	89
	0.2	92	93	94	94	94	94	92	90
	0.3	93	93	94	94	94	93	92	91
0.1	0.0	92	93	93	94	94	93	90	87
	0.1	93	95	96	97	97	96	92	87
	0.2	95	96	97	98	97	96	92	87
	0.3	96	98	98	98	98	95	92	88
0.2	0.0	93	93	93	93	93	93	92	89
	0.1	95	96	98	98	97	95	93	89
	0.2	97	97	98	98	97	95	92	89
	0.3	98	98	98	98	97	95	93	91
0.3	0.0	93	94	94	94	94	93	92	91
	0.1	97	97	98	98	97	95	93	89
	0.2	98	98	98	98	97	94	93	91
	0.3	99	99	99	98	98	96	95	94

Table 1: Coverage (1 - size) of 90% GMS joint test for α_0 and α_1: $n = 1000$. Based on 10000 simulation replications.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	90	90	90	91	90	91	90	90
	0.1	91	93	94	94	94	94	90	89
	0.2	92	93	94	94	94	94	92	90
	0.3	93	93	94	94	94	93	92	91
0.1	0.0	92	93	93	94	94	93	90	87
	0.1	93	95	96	97	97	96	92	87
	0.2	95	96	97	98	97	96	92	87
	0.3	96	98	98	98	98	95	92	88
0.2	0.0	93	93	93	93	93	93	92	89
	0.1	95	96	98	98	97	95	93	89
	0.2	97	97	98	98	97	95	92	89
	0.3	98	98	98	98	97	95	93	91
0.3	0.0	93	94	94	94	94	93	92	91
	0.1	97	97	98	98	97	95	93	89
	0.2	98	98	98	98	97	94	93	91
	0.3	99	99	99	98	98	96	95	94

Table 2: Coverage (1 - size) of 90% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	95	95	95	96	96	96	95	95
	0.1	96	97	97	97	97	97	95	94
	0.2	96	97	98	98	97	97	96	95
	0.3	97	97	97	98	97	97	96	95
0.1	0.0	96	97	97	97	97	97	95	93
	0.1	97	98	99	99	99	98	96	92
	0.2	98	99	99	99	99	98	96	93
	0.3	99	99	99	99	99	98	96	94
0.2	0.0	97	97	97	97	97	96	96	94
	0.1	98	99	99	99	99	98	96	94
	0.2	99	99	99	99	99	98	96	94
	0.3	99	100	100	99	99	98	97	95
0.3	0.0	97	97	97	97	97	96	96	95
	0.1	99	99	99	99	99	98	97	94
	0.2	99	99	99	99	99	98	97	96
	0.3	100	100	100	99	99	98	98	97

Table 3: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	95	95	95	96	96	96	95	95
	0.1	96	97	97	97	97	97	95	94
	0.2	96	97	98	98	97	97	96	95
	0.3	97	97	97	98	97	97	96	95
0.1	0.0	96	97	97	97	97	97	95	93
	0.1	97	98	99	99	99	98	96	92
	0.2	98	99	99	99	99	98	96	93
	0.3	99	99	99	99	99	98	96	94
0.2	0.0	97	97	97	97	97	96	96	94
	0.1	98	99	99	99	99	98	96	94
	0.2	99	99	99	99	99	98	96	94
	0.3	99	100	100	99	99	98	97	95
0.3	0.0	97	97	97	97	97	96	96	95
	0.1	99	99	99	99	99	98	97	94
	0.2	99	99	99	99	99	98	97	96
	0.3	100	100	100	99	99	98	98	97

Table 4: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.7	98.0	98.0	97.4	97.9
	0.1	98.0	98.7	98.8	99.1	98.8	98.4	97.1	96.4
	0.2	98.4	98.5	98.9	98.9	98.8	98.6	98.0	97.0
	0.3	98.5	98.8	98.8	99.0	98.7	98.4	97.8	97.5
0.1	0.0	98.1	98.5	98.3	98.8	98.8	98.4	96.8	95.7
	0.1	98.6	99.1	99.5	99.6	99.6	98.8	97.7	95.2
	0.2	99.0	99.3	99.7	99.8	99.7	98.9	97.5	95.7
	0.3	99.4	99.7	99.8	99.8	99.6	99.0	98.2	96.7
0.2	0.0	98.6	98.5	98.6	98.9	98.7	98.2	97.7	97.0
	0.1	99.0	99.5	99.7	99.7	99.4	99.0	98.1	96.5
	0.2	99.5	99.7	99.8	99.7	99.4	99.0	97.8	96.8
	0.3	99.7	99.8	99.8	99.8	99.5	99.0	98.7	97.7
0.3	0.0	98.7	98.7	98.8	98.7	98.7	98.2	98.1	97.6
	0.1	99.4	99.6	99.6	99.7	99.4	98.9	98.3	96.8
	0.2	99.8	99.8	99.7	99.8	99.5	99.1	98.5	97.8
	0.3	100.0	99.9	99.9	99.8	99.6	99.5	99.1	98.8

Table 5: Coverage (1 - size) of 97.5% GMS joint test for α_0 and α_1: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.7	98.0	98.0	97.4	97.9
	0.1	98.0	98.7	98.8	99.1	98.8	98.4	97.1	96.4
	0.2	98.4	98.5	98.9	98.9	98.8	98.6	98.0	97.0
	0.3	98.5	98.8	98.8	99.0	98.7	98.4	97.8	97.5
0.1	0.0	98.1	98.5	98.3	98.8	98.8	98.4	96.8	95.7
	0.1	98.6	99.1	99.5	99.6	99.6	98.8	97.7	95.2
	0.2	99.0	99.3	99.7	99.8	99.7	98.9	97.5	95.7
	0.3	99.4	99.7	99.8	99.8	99.6	99.0	98.2	96.7
0.2	0.0	98.6	98.5	98.6	98.9	98.7	98.2	97.7	97.0
	0.1	99.0	99.5	99.7	99.7	99.4	99.0	98.1	96.5
	0.2	99.5	99.7	99.8	99.7	99.4	99.0	97.8	96.8
	0.3	99.7	99.8	99.8	99.8	99.5	99.0	98.7	97.7
0.3	0.0	98.7	98.7	98.8	98.7	98.7	98.2	98.1	97.6
	0.1	99.4	99.6	99.6	99.7	99.4	98.9	98.3	96.8
	0.2	99.8	99.8	99.7	99.8	99.5	99.1	98.5	97.8
	0.3	100.0	99.9	99.9	99.8	99.6	99.5	99.1	98.8

