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: September 2, 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

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 $\mathbf{x}$

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

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

4

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 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

9

$1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x}) > 0$, 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 $B = 1 - 2\min\{\min_k \mathbb{P}(T = 1|\mathbf{x}, z = k), 1 - \max_k \mathbb{P}(T = 1|\mathbf{x}, z = k)\}$

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*

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

for all pairs (t, k) where

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

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

and we define

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

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

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 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_{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 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, perhaps along the lines of Lewbel (2007). 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."

12

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:

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

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

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

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.

---

[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 Appendix C

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 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.

3 Identification-Robust Inference

We now turn our attention to inference based on the identification results from above. As we explain below, inference under binary mis-classification is complicated by problems of weak identification and parameters on the boundary. For simplicity we fix the exogenous covariates at some specified level and suppress dependence on $\mathbf{x}$ in the notation. This is appropriate if the covariates have a discrete support. We discuss how to incorporate covariates more generally in Section 3.6.

3.1 The Non-standard Inference Problem

Lemmas 2.2–2.4 yield the following system of linear moment equalities in the reduced form parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ from Equations 11–13:

$$\text{Cov}(y, z) - \text{Cov}(T, z)\theta_1 = 0$$

$$\text{Cov}(y^2, z) - 2\text{Cov}(yT, z)\theta_1 + \text{Cov}(T, z)\theta_2 = 0$$

$$\text{Cov}(y^3, z) - 3\text{Cov}(y^2T, z)\theta_1 + 3\text{Cov}(yT, z)\theta_2 - \text{Cov}(T, z)\theta_3 = 0$$

Non-linearity arises solely through the relationship between the reduced from parameters $\boldsymbol{\theta}$ and the structural parameters $(\alpha_0, \alpha_1, \beta)$. To convert the preceding moment equations into unconditional moment equalities, we define the additional reduced form parameters $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3)$ as follows:

$$\kappa_1 = c - \alpha_0\theta_1$$
$$\kappa_2 = c^2 + \sigma_{\varepsilon\varepsilon} + \alpha_0(\theta_2 - 2c\theta_1)$$
$$\kappa_3 = c^3 + 3(c - \theta_1\alpha_0)\sigma_{\varepsilon\varepsilon} + \mathbb{E}[\varepsilon^3] - \alpha_0\theta_3 - 3c\alpha_0\left[\theta_1(c + \beta) - 2\theta_1^2(1 - \alpha_1)\right]$$

15

Building on this notation, let

$$\boldsymbol{\psi}_1' = (-\theta_1, 1, 0, 0, 0, 0), \quad \boldsymbol{\psi}_2' = (\theta_2, 0, -2\theta_1, 1, 0, 0), \quad \boldsymbol{\psi}_3' = (-\theta_3, 0, 3\theta_2, 0, -3\theta_1, 1) \quad (15)$$

and collect these in the matrix $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi_1} & \boldsymbol{\psi_2} & \boldsymbol{\psi_3} \end{bmatrix}$. Defining the observed data vector $\mathbf{w}_i' = (T_i, y_i, y_i T_i, y_i^2, y_i^2 T_i, y_i^3)$ for observation i, we can re-write the moment equations as:

$$\mathbb{E}\left[\left(\boldsymbol{\Psi}'(\boldsymbol{\theta})\mathbf{w}_i - \boldsymbol{\kappa} \right) \otimes \begin{pmatrix} 1 \\ z_i \end{pmatrix} \right] = \mathbf{0}. \quad (16)$$

Equation 16 is a just-identified, linear system of moment equalities in the reduced form parameters $(\boldsymbol{\theta}, \boldsymbol{\kappa})$ and yields explicit GMM estimators $(\widehat{\boldsymbol{\kappa}}, \widehat{\boldsymbol{\theta}})$. From Theorem 2.3, knowledge of $\boldsymbol{\theta}$ suffices to identify β. From the definitions of $\boldsymbol{\kappa}$ above and $\boldsymbol{\theta}$ in Equations 11–13, however, the moment equalities from Equation 16 do not depend on (α_0, α_1) if β equals zero. By continuity, they are *nearly* uninformative about the mis-classification probabilities if β is small. But unless $\beta = 0$, knowledge of (α_0, α_1) is necessary to recover β, via Lemma 2.2. Thus, we face a weak identification problem.[8] Indeed, the GMM estimator of $\widehat{\beta}$ based on Equation 16 may even fail to exist. Using arguments from the proof of Theorem 2.3, this estimator is given by is

$$\widehat{\beta} = \text{sign}(\widehat{\theta}_1)\sqrt{3\left(\widehat{\theta}_2/\widehat{\theta}_1\right)^2 - 2\left(\widehat{\theta}_3/\widehat{\theta}_1\right)}$$

Under our assumptions, $3(\theta_2/\theta_1)^2 > 2(\theta_3/\theta_1)$ provided that $\beta \neq 0$, but this may not be true of the sample analogue. Indeed, because $\widehat{\theta}_1$ appears in the denominator, the terms within the square root will be highly variable if β is small. Even if the GMM estimator exists, it may violate the partial identification bounds for (α_0, α_1) from Theorem 2.2, or imply that (α_0, α_1) are not valid probabilities. Importantly, the partial identification bounds remain informative even if β is small or zero: so long as Assumption 2.1 (ii) holds, the first-stage probabilities bound α_0 and α_1 from above.

Exactly the same inferential difficulties arise in the case where T^* and z are assumed to be jointly exogenous, as in Black et al. (2000); Frazis and Loewenstein (2003); Kane et al. (1999); Lewbel (2007); Mahajan (2006).[9] This issue, however, has received little attention in the literature. Kane et al. (1999) ensure that (α_0, α_1) are valid probabilities by employing a logit specification. Frazis and Loewenstein employ a pseudo-Bayesian approach to ensure

---

[8]This is essentially equivalent to the problem of estimating mixture probabilities when the means of the component distributions are very similar to each other.

[9]We provide details for Frazis and Loewenstein (2003) and Mahajan (2006) in Appendix D.

that α_0 and α_1 are valid probabilities, and to impose partial identification bounds related to those from our Theorem 2.1, i.e. without using the non-differential measurement error restrictions. Because they provide neither simulation evidence nor a theoretical justification for their procedure, however, it is unclear whether this method will yield valid Frequentist coverage. We are unaware of any papers in the related literature that discuss the weak identification problem arising when β is small.