Table 6: Coverage (1 - size) of 97.5% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	27	33	30	14	1	0	0	0
	0.1	27	32	29	13	2	0	0	0
	0.2	26	33	32	15	4	0	0	0
	0.3	26	34	30	17	5	0	0	0
0.1	0.0	26	32	31	14	2	0	0	0
	0.1	26	36	32	16	4	0	0	0
	0.2	27	35	31	18	8	0	0	0
	0.3	25	35	32	21	11	1	0	0
0.2	0.0	26	33	30	15	3	0	0	0
	0.1	26	33	30	19	6	0	0	0
	0.2	26	35	33	22	12	1	0	0
	0.3	26	35	33	26	15	3	0	0
0.3	0.0	26	32	32	16	6	0	0	0
	0.1	24	35	33	21	11	1	0	0
	0.2	26	32	35	27	15	4	0	0
	0.3	26	35	35	28	21	7	2	0

Table 7: Percentage of simulation replications for which the standard GMM confidence interval fails to exist, either becuase the point estimate is NaN or the asymptotic covariance matrix is numerically singular ($n = 1000$). Based on 2000 simulation replications.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	25	36	29	7	0	0	0	0
	0.1	28	36	29	7	0	0	0	0
	0.2	28	37	28	10	1	0	0	0
	0.3	27	36	28	12	2	0	0	0
0.1	0.0	27	36	27	10	0	0	0	0
	0.1	26	36	29	9	1	0	0	0
	0.2	28	38	29	13	2	0	0	0
	0.3	24	36	31	15	5	0	0	0
0.2	0.0	26	36	30	9	1	0	0	0
	0.1	25	37	29	12	2	0	0	0
	0.2	27	38	32	17	4	0	0	0
	0.3	25	39	34	20	9	1	0	0
0.3	0.0	26	37	30	10	2	0	0	0
	0.1	25	38	31	16	4	0	0	0
	0.2	27	38	34	19	9	0	0	0
	0.3	27	36	36	23	13	2	0	0

Table 8: Percentage of simulation replications for which the standard GMM confidence interval fails to exist, either becuase the point estimate is NaN or the asymptotic covariance matrix is numerically singular ($n = 2000$)

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	100	95	92	93	94	95	94	95
	0.1	100	94	91	93	94	95	96	95
	0.2	99	94	92	92	94	96	96	96
	0.3	99	94	92	92	94	96	96	95
0.1	0.0	99	95	92	92	94	95	96	96
	0.1	100	94	91	93	94	95	95	94
	0.2	100	93	91	93	94	95	95	94
	0.3	99	93	91	91	93	95	96	96
0.2	0.0	99	94	90	92	94	95	96	96
	0.1	100	93	91	92	93	95	96	94
	0.2	100	93	90	92	92	95	95	95
	0.3	100	93	90	91	93	95	95	96
0.3	0.0	100	94	91	92	95	95	96	96
	0.1	99	94	91	92	93	94	96	95
	0.2	99	93	91	92	93	94	96	96
	0.3	99	93	90	92	92	94	95	96

Table 9: Coverage of nominal 95% GMM Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	100	90	91	94	95	94	96	95
	0.1	100	90	91	93	95	95	95	96
	0.2	100	90	92	94	95	95	95	94
	0.3	100	90	92	93	95	95	95	94
0.1	0.0	100	90	92	92	94	95	94	96
	0.1	100	91	92	93	94	95	95	95
	0.2	100	90	92	92	95	96	95	95
	0.3	99	89	90	92	95	95	95	95
0.2	0.0	100	90	91	93	94	96	95	94
	0.1	99	91	92	93	94	96	95	96
	0.2	100	90	92	92	95	96	96	95
	0.3	99	90	91	93	95	95	96	96
0.3	0.0	100	90	91	93	94	97	95	95
	0.1	100	90	91	94	94	95	96	96
	0.2	99	90	91	92	94	96	96	96
	0.3	100	89	90	92	94	95	96	96

Table 10: Coverage of nominal 95% GMM Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	83.99	7.45	3.01	1.52	0.88	0.47	0.37	0.35
	0.1	85.51	7.2	3.01	1.62	1.01	0.61	0.51	0.46
	0.2	74.92	7.79	3.21	1.72	1.13	0.76	0.65	0.58
	0.3	76.66	8.02	3.19	1.78	1.29	0.91	0.79	0.7
0.1	0.0	76.59	7.46	3.2	1.59	0.99	0.61	0.51	0.46
	0.1	78.46	8.07	3.21	1.72	1.17	0.78	0.67	0.6
	0.2	77.79	7.9	3.26	1.95	1.33	0.97	0.85	0.75
	0.3	65.63	8.2	3.5	2.13	1.59	1.18	1.04	0.92
0.2	0.0	69.39	7.52	3.26	1.7	1.14	0.75	0.65	0.58
	0.1	81.48	7.79	3.27	1.95	1.34	0.97	0.84	0.75
	0.2	79.96	7.94	3.58	2.16	1.64	1.21	1.06	0.95
	0.3	85.95	8.14	3.7	2.54	1.96	1.53	1.33	1.19
0.3	0.0	87.95	7.44	3.17	1.84	1.31	0.9	0.79	0.7
	0.1	72.15	8.01	3.45	2.15	1.61	1.18	1.04	0.92
	0.2	67.84	7.6	3.75	2.63	2	1.55	1.35	1.19
	0.3	84.13	8.47	4.23	3.07	2.55	1.98	1.77	1.55