3.2 Overview of the Inference Procedure

In the following sections we develop a procedure for uniformly valid inference in models with a mis-classified binary regressor. Our purpose is to construct a confidence interval for β that is robust to possible weak identification, respects the restricted parameter space for (α_0, α_1), and incorporates both the information in the equality moment conditions from Equation 16 along with the partial identification bounds from Theorem 2.2.[10] As argued in the preceding section, our partial identification bounds remain informative even when the equality moment conditions contain essentially no information about (α_0, α_1).

To carry out identification-robust inference combining equality and inequality moment conditions, we adopt the generalized moment selection (GMS) approach of Andrews and Soares (2010). This procedure provides a uniformly valid test of a *joint* null hypothesis for the full parameter vector. In our model, this includes the parameter of interest β along with various nuisance parameters: the mis-classification probabilities α_0 and α_1, the reduced form parameters $\boldsymbol{\kappa}$, defined in Section 3.1, and a vector $\mathbf{q}$ of parameters that enter the moment inequalities.[11] Under a given joint null hypothesis for $(\beta, \alpha_0, \alpha_1)$, however, $\boldsymbol{\kappa}$ and $\mathbf{q}$ are strongly identified and lie on the interior their respective parameter spaces. Accordingly, in Section 3.4 we explain how to concentrate these parameters out of the GMS procedure, by deriving an appropriate correction to the asymptotic variance matrix for the test.[12]

This leaves us with a uniformly valid test of any joint null hypothesis for $(\beta, \alpha_0, \alpha_1)$. To construct a marginal confidence interval for β we proceed as follows. Suppose that z is a strong instrument. Then the usual IV estimator provides a valid confidence interval for the reduced from parameter θ_1. By Lemma 2.2, knowledge of $(1 - \alpha_0 - \alpha_1)$ suffices to determine β from θ_1. Thus, a valid confidence interval for $(1 - \alpha_0 - \alpha_1)$ can be combined with the IV interval for θ_1 to yield a corresponding interval for β, via a Bonferroni-type correction. To

---

[10]Note that $\beta = 0$ if and only if $\theta_1 = 0$. Thus, if one is merely interested in testing $H_0\colon \beta = 0$, one can ignore the mis-classification error problem and test $H_0\colon \theta_1 = 0$ using the standard IV estimator and standard error, provided that z is a strong instrument.

[11]These are defined below in Section 3.3.

[12]Note that we cannot take the same approach to concentrate out α_0 and α_1 because the mis-classification probabilities may be weakly identified or lie on the boundary of their parameter space.

construct the required interval for $(1 - \alpha_0 - \alpha_1)$, we note from Equations 11–13 that β only enters the moment equality conditions in Equation 16 through θ_1. But, again, inference for θ_1 is standard provided that z is a strong instrument. We can thus pre-estimate θ_1 along with $\boldsymbol{\kappa}$ and $\mathbf{q}$, yielding a uniformly valid GMS test of any joint null hypothesis for (α_0, α_1). By inverting this test, we construct a joint confidence set for (α_0, α_1) which we then project to obtain a confidence interval for $(1 - \alpha_0 - \alpha_1)$. Because the parameter space for (α_0, α_1) is bounded and two-dimensional, the projection step is computationally trivial.[13] If desired, one could also carry out a valid test of the null hypothesis that there is no mis-classification, $\alpha_0 = \alpha_1 = 0$, using the joint test for (α_0, α_1). In the following sections we provide full details of our Bonferroni-based confidence interval procedure for β.

3.3 Moment Inequalities

As noted above, the partial identification bounds from Theorems 2.1 and 2.2 remain informative about (α_0, α_1) even when β is small. To incorporate them in our inference procedure, we first express them as unconditional moment inequalities. The bounds from Theorem 2.1 are given by

$$p_k - \alpha_0 \geq 0, \qquad 1 - p_k - \alpha_1 \geq 0, \quad \text{for all } k$$

where the first-stage probabilities p_k are defined in Equation 9. We write these as

$$\mathbb{E}\left[m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta})\right] \geq \mathbf{0}, \quad m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta}) \equiv \begin{bmatrix} (1 - z_i)(T_i - \alpha_0) \\ (1 - z_i)(1 - T_i - \alpha_1) \\ z_i(T - \alpha_0) \\ z_i(1 - T_i - \alpha_1) \end{bmatrix} \tag{17}$$

The bounds derived in Theorem 2.2 by imposing assumption 2.2 (iii) are

$$\mu_k(\alpha_0) - \underline{\mu}_{tk}\big(\underline{q}_{tk}(\alpha_0, \alpha_1)\big) \geq 0, \qquad \overline{\mu}_{tk}\big(\overline{q}_{tk}(\alpha_0, \alpha_1)\big) - \mu_k(\alpha_0) \geq 0, \quad \text{for all } t, k$$

---

[13] We considered two alternatives to the Bonferroni-based inference procedure described here. The first constructs a marginal confidence interval for β by projecting a joint confidence set for $(\beta, \alpha_1, \alpha_0)$, i.e. *without* preliminary estimation of θ_1. This method is more computationally demanding than our two-dimensional projection and involves a parameter space that is unbounded along the β-dimension. From a practical perspective, the relevant question is whether the reduction in conservatism from projecting a lower dimensional set is outweighed by the additional conservatism induced by the Bonferroni correction. In our experiments, the full three-dimensional projection and Bonferroni procedure produced broadly similar results: neither reliably dominated in terms of confidence interval width. Given its substantially lower computational burden, we prefer the Bonferroni procedure. We also experimented with two recently proposed methods for sub-vector inference: Kaido et al. (2016) and Bugni et al. (2017). In both cases we obtained significant size distortions, suggesting that our model may not satisfy the high-level regularity conditions required by these papers.

where $\mu_k, \underline{\mu}_{tk}, \overline{\mu}_{tk}, \underline{q}_{tk}$ and $\overline{q}_{tk}$ are defined in the statement of the Theorem. Expressing these as unconditional moment inequalities, we have

$$\mathbb{E}[m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q})] \geq \mathbf{0}, \quad m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} m_{2,00}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,10}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,01}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,11}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \end{bmatrix} \tag{18}$$

where $\mathbf{q} \equiv (\underline{q}_{00}, \overline{q}_{00}, \underline{q}_{10}, \overline{q}_{10}, \underline{q}_{01}, \overline{q}_{01}, \underline{q}_{11}, \overline{q}_{11})$ and we define

$$m_{2,0k}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} y_i \mathbf{1}\,(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right) \right\} \\ -y_i \mathbf{1}(z_i = k)\left\{ (T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right) \right\} \end{bmatrix} \tag{19}$$

$$m_{2,1k}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} y_i \mathbf{1}\,(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{1k})T_i\left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right) \right\} \\ -y_i \mathbf{1}(z_i = k)\left\{ (T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{1k})T_i\left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right) \right\} \end{bmatrix}. \tag{20}$$

Finally we define $m^I = (m_1^{I\prime}, m_2^{I\prime})'$. Notice that the second set of inequalities, m_2^I, depends on the unknown parameter $\mathbf{q}$ which is in turn a function of (α_0, α_1). In the next section we discuss how $\mathbf{q}$ can be estimated under a given null hypothesis for (α_0, α_1).

3.4 Accounting for Preliminary Estimation

Let $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\gamma} = (\boldsymbol{\kappa}, \theta_1)$ where θ_1 is defined in Equation 11 and $\boldsymbol{\kappa}$ in Section 3.1. Our moment conditions take the form

$$\mathbb{E}[m^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)] \geq \mathbf{0}, \quad \mathbb{E}[m^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)] = \mathbf{0} \tag{21}$$

where $m^I = (m_1^{I\prime}, m_2^{I\prime})'$, defined in Section 3.3, and

$$m^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) = \begin{bmatrix} \{\boldsymbol{\psi}_2'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \kappa_2\}\,z_i \\ \{\boldsymbol{\psi}_3'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \kappa_3\}\,z_i \end{bmatrix}. \tag{22}$$

Notice that we now write $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$, defined in Equation 15, as explicit functions of $(\theta_1, \alpha_0, \alpha_1)$, using the definitions of (θ_2, θ_3) from Equations 12–13. To construct a GMS test of the null hypothesis $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta_0}$ based on Equation 21, we require preliminary estimators $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ that are consistent and asymptotically normal *under the null*. We now provide full details of the construction and derive the associated adjustment to the

asymptotic variance matrix.

Consider first the equality moment conditions m^E. For these we require preliminary estimators of θ_1, κ_2, and κ_3. Recall that θ_1 is simply the IV estimand: it can be consistently estimated directly from observations of (y, T, z) without knowledge of α_0 or α_1. Note, moreover, from Equation 16 that $\boldsymbol{\kappa}$ is simply a vector of *intercepts*. These can be directly estimated from observations of $\mathbf{w}$ because $\boldsymbol{\Psi}(\theta_1, \alpha_0, \alpha_1)$ is consistently estimable under the null $H_0 \colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$: the hypothesis specifies α_0 and α_1 and IV provides a consistent estimator of θ_1. Accordingly, define

$$h^E(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\gamma}) = \left[\begin{array}{c} \boldsymbol{\Psi}'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \boldsymbol{\kappa} \\ \left\{ \boldsymbol{\psi}'_1(\theta_1)\mathbf{w}_i - \kappa_1 \right\} z_i \end{array} \right]. \tag{23}$$

Under $H_0 \colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, the just-identified GMM-estimator based on $\mathbb{E}[h^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)] = \mathbf{0}$ yields a consistent and asymptotically normal estimator of $\boldsymbol{\gamma}_0$ under the usual regularity conditions.

Now consider the inequality moment conditions m^I. From Section 3.3 we see that m_2^I depends on $\mathbf{q}$, the vector of conditional quantiles $\overline{q}_{tk}$ and $\underline{q}_{tk}$ defined in Theorem 2.2. Under the assumption that y follows a continuous distribution, as maintained in Theorem 2.2, these can be expressed as conditional moment equalities as follows:

$$\mathbb{E}\left[\mathbf{1}(y \le \underline{q}_{tk})|T = t, z = k\right] - r_{tk}(\alpha_0, \alpha_1) = 0 \tag{24}$$

$$\mathbb{E}\left[\mathbf{1}(y \le \overline{q}_{tk})|T = t, z = k\right] - \left(1 - r_{tk}(\alpha_0, \alpha_1)\right) = 0 \tag{25}$$

where r_{tk} is defined in Theorem 2.2 and $t, k = 0, 1$. Under $H_0 \colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, a consistent estimator $\widehat{r}_{tk}$ of r_{tk} can be obtained directly from $\widehat{p}_k$, the sample analogue of p_k based on iid observations of $\mathbf{w}_i$. In turn, the $(\widehat{r}_{tk})^{\text{th}}$ and $(1 - \widehat{r}_{tk})^{\text{th}}$ sample conditional quantiles of y provide consistent estimates of $\underline{q}_{tk}$ and $\overline{q}_{tk}$.[14] Collecting these for all (t, k) gives $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$. Now, define

$$h^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) = \left[\begin{array}{c} h_0^I(\mathbf{w}, \boldsymbol{\vartheta}, \mathbf{q}) \\ h_1^I(\mathbf{w}, \boldsymbol{\vartheta}, \mathbf{q}) \end{array} \right] \tag{26}$$

---

[14]Consistency of the sample quantiles requires $0 < r_{tk} < 1$. If $r_{tk} = 0$ or 1 for some (t, k), however, then the associated moment inequality is trivially satisfied and we no longer require estimates of $\underline{q}_{tk}, \overline{q}_{tk}$.

where

$$h_k^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) = \begin{bmatrix} \mathbf{1}(y_i \leq \underline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{1-\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \\ \mathbf{1}(y_i \leq \underline{q}_{1k})\mathbf{1}(z_i = k)T_i - \left(\frac{1-\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{1k})\mathbf{1}(z_i = k)T_i - \left(\frac{\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{bmatrix}. \quad (27)$$

Equation 27 gives the unconditional version of Equations 24–25. Now, under the null $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ converges in probability to $\mathbf{q}_0$, which satisfies the just-identified collection of moment equalities $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)] = \mathbf{0}$. Although h^I is a discontinuous function of $\mathbf{q}$, it is bounded for any fixed (α_0, α_1). Moreover, since $y|(T = t, z = k)$ is a continuous random variable, $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q})]$ is continuously differentiable with respect to $\mathbf{q}$. Hence, $\widehat{\mathbf{q}}$ is asymptotically normal under mild regularity conditions.[15] To account for the effect of preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ on the asymptotic variance matrix used in the GMS test, we rely on the following Lemma.