Table 11: Median Width of nominal 95% GMM Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	74.47	5.17	2.16	1.06	0.62	0.33	0.27	0.24
	0.1	86.02	5.07	2.25	1.13	0.7	0.43	0.36	0.33
	0.2	87.06	5.28	2.3	1.24	0.81	0.53	0.46	0.41
	0.3	83.3	5.08	2.33	1.34	0.92	0.65	0.56	0.5
0.1	0.0	71.27	5.38	2.24	1.14	0.71	0.43	0.36	0.33
	0.1	64.58	4.95	2.41	1.25	0.83	0.56	0.48	0.43
	0.2	80.03	5.31	2.39	1.38	0.98	0.69	0.6	0.53
	0.3	70.37	4.78	2.54	1.55	1.14	0.84	0.73	0.65
0.2	0.0	65.65	4.99	2.44	1.23	0.81	0.54	0.46	0.41
	0.1	71.25	5.31	2.49	1.35	0.98	0.69	0.6	0.54
	0.2	83.96	5.57	2.63	1.61	1.17	0.86	0.76	0.67
	0.3	75.88	5.83	2.88	1.85	1.44	1.09	0.95	0.85
0.3	0.0	74.62	5.23	2.41	1.32	0.92	0.65	0.56	0.5
	0.1	76.36	5.69	2.54	1.57	1.15	0.84	0.74	0.65
	0.2	91.87	5.44	2.96	1.82	1.42	1.08	0.96	0.85
	0.3	73.3	5.17	3.16	2.24	1.85	1.43	1.24	1.1

Table 12: Median Width of nominal 95% GMM Intervals with $n = 2000$

α_0	α_1	β = 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	96
	0.1	97	99	99	99	99	100	100	99
	0.2	98	99	99	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.1	0.0	97	99	99	99	100	100	100	98
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.2	0.0	97	99	99	100	100	100	100	100
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.3	0.0	97	99	100	100	100	100	100	100
	0.1	97	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100

Table 13: Coverage of nominal $> 95\%$ Bonferroni Intervals with $n = 1000$

α_0	α_1	β = 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	96	97	96	96	95	95
	0.1	97	98	99	100	100	100	100	99
	0.2	97	99	99	100	100	100	100	100
	0.3	97	99	100	100	100	100	100	100
0.1	0.0	97	99	99	99	100	100	100	99
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.2	0.0	97	99	99	100	100	100	100	99
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.3	0.0	97	100	100	100	100	100	100	100
	0.1	97	100	100	100	100	100	100	100
	0.2	97	100	100	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100

Table 14: Coverage of nominal $> 95\%$ Bonferroni Intervals with $n = 2000$

α_0	α_1	β 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.41	0.41
	0.1	0.45	0.47	0.54	0.59	0.63	0.7	0.75	0.86
	0.2	0.51	0.54	0.65	0.76	0.85	0.95	1.01	1.17
	0.3	0.58	0.62	0.79	0.95	1.07	1.17	1.24	1.48
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.7	0.76	0.88
	0.1	0.51	0.54	0.66	0.77	0.86	1.03	1.18	1.46
	0.2	0.58	0.63	0.8	0.98	1.12	1.38	1.55	1.88
	0.3	0.67	0.75	1	1.25	1.46	1.74	1.94	2.4
0.2	0.0	0.51	0.54	0.65	0.76	0.86	0.96	1.02	1.19
	0.1	0.58	0.63	0.81	0.99	1.14	1.42	1.64	2.08
	0.2	0.67	0.75	1.01	1.29	1.54	1.97	2.33	2.9
	0.3	0.81	0.91	1.3	1.7	2.09	2.73	3.13	3.9
0.3	0.0	0.58	0.62	0.8	0.95	1.09	1.18	1.25	1.5
	0.1	0.68	0.74	1.01	1.26	1.49	1.84	2.13	2.78
	0.2	0.81	0.91	1.3	1.7	2.11	2.8	3.4	4.48
	0.3	1.01	1.16	1.74	2.35	2.93	4.17	5.2	6.85

Table 15: Median Width of nominal $> 95\%$ Bonferroni Intervals with $n = 1000$

α_0	α_1	β 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.29	0.3	0.31	0.31	0.31	0.3	0.29	0.29
	0.1	0.32	0.35	0.4	0.44	0.48	0.53	0.55	0.61
	0.2	0.36	0.41	0.51	0.59	0.65	0.67	0.69	0.81
	0.3	0.41	0.48	0.64	0.76	0.81	0.8	0.85	1.01
0.1	0.0	0.32	0.35	0.4	0.44	0.48	0.53	0.56	0.62
	0.1	0.36	0.41	0.51	0.6	0.69	0.82	0.88	1.02
	0.2	0.41	0.48	0.64	0.79	0.91	1.04	1.08	1.27
	0.3	0.48	0.59	0.82	1.02	1.16	1.25	1.33	1.61
0.2	0.0	0.36	0.41	0.51	0.59	0.65	0.67	0.7	0.82
	0.1	0.41	0.48	0.65	0.79	0.92	1.09	1.21	1.52
	0.2	0.48	0.59	0.83	1.05	1.24	1.49	1.61	1.96
	0.3	0.57	0.73	1.09	1.43	1.69	1.9	2.08	2.6
0.3	0.0	0.41	0.48	0.64	0.77	0.82	0.78	0.84	1.02
	0.1	0.48	0.59	0.83	1.03	1.18	1.36	1.57	2.06
	0.2	0.57	0.73	1.1	1.43	1.71	2.11	2.45	3.18
	0.3	0.72	0.95	1.5	2.03	2.53	3.15	3.56	4.56

Table 16: Median Width of nominal $> 95\%$ Bonferroni Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	93
	0.1	97	99	99	99	99	98	96	95
	0.2	98	99	99	100	100	97	96	96
	0.3	97	100	100	100	99	96	96	96
0.1	0.0	97	99	99	99	100	98	97	95
	0.1	98	100	100	100	100	96	96	96
	0.2	98	100	100	100	99	96	96	95
	0.3	97	100	100	100	97	95	96	96
0.2	0.0	97	99	99	100	100	96	96	96
	0.1	98	100	100	100	99	96	96	96
	0.2	98	100	100	100	96	95	95	96
	0.3	98	100	100	98	95	95	95	96
0.3	0.0	97	99	100	100	100	95	96	97
	0.1	97	100	100	100	97	94	96	96
	0.2	98	100	100	98	94	94	96	96
	0.3	98	100	99	96	92	94	95	96