Lemma 3.1. *Let $\bar{m}_{1,n}^I(\boldsymbol{\vartheta}) = n^{-1}\sum_{i=1}^n m_{1,n}^I(\mathbf{w}_i, \boldsymbol{\vartheta})$ and define $\bar{m}_{2,n}^I, \bar{m}_n^E, \bar{h}_n^I, \bar{h}_n^E$ analogously. Further let $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0) = \arg\min_{\boldsymbol{\gamma}}\|\bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma})\|$ and $\|h_n^I(\boldsymbol{\vartheta}_0, \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0))\| \leq \inf_{\mathbf{q}}\|h_n^I(\boldsymbol{\vartheta}_0, \mathbf{q})\| + o_p(1)$. Then, under standard regularity conditions*

$$\sqrt{n}\begin{bmatrix} \bar{m}_{1,n}^I(\boldsymbol{\vartheta}_0) \\ \bar{m}_{2,n}^I(\boldsymbol{\vartheta}_0, \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)) \\ \bar{m}_n^E(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)) \end{bmatrix} \to_p \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & B^I(\boldsymbol{\vartheta}_0, \mathbf{q}_0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & B^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \end{bmatrix} \sqrt{n}\begin{bmatrix} \bar{m}_{1,n}^I(\boldsymbol{\vartheta}_0) \\ \bar{m}_{2,n}^I(\boldsymbol{\vartheta}_0, \mathbf{q}_0) \\ \bar{m}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \\ \bar{h}_n^I(\boldsymbol{\vartheta}_0, \mathbf{q}_0) \\ \bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \end{bmatrix}$$

where $B^I(\boldsymbol{\vartheta}, \boldsymbol{q}) = (1 - \alpha_0 - \alpha_1)\left[diag(\mathbf{a})\right]^{-1}\mathbf{q}$ with

$$\mathbf{a}' = (\alpha_1, \alpha_1, 1 - \alpha_1, 1 - \alpha_1, \alpha_1, \alpha_1, 1 - \alpha_1, 1 - \alpha_1)$$

---

[15] Cite Andrews/Newey-McFadden.

and $B^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) = -M^E(H^E)^{-1}$, with

$$
M^E = \begin{bmatrix} 0 & -\mathbb{E}[z_i] & 0 & \left(\frac{\partial \psi_2}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \\ 0 & 0 & -\mathbb{E}[z_i] & \left(\frac{\partial \psi_3}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \end{bmatrix}, \quad H^E = \begin{bmatrix} -1 & 0 & 0 & \left(\frac{\partial \psi_1}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ 0 & -1 & 0 & \left(\frac{\partial \psi_2}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ 0 & 0 & -1 & \left(\frac{\partial \psi_3}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ -\mathbb{E}[z_i] & 0 & 0 & \left(\frac{\partial \psi_1}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \end{bmatrix}.
$$

Lemma 3.1 relates the sample analogues $\bar{m}_{2,n}^I$ and $\bar{m}_n^E$ evaluated at the preliminary estimators $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ to their counterparts evaluated at the true parameter values $\mathbf{q}_0$ and $\boldsymbol{\gamma}_0$. The estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ exactly solves $h_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}) = 0$ while $\widehat{q}(\boldsymbol{\vartheta}_0)$, constructed as described immediately before the statement of the Lemma, *approximately* solves $\bar{h}_n^I(\boldsymbol{\vartheta}_0, \mathbf{q}) = 0$. A few lines of matrix algebra show that the determinant of H^E equals $\mathrm{Cov}(z, T)$. Hence, B^E is well-defined if z is a relevant instrument. The matrix B^I is likewise well-defined provided that $\alpha_1 \neq 0$ and the elements of $\mathbf{q}_0$ are computed for probabilities strictly between zero and one. If either of these conditions fails, however, some of the moment inequalities in m_2^I are trivially satisfied and can be dropped (see Footnote 14). After removing the corresponding elements of $\mathbf{q}_0$ and $\mathbf{a}$, B^I becomes well-defined. The regularity conditions required for Lemma 3.1 are mild. The result relies on a number of mean-value expansions: $\bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)$ and $\bar{m}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)$ are expanded around $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ while $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)]$ and $\mathbb{E}[m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)]$ are expanded around $\mathbf{q} = \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$. These expansions, in turn, rely on the fact that $\mathbf{q}$ and $\boldsymbol{\gamma}$ are interior to their respective parameter spaces and the relevant functions are all continuously differentiable in our example.

We now have all the ingredients required to construct an asymptotic variance matrix for the GMS test that accounts for preliminary estimation of $\boldsymbol{\gamma}$ and $\mathbf{q}$. Let $m' = (m_1^{I'}, m_2^{I'}, m^{E'})$, $h' = (h^{I'}, h^{E'})$, and define the shorthand $\boldsymbol{\tau}_0' = (\boldsymbol{\gamma}_0', \mathbf{q}_0')$ and $\widehat{\boldsymbol{\tau}}_0' = (\widehat{\boldsymbol{\gamma}}'(\boldsymbol{\vartheta}_0), \widehat{\mathbf{q}}'(\boldsymbol{\vartheta}_0))$. Given a collection of iid observations $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$, we have

$$
\sqrt{n} \begin{bmatrix} \bar{m}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \\ \bar{h}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \end{bmatrix} \to_d N\big(\mathbf{0}, \mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\big), \quad \mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) = \mathrm{Var} \begin{bmatrix} m(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \end{bmatrix} \tag{28}
$$

under $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, by an appropriate central limit theorem. What we require for the test, however, is the asymptotic variance matrix of $\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$. Combining Equation 28 with Lemma 3.1, we obtain

$$
\mathrm{Avar}\big(\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\big) = \Xi(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\, \mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\, \Xi'(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \tag{29}
$$

with

$$\Xi(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \begin{bmatrix} \mathbf{I} & B(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{bmatrix}, \quad B(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B^I(\boldsymbol{\vartheta}, \mathbf{q}) & \mathbf{0} \\ \mathbf{0} & B^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{bmatrix} \tag{30}$$

where $B^I(\cdot, \cdot)$ and $B^E(\cdot, \cdot)$ are defined in Lemma 3.1. Finally, we construct a consistent estimator $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$ of the asymptotic variance matrix of $\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ under the null:

$$\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0) \equiv \Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\,\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}})\,\Xi'(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0) \tag{31}$$

where

$$\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \equiv \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} m(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{h}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{bmatrix} \begin{bmatrix} m(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{h}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{bmatrix}'. \tag{32}$$

In the following section we provide a step-by-step description of our inference procedure.