Table 17: Coverage of two-step CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	96	97	96	96	95	93
	0.1	97	98	99	100	100	98	97	96
	0.2	97	99	99	100	100	97	96	95
	0.3	97	99	100	100	99	96	96	96
0.1	0.0	97	99	99	99	100	98	96	95
	0.1	98	100	100	100	100	96	96	97
	0.2	98	100	100	100	99	96	96	97
	0.3	98	100	100	99	97	95	96	96
0.2	0.0	97	99	99	100	100	97	96	95
	0.1	98	100	100	100	98	96	96	97
	0.2	98	100	100	100	96	96	96	96
	0.3	98	100	100	97	95	95	96	96
0.3	0.0	97	100	100	100	99	98	97	96
	0.1	97	100	100	100	96	95	96	97
	0.2	97	100	100	97	94	96	96	97
	0.3	97	100	100	94	94	95	96	96

Table 18: Coverage of two-step CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 2000$

α_0	α_1	β 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.4	0.35
	0.1	0.45	0.47	0.54	0.59	0.63	0.67	0.52	0.46
	0.2	0.51	0.54	0.65	0.76	0.84	0.82	0.65	0.58
	0.3	0.58	0.62	0.79	0.95	1.05	0.96	0.79	0.7
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.67	0.51	0.46
	0.1	0.51	0.54	0.66	0.77	0.86	0.92	0.69	0.61
	0.2	0.58	0.63	0.8	0.97	1.11	1.17	0.87	0.75
	0.3	0.67	0.75	1	1.25	1.4	1.4	1.06	0.92
0.2	0.0	0.51	0.54	0.65	0.76	0.85	0.83	0.65	0.58
	0.1	0.58	0.63	0.81	0.99	1.12	1.18	0.86	0.75
	0.2	0.67	0.75	1.01	1.29	1.48	1.56	1.08	0.95
	0.3	0.81	0.91	1.3	1.67	1.95	1.77	1.35	1.2
0.3	0.0	0.58	0.62	0.8	0.95	1.07	0.95	0.8	0.7
	0.1	0.68	0.74	1.01	1.26	1.43	1.48	1.06	0.93
	0.2	0.81	0.91	1.3	1.66	1.98	1.94	1.37	1.19
	0.3	1.01	1.16	1.73	2.24	2.71	2.33	1.78	1.55

Table 19: Median width of two-step CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 1000$

α_0	α_1	β 0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.29	0.3	0.31	0.31	0.31	0.3	0.29	0.25
	0.1	0.32	0.35	0.4	0.44	0.48	0.48	0.36	0.33
	0.2	0.36	0.41	0.51	0.59	0.65	0.57	0.46	0.41
	0.3	0.41	0.48	0.64	0.76	0.79	0.68	0.56	0.5
0.1	0.0	0.32	0.35	0.4	0.44	0.48	0.48	0.37	0.33
	0.1	0.36	0.41	0.51	0.6	0.68	0.65	0.48	0.43
	0.2	0.41	0.48	0.64	0.78	0.89	0.83	0.61	0.54
	0.3	0.48	0.59	0.82	1.02	1.09	0.98	0.75	0.65
0.2	0.0	0.36	0.41	0.51	0.59	0.65	0.58	0.46	0.41
	0.1	0.41	0.48	0.65	0.79	0.9	0.89	0.61	0.54
	0.2	0.48	0.59	0.83	1.05	1.2	1.22	0.77	0.67
	0.3	0.57	0.73	1.09	1.4	1.58	1.53	0.97	0.85
0.3	0.0	0.41	0.48	0.64	0.77	0.8	0.69	0.56	0.5
	0.1	0.48	0.59	0.83	1.02	1.13	1.19	0.75	0.65
	0.2	0.57	0.73	1.1	1.4	1.62	1.79	0.97	0.85
	0.3	0.72	0.95	1.49	1.93	2.36	1.58	1.25	1.1

Table 20: Median width of two-step CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 2000$

5.2 Exogenous Regressor

6 Conclusion

A Proofs

Lemma A.1 (Lemma for Appendix only with Bayes' Rule)**.** *For mis-classification probabilities*

$$
\begin{aligned}
P(T^* = 1|T = 1, Z = k) &= P(T = 1|T^* = 1)\left(\frac{p_k^*}{p_k}\right) = (1-\alpha_1)\left(\frac{p_k^*}{p_k}\right)\\
P(T^* = 1|T = 0, Z = k) &= P(T = 0|T^* = 1)\left(\frac{p_k^*}{1-p_k}\right) = \alpha_1\left(\frac{p_k^*}{1-p_k}\right)\\
P(T^* = 0|T = 1, Z = k) &= P(T = 1|T^* = 0)\left(\frac{1-p_k^*}{p_k}\right) = \alpha_0\left(\frac{1-p_k^*}{p_k}\right)\\
P(T^* = 0|T = 0, Z = k) &= P(T = 0|T^* = 0)\left(\frac{1-p_k^*}{1-p_k}\right) = (1-\alpha_0)\left(\frac{1-p_k^*}{1-p_k}\right)
\end{aligned}
$$

Proof of Lemma 2.1. The result follows from a simple calculation using the law of total probability. □

Proof of Lemma 2.2. Immediate since $\mathrm{Cov}(T, y|\mathbf{x}) = [1-\alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\,\mathrm{Cov}(T^*, z|\mathbf{x})$ by Lemma 2.1. □

Proof of Theorem 2.1. Throughout this argument we suppress dependence on $\mathbf{x}$ for simplicity. Define $p_k = \mathbb{P}(T = 1|z = k)$ and $p_k^* = \mathbb{P}(T^* = 1|z_k)$.

We first show that so long as $\alpha_0 \leq p_k \leq 1 - \alpha_1$ then we can construct a valid joint probability distribution for (T^*, T, z) that satisfies our assumptions. First decompose the joint probability mass function as

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

By Assumption 2.2 (ii), $p(T|T^*, z) = p(T|T^*)$ and thus α_0 and α_1 fully determine $p(T|T^*, z)$. Under the proposed bounds, α_0 and α_1 are clearly valid probabilities. Since $p(z)$ is observed, it thus suffices to ensure that $p(T^*|z)$ is a valid probability mass function. By the law of total probability and Assumption 2.2 (ii),

$$p_k^* = \frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}.$$

and $0 \leq p_k^* \leq 1$ if and only if $\alpha_0 \leq p_k \leq 1 - \alpha_1$. Since $(p_k - p_\ell) = (p_k^* - p_\ell^*)(1 - \alpha_0 - \alpha_1)$, provided that $p_k - p_\ell \neq 0$ we have $p_k^* \neq p_\ell^*$.

We now show how to construct a valid conditional distribution for y given (T^*, T, z) that satisfies our assumptions if $\beta(p_k - \alpha_0) = (1 - \alpha_0 - \alpha_1)[\mathbb{E}(y|z_k) - c]$ for all k. Define

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

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

$$G_{tk}^0(\tau) = F_{tk}^0(\tau + c), \quad G_{tk}^1(\tau) = F_{tk}^1(\tau + c + \beta) \tag{15}$$

and thus we see that

$$\begin{aligned} G_k(\tau) &= r_{1k}p_k F_{1k}^1(\tau + c + \beta) + r_{0k}(1 - p_k)F_{0k}^1(\tau + c + \beta) \\ &\quad + (1 - r_{1k})p_k F_{1k}^0(\tau + c) + (1 - r_{0k})(1 - p_k)F_{0k}^0(\tau + c) \end{aligned} \tag{16}$$

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

$$F_{tk}(\tau) = r_{tk}F_{tk}^1(\tau) + (1 - r_{tk})F_{tk}^0(\tau) \tag{17}$$

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

$$r_{1k} = \frac{(1 - \alpha_1)p_k^*}{p_k}, \quad r_{0k} = \frac{\alpha_1 p_k^*}{1 - p_k}. \tag{18}$$

There are four cases, corresponding to different possibilities for the r_{tk}.

Case I: $r_{1k} = 0, r_{0k} \neq 0$ By Equation 18, this requires $\alpha_1 = 1$ which is ruled out by Assumption 2.2 (ii).

Case II: $r_{0k} = r_{1k} = 0$ By Equation 18, this requires $p_k^* = 0$ which in turn requires $p_k = \alpha_0$. Moreover, by Equation 17 we have $F_{tk}^0 = F_{tk}$, while F_{tk}^1 is undefined. Substituting into Equation 16,

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

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 $\mathbb{E}(y|z = k) = c$. But since $p_k = \alpha_0$ in this case, $c = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$.

Case III: $r_{1k} \neq 0, r_{0k} = 0$ By Equation 18 this requires $\alpha_1 = 0$ and $p_k^* \neq 0$. By Equation 17 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 16, we obtain

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

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, $F_{0k}^0 = F_{0k}$ in the present case while F_{0k}^1 is undefined. We are free to choose any distributions for F_{1k}^0 and F_{1k}^1 that satisfy Equation 17, for example $F_{1k}^0 = F_{1k}^1 = F_{1k}$.

Case IV: $r_{1k} \neq 0, r_{0k} \neq 0$ In this case, we can solve Equation 17 to obtain

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

Substituting this into Equation 16, we have

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

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 $F_{tk}^0(\tau+c)$ is the conditional CDF of ε given $(T=t, z=k)$ and $F_{tk}^0(\tau+c+\beta)$ 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 simplifies to

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

Thus, in this case we are free to choose *any* distributions for F_{tk}^0 and F_{tk}^1 that satisfy Equation 17. For example we could take $F_{tk}^0 = F_{tk}^1 = F_{tk}$. $\square$

Proof of Corollary 2.1. Follows by plugging in the largest and smallest possible values for $\alpha_0+\alpha_1$ and taking the difference of the expressions for $\mathbb{E}[y|z=k]$ $\square$

Proof of Theorem 2.2. Under Assumption 2.1 (i) and Assumption 2.2 (iii), we obtain $\mathbb{E}(y|T^*, T, z) = \mathbb{E}(y|T^*, z)$. Hence, by iterated expectations

$$
\begin{aligned}
\mathbb{E}(y|T=0,z=k) = (1-r_{0k})\mathbb{E}(y|T^*=0,z=k) + r_{0k}\mathbb{E}(y|T^*=1,z=k) \\
\mathbb{E}(y|T=1,z=k) = (1-r_{1k})\mathbb{E}(y|T^*=0,z=k) + r_{1k}\mathbb{E}(y|T^*=1,z=k)
\end{aligned}
$$

where r_{tk} is defined as in the proof of Theorem 2.1. This is system of two linear equations in two unknowns: $\mathbb{E}(y|T^* = 0, z = k)$ and $\mathbb{E}(y|T^* = 1, z = k)$. After some algebra, we find that the determinant is

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

and thus a unique solution exists provided that $\alpha_0 \neq p_k$ and $\alpha_1 \neq 1 - p_k$. By our assumption that $\mathbb{E}[y|T = 0, z = k] \neq \mathbb{E}[y|T = 1, z = k]$, the system has no solution when the determinant condition fails. Thus, Assumption 2.2 (iii) rules out $\alpha_0 = p_k$ and $\alpha_1 = 1 - p_k$. Solving,

$$\mu_k^0 \equiv \mathbb{E}(y|T^* = 0, z = k) = \left(\frac{1}{1 - p_k - \alpha_1}\right)[(1 - p_k)\mathbb{E}(y|T = 0, z_k) - \alpha_1 \mathbb{E}(y|z = k)]$$
$$\mu_k^1 \equiv \mathbb{E}(y|T^* = 1, z = k) = \left(\frac{1}{p_k - \alpha_0}\right)[p_k \mathbb{E}(y|T = 1, z_k) - \alpha_0 \mathbb{E}(y|z = k)]$$

Given (α_0, α_1), we see that r_{tk}, μ_k^0, and μ_k^1 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_{tk}^0, F_{tk}^1 such that

$$\int_{\mathbb{R}} \tau F_{tk}^0(d\tau) = \mu_k^0, \quad \int_{\mathbb{R}} \tau F_{tk}^1(d\tau) = \mu_k^1, \quad F_{tk}(\tau) = r_{tk} F_{tk}^1(\tau) + (1 - r_{tk}) F_{tk}^0(\tau)$$

where F_{tk} and $F_{tk}^{t^*}$ are as defined in the proof of Theorem 2.2. For a given pair (t, k), there are two cases: $0 < r_{tk} < 1$ and $r_{tk} \in \{0, 1\}$.