3.5   Details of the Inference Procedure

In this section we provide full details of our Bonferroni-based inference procedure. We begin by defining some notation. Let J denote the total number of inequality moment conditions, K denote the total number of equality moment conditions, and define

$$\bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \begin{bmatrix} \bar{m}_n^I(\boldsymbol{\vartheta}, \mathbf{q}) \\ \bar{m}_n^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{bmatrix} = \begin{bmatrix} \bar{m}_{n,1}^I(\boldsymbol{\vartheta}) \\ \bar{m}_{n,2}^I(\boldsymbol{\vartheta}, \mathbf{q}) \\ \bar{m}_n^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta}) \\ m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m^E(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{bmatrix} \tag{33}$$

with m_1^I as defined in Equation 17, m_2^I in Equations 18–20 and m^E in Equation 22.[16] Now let S be the function

$$S(\mathbf{x}, \mathbf{y}) = \sum_j \min\{0, x_j^2\} + \mathbf{y}'\mathbf{y} \tag{34}$$

where $\mathbf{x}, \mathbf{y}$ are two arbitrary real vectors. This function will be used to calculate the modified method of moments (MMM) test statistic as part of the GMS test below. The argument $\mathbf{x}$ stands in for the moment inequalities, which only contribute to the test statistic when they are violated, i.e. take on a *negative* value. Using this notation, we now detail the first step of our inference procedure: a GMS test for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ with preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ under the null.

---

[16]In our problem $K = 2$ and J is at most 12. Under certain nulls for (α_0, α_1), however, we drop components of m_2^I as they are trivially satisfied. See footnote 14 and Section 3.4 for further discussion.

Algorithm 3.1 (GMS Test for α_0 and α_1).

Inputs: hypothesis $\boldsymbol{\vartheta}_0$; iid dataset $\{\mathbf{w}_i\}_{i=1}^n$; simulations $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R \sim$ iid $N_{J+K}(\mathbf{0}, \mathbf{I})$.

1. Calculate the variance matrix estimator $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$.

   (i) Calculate $\widehat{\boldsymbol{\tau}}_0 = (\widehat{\mathbf{q}}_0', \widehat{\boldsymbol{\gamma}}_0')'$ where $\widehat{\boldsymbol{\gamma}}_0 = \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\mathbf{q}}_0 = \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ from Section 3.4.

   (ii) Calculate $\Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 30.

   (iii) Calculate $\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 32.

   (iv) Set $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0) = \Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\, \widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}})\, \Xi'(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$.

2. Calculate the test statistic $T_n(\boldsymbol{\vartheta}_0)$.

   (i) Calculate $\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 33.

   (ii) Set $\boldsymbol{\nu}_n(\boldsymbol{\vartheta}_0) = \left[\text{diag}\left\{\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)\right\}\right]^{-1/2} [\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)]$.

   (iii) Let $\boldsymbol{\nu}_n^I(\boldsymbol{\vartheta}_0)$ denote the first J elements of $\boldsymbol{\nu}_n$ and $\boldsymbol{\nu}_n^E(\boldsymbol{\vartheta}_0)$ the last K elements.

   (iv) Set $T_n(\boldsymbol{\vartheta}_0) = S\big(\boldsymbol{\nu}_n^I(\boldsymbol{\vartheta}_0), \boldsymbol{\nu}_n^E(\boldsymbol{\vartheta}_0)\big)$ using Equation 34.

3. Construct the moment selection matrix Φ.

   (i) For $j = 1, \ldots, J$ set $\varphi_j^I = \mathbf{1}\left\{\nu_{n,j}^I(\boldsymbol{\vartheta}_0) \le \sqrt{\log n}\right\}$ and let $\widetilde{J} = \sum_{j=1}^J \varphi_j^I$.

   (ii) For $j = 1, \ldots, K$ set $\varphi_j^E = 1$.

   (iii) Set $\boldsymbol{\varphi} = (\varphi_1^I, \ldots, \varphi_J^I, \varphi_1^E, \ldots, \varphi_K^E)'$.

   (iv) Let Φ be the $(\widetilde{J} + K) \times (J + K)$ of zeros and ones that selects those elements x_j of an arbitrary vector $\mathbf{x}$ that correspond to $\varphi_j = 1$.

4. Simulate the sampling distribution of $T_n(\boldsymbol{\vartheta}_0)$ under the null.

   (i) Let $\widehat{\Omega}$ be the correlation matrix that corresponds to $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$.

   (ii) For each $r = 1, \ldots, R$ set $\boldsymbol{\xi}^{(r)} = \left[\Phi\, \widehat{\Omega}\, \Phi'\right]^{1/2} \Phi \boldsymbol{\zeta}^{(r)}$.

   (iii) Let $\boldsymbol{\xi}_I^{(r)}$ denote the first $\widetilde{J}$ and $\boldsymbol{\xi}_E^{(r)}$ the last K elements of $\boldsymbol{\xi}^{(r)}$.

   (iv) For each $r = 1, \ldots, R$ set $T_n^{(r)}(\boldsymbol{\vartheta}_0) = S\big(\boldsymbol{\xi}_I^{(r)}, \boldsymbol{\xi}_E^{(r)}\big)$ using Equation 34.

5. Calculate the p-value of the test: $\widehat{p}(\boldsymbol{\vartheta}_0) = \dfrac{1}{R} \sum_{r=1}^R \mathbf{1}\left\{T_n^{(r)}(\boldsymbol{\vartheta}_0) > T_n(\boldsymbol{\vartheta}_0)\right\}$.

Algorithm 3.1 corresponds to the asymptotic version of the GMS test from Andrews and Soares (2010), based on the MMM test statistic – S_1 in Andrews and Soares (2010) – and the "BIC choice" $\kappa_n = \sqrt{\log n}$ for the sequence of constants κ_n used for moment selection. The procedure is as follows. In Step 1, we compute a consistent estimator of the asymptotic variance matrix of the full set of moment conditions, under the null, accounting for preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ as explained in Section 3.4. In step 2, we calculate the observed value of the MMM test statistic. Note that this test statistic uses only the diagonal elements of $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$. Moreover, the moment inequalities only contribute to T_n if they are violated, i.e. if they take on a negative value. In step 3 we determine which moment inequalities are "far from binding," defined as having a t-ratio greater than $\sqrt{\log n}$. These moment inequalities will be excluded when approximating the large-sample distribution of the test statistic. The matrix Φ encodes the results of the moment selection step. Pre–multiplying a $(J + K)$–vector $\mathbf{x}$ by Φ results in a $(\widetilde{J} \times K)$–vector $\widetilde{\mathbf{x}}$ whose last K elements match the last K elements of $\mathbf{x}$ but whose first $\widetilde{J}$ elements contain the subset of $(x_1, \ldots, x_J)$ whose indices match those of the moment inequalities with t-ratios less than or equal to $\sqrt{\log n}$, i.e. those that are *not* far from binding.[17] Step 4 uses a collection of iid normal draws, $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R$, to approximate the large-sample distribution of T_n under the null. The appropriate multiplications by Φ ensure that this approximation includes all moment equalities, but excludes any moment inequality judged to be far from binding in step 3. Finally, step 5 computes the p-value of the test by comparing the actual test statistic $T_n(\boldsymbol{\vartheta}_0)$ to the collection of simulated test statistics $\{T_n^{(r)}(\boldsymbol{\vartheta}_0)\}_{r=1}^R$ from step 4. We now detail our Bonferroni-based confidence interval for β.

Algorithm 3.2 (Bonferroni-based Confidence Interval for β).
Inputs: significance levels (δ_1, δ_2); iid dataset $\{\mathbf{w}_i\}_{i=1}^n$; simulations $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R \sim$ iid $N_{J+K}(\mathbf{0}, \mathbf{I})$.

1. Construct a $(1 - \delta_1) \times 100\%$ joint confidence set $\mathcal{C}(\delta_1)$ for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)'$.

    (i) Let $\Lambda_N = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-2}{N}, \frac{N-1}{N}\right\}$ where $N > 1$ is a natural number.

    (ii) Set $\Delta_N = \{(\alpha_0, \alpha_1) \in (\Lambda_N \times \Lambda_N): \alpha_0 + \alpha_1 < 1\}$.

    (iii) For each $\boldsymbol{\vartheta} \in \Delta_N$ calculate $\widehat{p}(\boldsymbol{\vartheta})$ using Algorithm 3.1, holding $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R$ *fixed*.

---

[17]Although this does not affect the results of the procedure, notice that Algorithm 3.1 carries out moment selection in a slightly different way from the steps given by Andrews and Soares (2010). In particular, before carrying out any further calculations, we *subset* the correlation matrix $\widehat{\Omega}$ and normal vectors $\boldsymbol{\zeta}^{(r)}$ to remove elements corresponding to moment inequalities deemed far from binding. In contrast, Andrews and Soares (2010) carry along the full set of inequalities throughout, but add $+\infty$ to the appropriate elements when computing $T_n^{(r)}$ to ensure that only the moment inequalities that are not far from binding affect the results. Although it requires more notation to describe, sub-setting is substantially faster, as it avoids carrying out computations for inequalities that cannot affect the result.

(iv) Set $\mathcal{C}(\delta_1) = \{\boldsymbol{\vartheta} \in \Delta_N \colon \widehat{p}(\boldsymbol{\vartheta}) \geq \delta_1\}$.

2. Construct a $(1 - \delta_1) \times 100\%$ confidence interval $[\underline{s}(\delta_1), \overline{s}(\delta_1)]$ for $s \equiv (1 - \alpha_0 - \alpha_1)$.

    (i) Set $\underline{s}(\delta_1) = \min\left\{(1 - \alpha_0 - \alpha_1) \colon (\alpha_0, \alpha_1) \in \mathcal{C}(\delta_1)\right\}$.

    (ii) Set $\overline{s}(\delta_1) = \max\left\{(1 - \alpha_0 - \alpha_1) \colon (\alpha_0, \alpha_1) \in \mathcal{C}(\delta_1)\right\}$.

3. Construct a $(1 - \delta_2) \times 100\%$ confidence interval $\left[\underline{\theta}_1(\delta_2), \overline{\theta}_1(\delta_2)\right]$ for θ_1.

    (i) Use the standard IV interval from a regression of y on T with instrument z.

4. Construct the $(1 - \delta) \times 100\%$ Bonferroni-based confidence interval $\left[\underline{\beta}(\delta), \overline{\beta}(\delta)\right]$ for β.

    (i) Let $\delta = \delta_1 + \delta_2$.

    (ii) Set $\underline{\beta}(\delta) = \min\left\{\underline{s}(\delta_1) \times \underline{\theta}_1(\delta_2),\ \overline{s}(\delta_1) \times \underline{\theta}_1(\delta_2)\right\}$.

    (iii) Set $\overline{\beta}(\delta) = \min\left\{\underline{s}(\delta_1) \times \overline{\theta}_1(\delta_2),\ \overline{s}(\delta_1) \times \overline{\theta}_1(\delta_2)\right\}$.

Step 1 of Algorithm 3.2 constructs a $(1 - \delta_1) \times 100\%$ joint confidence set $\mathcal{C}(\delta_1)$ for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ by inverting the GMS test from Algorithm 3.1 over a discretized parameter space Δ_N. Because the parameter space for (α_0, α_1) is bounded, this is computationally straightforward. Note that the *same* normal draws $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R$ are used to test each null hypothesis contained in Δ_N. Step 2 projects $\mathcal{C}(\delta_1)$ to yield a $(1 - \delta_1) \times 100\%$ confidence interval for $s \equiv (1 - \alpha_0 - \alpha_1)$, simply taking the maximum and minimum values of s in the discrete set $\mathcal{C}(\delta_1)$. Step 3 constructs the usual IV confidence interval for the reduced form parameter θ_1, and step 4 combines the results of steps 2–3 with Bonferroni's inequality to yield a $(1 - \delta_1 - \delta_2) \times 100\%$ confidence interval for β. For some discussion of alternatives to Algorithm 3.2, see Footnote 13. Under mild regularity conditions, the confidence interval from Algorithm 3.2 is uniformly asymptotically valid.

> Should add at least a couple of sentences about how Bonferroni relates to the IV interval: in particular it only excludes zero if the CI for θ_1 excludes zero.