Case I: $r_{tk} \in \{0, 1\}$ Suppose that $r_{tk} = 1$. Then, $\mu_k^1 = \mathbb{E}[y|T = t, z = k]$ so we can simply set $F_{tk}^1 = F_{tk}$. In this case F_{tk}^0 is undefined. If instead $r_{tk} = 0$, then $\mu_k^0 = \mathbb{E}[y|T = t, z = k]$ so we can simply set $F_{tk}^0 = F_{tk}$. In this case F_{tk}^1 is undefined.

Case II: $0 < r_{tk} < 1$ Define

$$m_{tk}(\xi) = \mathbb{E}[y|y \in I_{tk}(\xi), T = t, z = k]$$
$$I_{tk}(\xi) = \left[F_{tk}^{-1}(1 - \xi - r_{tk}), F_{tk}^{-1}(1 - \xi)\right]$$

for $t, k = 0, 1$ where $0 \leq \xi \leq 1 - r_{tk}$ and F_{tk}^{-1} is the quantile function of y given $(T = t, z = k)$. We see that m_{tk} is a decreasing function of ξ that attains its maximum at $\xi = 0$ and minimum at $\xi = 1 - r_{tk}$. Define these extrema as $\underline{m}_{tk} = m_{tk}(1 - r_{tk})$ and $\overline{m}_{tk} = m_{tk}(0)$.

Suppose first that μ_k^1 does *not* lie in the interval $[\underline{m}_{tk}, \overline{m}_{tk}]$. We show that it is impossible to construct valid CDFs F_{tk}^0 and F_{tk}^1 that satisfy $F_{tk}(\tau) = r_{tk} F_{tk}^1(\tau) + (1 - r_{tk}) F_{tk}^0(\tau)$ where F_{tk} and $F_{tk}^{t^*}$ are as defined in the proof of Theorem 2.2. 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

$$\frac{F_{tk}(\tau) - (1 - r_{tk})}{r_{tk}} \leq F_{tk}^1(\tau) \leq \frac{F_{tk}(\tau)}{r_{tk}} \tag{19}$$

Now define

$$\underline{F}_{tk}^1(\tau) = \min\{1,\ F_{tk}(\tau)/r_{tk}\}$$
$$\overline{F}_{tk}^1(\tau) = \max\{0,\ F_{tk}(\tau)/r_{tk} - (1 - r_{tk})/r_{tk}\}$$

Combining Equation 19 with the requirement that $0 \leq F_{tk}^1(\tau) \leq 1$, we see that

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

Hence $\overline{F}_{tk}^1$ first-order stochastically dominates F_{tk}^1 which in turn first-order stochastically dominates $\underline{F}_{tk}^1$. It follows that

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

But notice that

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

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

Now suppose that $\mu_k^1 \in [\underline{m}_{tk}, \overline{m}_{tk}]$. Since y is assumed to follow a continuous distribution conditional on (T, z), m_{tk} is continuous on its domain and takes on all values in $[\underline{m}_{tk}, \overline{m}_{tk}]$ by the intermediate value theorem. Thus, there exists a ξ^* such that $m_{tk}(\xi^*) = \mu_k^1$. Now let $f_{tk}(\tau) = dF_{tk}(\tau)/d\tau$ which is non-negative by the assumption that y is continuously distributed. Define the densities

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

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

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

where I_{tk}^C is the complement of I_{tk}. And, 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,

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

Some discussion putting all of the pieces together and explaining what the identified set looks like. In particular, what are the "edge" cases? Can't rule out $\alpha_0 = \alpha_1 = 0$.

□

Proof of Lemma 2.3. Throughout this argument we suppress dependence on $\mathbf{x}$ for simplicity. By Assumption 2.1 (i) and the basic properties of covariance,

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

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,

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

but since $\theta_2 = \theta_1^2 [(1 - \alpha_1) + \alpha_0]$, we see that $[2\theta_1^2(1 - \alpha_1) - \theta_2] = \theta_1\beta$. 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 $m^*_{tk} = \mathbb{E}[\varepsilon|T^* = t, z = k]$ 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. Throughout this argument we suppress dependence on $\mathbf{x}$ for simplicity. 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

$$\text{Cov}(T\varepsilon^2, z) = (1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon^2, z). \tag{20}$$

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

$$\text{Cov}(TT^*\varepsilon, z) = q(1 - q)(1 - \alpha_1)(p^*_1 m^*_{1k} - p^*_0 m^*_{10}) = (1 - \alpha_1)\text{Cov}(T^*\varepsilon, z) \tag{21}$$

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

$$\begin{aligned} \text{Cov}(TT^*, z) &= (1 - \alpha_1)\text{Cov}(T^*, z) \\ \text{Cov}(T^*, z) &= \pi/(1 - \alpha_0 - \alpha_1) \\ \text{Cov}(T^*\varepsilon, z) &= \text{Cov}(T\varepsilon, z)/(1 - \alpha_0 - \alpha_1) \end{aligned}$$

and combining these equalities with Equations 20 and 21, 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

$$\begin{aligned}
K_1 &\equiv 3\theta_1(1-\alpha_0-\alpha_1) \\
K_2 &\equiv 6\theta_1\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] - 3\theta_2(1-\alpha_0-\alpha_1) \\
K_3 &\equiv 3\theta_1\left[(1-\alpha_1)(c+\beta)^2 - c^2\alpha_0\right] - 3\theta_2\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] + \theta_3(1-\alpha_0-\alpha_1)
\end{aligned}$$

Clearly $K_1 = 3\beta$. Substituting the definitions of θ_1, θ_2, and θ_3 from Equations 11–13, tedious but straightforward algebra likewise 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. For ease of notation we suppress dependence on $\mathbf{x}$ throughout. Collecting the results of Lemmas 2.2–2.4, we have

$$\begin{aligned}
\eta_1 &= \pi\theta_1 \\
\eta_2 &= 2\tau_1\theta_1 - \pi\theta_2 \\
\eta_3 &= 3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3
\end{aligned}$$