Theorem 3.1. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be an iid collection of observations satisfying the conditions of Theorems 2.2 and 2.3, and let z be a strong instrument. Then, under standard regularity conditions, the confidence interval for β from Algorithm 3.2 has asymptotic coverage probability no less than $1 - (\delta_1 + \delta_2)$ as $R, N, n \to \infty$ uniformly over the parameter space.*

Theorem 3.1 is effectively a corollary of Theorem 1 from Andrews and Soares (2010), which establishes the uniform asymptotic validity of the GMS test, and Lemma 3.1, which

accounts for preliminary estimation of $\boldsymbol{\gamma}$ and $\mathbf{q}$. Given iid observations $\mathbf{w}_i$, the only substantive condition required for Theorem 3.1 is the joint asymptotic normality of $\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)$ and $\sqrt{n}\bar{h}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)$, where $\bar{h}_n$ denotes the sample analogues for full set of auxiliary moment conditions (h^I, h^E) defined in Section 3.4. For further discussion of the regularity conditions required for the GMS procedure, see Appendix A3 of Andrews and Soares (2010). For some discussion of the regularity conditions required for Lemma 3.1, see Section 3.4.

Theorem 3.1 invokes the higher-moment assumptions (Assumptions 2.5–2.6) under which we establish global identification of β in Theorem 2.3, and Algorithm 3.1 likewise incorporates the higher-moment equality conditions that arise from this result. To proceed without these conditions, simply remove m^E from the set of moment conditions used in the algorithm and leave the steps unchanged. In this case β is no longer point identified but the inference procedure provides valid inference for the points in the sharp identified set from Theorem 2.2. Algorithm 3.2 can likewise be used in the case of an exogenous T^*, as in Mahajan (2006) and Frazis and Loewenstein (2003). As mentioned above in Section 3.1, the exogenous regressor case is subject to the same inferential difficulties and the endogenous case on which we focuse in this paper. To accommodate an exogenous regressor,simply replace m^E with the moment equalities described in Appendix D.

3.6 Incorporating Covariates

Just give a quick overview of how one can handle them more generally. There several routes: one is to be more restrictive and do what FL do. The other is to try to be fully nonparametric. There's a paper by Andrews that gives some more details. This is likely to be complicated in general. There may be a "halfway house" using a semiparametric model. Perhaps look at what Lewbel said about this. This Section should really just be a paragraph or so.

4 Simulation Study

In this section we present results from a simulation study using the inference procedure described in Section 3.5 above. Unless otherwise specified, all calculations are based on 2000 simulation replications with $n = 1000$ using Algorithm 3.2 with $R = 5000$ simulation draws. Supplementary simulation results appear in Appendix E.

4.1 Simulation DGP

Our simulation design generates n iid observations of the observables (y_i, T_i, z) as follows:

1. Generate the instrumental variable z.

   (i) For each $1 \le i \le n/2$ set $z_i = 0$.

   (ii) For each $n/2 < i \le n$, set $z_i = 1$.

2. Generate the error terms:

$$\begin{bmatrix} \eta_i \\ \varepsilon_i \end{bmatrix} \sim \text{iid N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$

3. Generate the unobserved regressor: $T_i^* = \mathbf{1}\{d_0 + d_1 z_i + \eta_i > 0\}$.

4. Generate the outcome: $y_i = c + \beta T_i^* + \varepsilon_i$.

5. Generate the observed, mis-classified regressor T.

   (i) For all i with $T_i^* = 0$ draw $T_i \sim$ iid Bernoulli(α_0).

   (ii) For all i with $T_i^* = 1$ draw $T_i \sim$ iid Bernoulli($1 - \alpha_1$).

This DGP generates random variables that satisfy the conditions of Theorems 2.2 and 2.3. Thus β is point identified, and all moment equalities and inequalities from Section 3 hold at the true parameter values of the DGP. Note from step 1 that we hold *condition* on the instrument z, holding it fixed in repeated samples. Our simulation varies the parameters $(\alpha_0, \alpha_1, \beta, n)$ over a grid. Because ε has unit variance, values for β are measured in standard deviations of the error. For simplicity we present results for $c = 0, d_0 = \Phi^{-1}(0.15)$, and $d_1 = \Phi^{-1}(0.85) - \Phi^{-1}(0.15)$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of a standard normal random variable. Using these values for (d_0, d_1) holds the unobserved first stage probabilities fixed: $p_0^* = 0.15$ and $p_1^* = 0.85$. In contrast the *observed* first-stage probabilities p_0 and p_1 vary with (α_0, α_1) according to Lemma 2.1.

4.2 Simulation Results

As explained in Section 3.1 above, the just-identified, unconstrained GMM estimator based on Equation 16 suffers from weak identification and boundary value problems. Moreover, the estimator may not even exist in finite samples. Even when the GMM estimator exists, its asymptotic variance matrix could be numerically singular, so that the standard GMM

confidence interval is undefined. Table 1 reports the percentage of simulation draws for which the standard GMM confidence interval is undefined, while Table 2 reports the coverage probability of a nominal 95% GMM confidence interval, conditional on its existence.

α_0	α_1	β							
		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 1: Percentage of replications for which the standard GMM confidence interval based on Equation 16 fails to exist, either because the point estimate is NaN or the asymptotic covariance matrix is numerically singular. Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

We see from Table 1 that when β is small compared to the error variance, the GMM confidence interval fails to exist with high probability. When $\beta = 0.5$, for example, the interval is undefined approximately 30% of the time. As β increases, however, it becomes less likely that the GMM interval is undefined. All else equal, larger amounts of mis-classification, i.e. higher values for (α_0, α_1), increase the probability that the GMM interval fails to exist. Turning our attention to the simulation draws for which it is well-defined, we see from Tables 2 and 6 that the GMM confidence interval performs extremely poorly when β is small. Substantial size distortions persist until β is 1.5 or larger. All else equal, the size distortions are more severe the larger the amount of mis-classification error. For sufficiently large β, however, standard GMM inference performs well. As β grows, the weak identification problem vanishes. For large enough β the inference problem in effect becomes standard.