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 12 and 13 to obtain

$$A = \theta_2/\theta_1^2 = 1 + (\alpha_0 - \alpha_1) \tag{22}$$

$$B = \theta_3/\theta_1^3 = (1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1) \tag{23}$$

Equation 22 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 23 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,

$$r = \frac{A}{2} \pm \sqrt{3A^2 - 2B} = \frac{1}{\theta_1^2}\left(\frac{\theta_2}{2} \pm \sqrt{3\theta_2^2 - 2\theta_1\theta_3}\right). \tag{24}$$

By Equations 12 and 13,

$$\begin{aligned}
3\theta_2^2 - 2\theta_1\theta_3 &= 3\left[\theta_1^2(1+\alpha_0-\alpha_1)\right]^2 - 2\theta_1\left\{\theta_1^3\left[(1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1)\right]\right\} \\
&= \theta_1^4\left\{3(1+\alpha_0-\alpha_1)^2 - 2\left[(1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1)\right]\right\}.
\end{aligned}$$

Expanding the first term we find that

$$\begin{aligned}3(1+\alpha_0-\alpha_1)^2 &= 3\left[1+2(\alpha_0-\alpha_1)+(\alpha_0-\alpha_1)^2\right]\\ &= 3+6\alpha_0-6\alpha_1+3\alpha_0^2+3\alpha_1^2-6\alpha_0\alpha_1\end{aligned}$$

and expanding the second

$$\begin{aligned}2\left[(1-\alpha_0-\alpha_1)^2+6\alpha_0(1-\alpha_1)\right] &= 2\left[1-2(\alpha_0+\alpha_1)+(\alpha_0+\alpha_1)^2+6\alpha_0-6\alpha_0\alpha_1\right]\\ &= 2+8\alpha_0-4\alpha_1+2\alpha_0^2+2\alpha_1^2-8\alpha_0\alpha_1.\end{aligned}$$

Therefore

$$\begin{aligned}3\theta_2^2-2\theta_1\theta_3 &= \theta_1^4\left\{1-2\alpha_0-2\alpha_1+\alpha_0^2-\alpha_1^2+2\alpha_0\alpha_1\right\}\\ &= \theta_1^4\left[(1-\alpha_0-\alpha_1)^2\right]\end{aligned}$$

which is strictly greater than zero since $\theta_1 \neq 0$ and $\alpha_0+\alpha_1 \neq 0$. 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 11, 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. $\square$

B Mahajan's Approach

Here we show that Mahajan's proof of identification for an endogenous treatment is incorrect. The problem is subtle so we give his argument in full detail. We continue to supress dependence on the exogenous covariates $\mathbf{x}$.

The first step of Mahajan's argument is to show that if one could recover the conditional mean function of y given T^*, then a valid and relevant binary instrument would suffice to identify the treatment effect.

Assumption B.1 (Mahajan A2)**.** *Suppose that $y = c + \beta T^* + \varepsilon$ where*

(i) $\mathbb{E}[\varepsilon|z] = 0$

(ii) $\mathbb{P}(T^*=1|z_k) \neq \mathbb{P}(T^*=1|z_\ell)$ *for all* $k \neq \ell$

(iii) $\mathbb{P}(T=1|T^*=0,z)=\alpha_0$, $\mathbb{P}(T=0|T^*=1,z)=\alpha_1$

(iv) $\alpha_0+\alpha_1 \neq 1$

(v) $\beta \neq 0$

Lemma B.1 (Mahajan A2)**.** *Under Assumption ???, knowledge of the mis-classification error rates α_0, α_1 suffices to identify β.*

In his Theorem 1, Mahajan (2006) proves that α_0, α_1 can in fact be identified under the following assumptions.[8]

[8] Technically, one additional assumption is required, namely that the conditional mean of y given T^* and any covariates would be identified if T^* were observed.

Assumption B.2 (Mahajan A1)**.** *Define* $\nu = y - \mathbb{E}[y|T^*]$ *so that by construction we have* $\mathbb{E}[\nu|T^*] = 0$*. Assume that*

(i) $\mathbb{E}[\nu|T^*, T, z] = 0$.[9]

(ii) $\mathbb{P}(T^* = 1|z_k) \neq \mathbb{P}(T^* = 1|z_\ell)$ *for all* $k \neq \ell$

(iii) $\mathbb{P}(T = 1|T^* = 0, z) = \alpha_0$*,* $\mathbb{P}(T = 0|T^* = 1, z) = \alpha_1$

(iv) $\alpha_0 + \alpha_1 < 1$

(v) $\mathbb{E}[y|T^* = 0] \neq \mathbb{E}[y|T^* = 1]$

Lemma B.2 (Mahajan Theorem 1)**.** *Under Assumptions ???, the error rates* α_0, α_1 *are identified as is the conditional mean function* $\mathbb{E}[y|T^*]$*.*

Notice that the identification of the error rates in Lemma ??? does not depend on the interpretation of the conditional mean function $\mathbb{E}[y|T^*]$. If T^* is an exogenous treatment, the conditional mean coincides with the treatment effect; if it is endogenous, this is not the case. Either way, the meaning of α_0, α_1 is unchanged: these parameters simply characterize the mis-classification process. Based on this observation, Mahajan (2006) claims that he can rely on Lemma ??? to identify α_0, α_1 and thus the causal effect β when the treatment is endogenous via Lemma ???. To do this, he must build a bridge between Assumption ??? and Assumption ??? that allows T^* to be endogenous. Mahajan (2006) does this by imposing one additional assumption: Equation 11 in his paper.

Assumption B.3 (Mahajan Equation 11)**.** *Let* $y = c + \beta T^* + \varepsilon$ *where* $\mathbb{E}[\varepsilon|T^*]$ *may not be zero and suppose that*

$$\mathbb{E}[\varepsilon|T^*, T, z] = \mathbb{E}[\varepsilon|T^*].$$

Lemma B.3. *Suppose that* $y = c + \beta T^* + \varepsilon$ *where* $E[\varepsilon|z] = 0$ *and define the unobserved projection error* $\nu = y - \mathbb{E}[y|T^*]$*. Then Assumption ??? implies that* $E[\nu|T^*, T, z] = 0$*, which is Assumption ???.*

To summarize, Mahajan's claim is equivalent to the proposition that under Assumptions ??? β is identified even if T^* is endogenous. although Lemmas ??? are all correct, Mahajan's claim is not.[10] While Assumption ??? does guarantee that Assumption ??? holds, when combined with Assumption ??? it also implies that ??? fails if T^* is endogenous. The failure of Assumption ??? in turn leads to a division by zero in the solution to the linear system following Mahajan's displayed Equation 26: the system no longer has a unique solution so identification fails.[11]

Proposition B.1 (Lack of a First Stage)**.** *Suppose that Assumptions ??? hold and* $\mathbb{E}[\varepsilon|T^*] \neq 0$*. Then* $\mathbb{P}(T^* = 1|z_1) = \mathbb{P}(T^* = 1|z_2)$*, violating Assumption ???.*

[9]This is Mahajan's Equation (I).

[10]Our Lemma ??? does not in fact appear in Mahajan (2006), but it is an implicit step in his proof in Appendix A2.

[11]Notice that the root of the problem is the attempt to use *one* instrument to solve both the measurement error and endogeneity problems. In a setting where one had a second mis-measured surrogate for T^* *in addition* to an instrument that is conditionally mean independent of ε one could use the second surrogate as an instrument for the first to estimate α_0 and α_1 via Lemma ??? and then use the additional instrumental variable to estimate $\beta/(1 - \alpha_0 - \alpha_1)$ via the familiar Wald IV estimator. This is effectively the approach used by Battistin et al. (2014) to evaluate the returns to schooling in a setting with multiple misreported measures of educational qualifications.

Proof. By the Law of Iterated Expectations,

$$\mathbb{E}[\varepsilon|T^*, z] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*, T, z\right)\right] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*\right)\right] = \mathbb{E}\left[\varepsilon|T^*\right] \tag{25}$$

where the second equality follows from Assumption 2.4 and the final equality comes from the fact that $\mathbb{E}[\varepsilon|T^*]$ is (T^*, z)–measurable. Using our notation from above let $u = c + \varepsilon$ and define $m^*_{tk} = \mathbb{E}[u|T^* = t, z = z_k]$. Since c is a constant, by Equation ??? we see that $m^*_{01} = m^*_{02}$ and $m^*_{11} = m^*_{12}$. Now, by Assumption ??? we have $\mathbb{E}[\varepsilon|z] = 0$ so that $\mathbb{E}[u|z_1] = \mathbb{E}[u|z_2] = c$. Again using iterated expectations,

$$\mathbb{E}\left[u|z_1\right] = \mathbb{E}_{T^*|z_1}\left[\mathbb{E}\left(u|T^*, z_1\right)\right] = (1 - p^*_1)m^*_{01} + p^*_1 m^*_{11} = c$$
$$\mathbb{E}\left[u|z_2\right] = \mathbb{E}_{T^*|z_2}\left[\mathbb{E}\left(u|T^*, z_2\right)\right] = (1 - p^*_2)m^*_{02} + p^*_2 m^*_{12} = c$$

The preceding two equations, combined with $m^*_{01} = m^*_{02}$ and $m^*_{11} = m^*_{12}$ imply that $p^*_1 = p^*_2$ unless $m^*_{01} = m^*_{11} = m^*_{02} = m^*_{12} = c$. But this four-way equality is ruled out by the assumption that $\mathbb{E}[\varepsilon|T^*] \neq 0$. $\square$

To understand the economic intuition behind Proposition ???, consider a simple example in which we randomize the offer of a job training program to a sample of workers to study the impact on future earnings. In this context z indicates whether a particular individual is *offered* job training by the experimenter while T^* indicates whether she actually *obtains* job training from any source, inside or outside of the experiment. We observe not T^* but a self-report T that is measured with error. In this example u contains all of the unobservable factors that determine an individual's wage.

Assumption ??? allows for endogenous treatment receipt: $\mathbb{E}[u|T^* = 1]$ may be different from $\mathbb{E}[u|T^* = 0]$. We might expect, for example, that individuals who obtain job training are more motivated than those who do not, and hence earn higher wages on average. However, Assumption ??? imposes that $\mathbb{E}\left[u|T^* = t, z_1\right] = \mathbb{E}\left[u|T^* = t, z_2\right]$ for $t = 0, 1$. This has two implications. First, it means that, among those who do not obtain job training, the average value of u is the same for those who were offered training and those who were not. Second, it means that, among those who *did* obtain job training, the average value of u is the same for those who were offered training and those who were not. In other words, Assumption ??? requires that there is *no selection on unobservables.* This is exactly the opposite of what we would expect in the job training setting. For example, individuals who are offered job training but refuse it, are likely to be very different from those who are not offered training and fail to obtain it from an outside source. And herein lies the problem: Assumption ??? simultaneously allows endogeneity and rules out selection. Given that the offer of job training is randomly assigned, and hence a valid instrument, the only way to avoid a contradiction is if there is no first stage: the fraction of individuals who take up job training cannot depend on the offer of training.

References

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

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.

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

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.

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

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

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., 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 (Forthcoming).

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.

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., 1997. Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. Econometrica, 1201–1213.

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

Lewbel, A., 2012. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. Journal of Business & Economic Statistics 30 (1), 67–80.

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.

Schennach, S. M., 2004. Estimation of nonlinear models with measurement error. Econometrica 72 (1), 33–75.

Schennach, S. M., 2007. Instrumental variable estimation of nonlinear errors-in-variables models. Econometrica 75 (1), 201–239.

Schennach, S. M., 2013. Measurement error in nonlinear models – a review. In: Acemoglu, D., Arellano, M., Dekel, E. (Eds.), Advances in Economics and Econometrics. Vol. 3. Cambridge University Press, pp. 296–337.

Shiu, J.-L., 2015. Identification and estimation of endogenous selection models in the presence of misclassification errors. Economic Modelling (Forthcoming).

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.

Song, S., Schennach, S. M., White, H., 2015. Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors. Quantitative Economics (Forthcoming).

Ura, T., November 2015. Heterogeneous treatment effects with mismeasured endogenous treatment. Tech. rep., Duke University Department of Economics.

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