We now examine the performance of the Bonferroni-based confidence interval from Algorithm 3.2, beginning with its first step: a joint GMS confidence set for (α_0, α_1). Table 3 presents coverage probabilities for a nominal 97.5% GMS confidence set for (α_0, α_1). Be-

α_0	α_1	β							
		0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	72	62	62	80	92	95	94	95
	0.1	72	62	63	79	92	95	96	95
	0.2	73	61	61	77	90	96	96	96
	0.3	73	59	62	76	88	95	96	95
0.1	0.0	73	63	60	78	91	95	96	96
	0.1	73	58	59	77	90	95	95	94
	0.2	73	59	61	75	86	95	95	94
	0.3	74	59	58	71	82	94	96	96
0.2	0.0	74	62	60	78	91	95	96	96
	0.1	73	60	61	74	87	95	96	94
	0.2	73	58	57	70	81	93	95	95
	0.3	73	58	56	66	78	92	95	96
0.3	0.0	74	62	60	76	89	95	96	96
	0.1	75	59	58	71	82	93	96	95
	0.2	74	61	56	65	78	90	96	96
	0.3	73	58	55	64	71	88	93	96

Table 2: Coverage (%) of the standard nominal 95% GMM confidence interval for β based on Equation 16. Coverage is calculated only for those simulation draws for which the interval exists. (See Table 1.) Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

cause these results are extremely fast to compute, Table 3 is based on 10,000 simulation replications. Aside from some slight under-coverage at intermediate values of (α_0, α_1) when $\beta = 3$, the GMS interval makes good on its promise of uniformly valid inference. As shown in Appendix E, the under-coverage problem appears to be a finite-sample artifact: if we increase n to 2000, the maximum size distortion becomes negligible. The GMS test tends, however, to be fairly conservative, particularly for larger values of (α_0, α_1). When there is no mis-classification error, the GMS confidence sets are very nearly exact. Results for nominal 95% and 90% intervals are qualitatively similar: see Appendix E.

We now present results for the Bonferroni interval from Algorithm 3.2, setting $\delta_1 = \delta_2 = 0.025$ to yield an interval with asymptotic coverage no less that 95%.[18] Table 4 presents coverage probabilities in percentage points and Table 4 median widths.

The Bonferroni interval achieves its stated minimum coverage uniformly over the parameter space. When there is no mis-classification, $\alpha_0 = \alpha_1$, its actual coverage is close or equal to 95%. In the presence of mis-classification, however, the interval can be quite conservative, particularly for larger values of β. For smaller but nonzero values of β, this conservatism reflects the fact that the model is *effectively* partially identified: although Theorem 2.3 shows

---

[18]In principle, one could optimize the choice of δ_1 subject to the constraint $\delta_1 + \delta_2 = 0.95$ to reduce the width of the resulting interval. In our experiments, there was no choice of δ_1 that uniformly dominated for all values of $(\alpha_0, \alpha_1, \beta)$ so we report only results for $\delta_1 = \delta_2$ here.

α_0	α_1	β							
		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 3: Coverage probability (1 - size) in percentage points of a 97.5% GMS joint test for α_0 and α_1 using Algorithm 3.1 with $n = 1000$. Calculations are based on 10,000 replications of the DGP from Section 4.1.

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 4: Coverage of probability in percentage points of a nominal $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. Calculations are based on 2000 replications of the DGP from Section 4.1.

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 5: Median width of a nominal $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. Calculations are based on 2000 replications of the DGP from Section 4.1.

that (α_0, α_1) are point identified for any $\beta \neq 0$, the amount of data required to distinguish one pair of alphas from another when β is small would be astronomical.

In spite of its conservatism, the Bonferroni interval is informative, as we see from the median widths in Table 5. Because median widths provide only a limited pictures of the behavior of a confidence interval, Figures 1–3 present further evidence in the form of coverage functions (1 - power) for $\beta = 0.5, 1, 3$. Coverage curves for additional values of β and n appear in Appendix E. Each figure holds the true value of β fixed and varies (α_0, α_1) over the grid $\{0, 0.1, 0.2\} \times \{0, 0.2, 0.2\}$. The plots within each Figure give coverage in percentage points as a function of the specified alternative for β. Solid curves are computed using the full set of inequality moment conditions from 3.3, while dashed curves use only m_1^I, i.e. they do not impose the restrictions implied by non-differential measurement error. In each figure, the dashed horizontal line gives the nominal coverage probability, 95%, while the dashed vertical lines are the reduced from and instrumental variables estimands: for $\beta \geq 0$ the reduced form is always smaller than the IV.

As seen from Figures 1–3, and their counterparts in Appendix E, the Bonferroni procedure has power against the alternative $\beta = 0$, even when the true value of β is small. As described in Section 3.5, the Bonferroni interval excludes zero if and only if the confidence interval for θ_1 from which it is constructed also excludes zero. These figures also indicate the gains

from including m_2^I, the moment inequalities that emerge from assuming non-differential measurement error: substantial increases in power against alternatives between the true parameter value and zero, particularly for larger values of β. Note moreover that the excellent performance of Bonferroni in the zero mis-classification case (α_0, α_1) depends crucially on imposing the assumption of non-differential measurement error. As the true value of β increases, the Bonferroni interval begins to have power against both the reduced form and IV estimands.

					β				
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	19.07	3.44	1.86	1.32	0.87	0.47	0.37	0.35
	0.1	17.52	3.47	1.92	1.41	1	0.61	0.51	0.46
	0.2	17.41	3.51	1.9	1.45	1.1	0.76	0.65	0.58
	0.3	18.23	3.34	1.92	1.48	1.24	0.91	0.79	0.7
0.1	0.0	17.13	3.51	1.86	1.38	0.97	0.61	0.51	0.46
	0.1	17.88	3.33	1.85	1.45	1.13	0.78	0.67	0.6
	0.2	17.37	3.36	1.95	1.54	1.24	0.97	0.85	0.75
	0.3	18.07	3.33	1.98	1.63	1.41	1.17	1.04	0.92
0.2	0.0	17.79	3.39	1.92	1.45	1.11	0.75	0.65	0.58
	0.1	18.98	3.43	1.96	1.54	1.26	0.97	0.84	0.75
	0.2	18.25	3.26	1.92	1.64	1.45	1.2	1.06	0.95
	0.3	19.03	3.31	2.02	1.75	1.66	1.49	1.33	1.19
0.3	0.0	18.27	3.48	1.87	1.5	1.25	0.9	0.79	0.7
	0.1	19.4	3.41	1.96	1.63	1.43	1.18	1.04	0.92
	0.2	18.22	3.56	1.96	1.74	1.67	1.49	1.35	1.19
	0.3	17.56	3.55	2.13	1.96	1.86	1.86	1.74	1.55

Table 6: Median width of the standard nominal 95% GMM confidence interval for β based on Equation 16. Coverage is calculated only for those simulation draws for which the interval exists. (See Table 1.) Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

A drawback of the identification-robust inference procedure from 3.2 becomes apparent when both β and the mis-classification probabilities are large. In this case the confidence interval for β is excessively wide, as we see from Table 5 and Figure 3.[19] Note from Tables 1 and 2, that this is a region of the parameter space in which the plain-vanilla GMM confidence interval yields valid inference. Moreover, we see from Table 6 that the median width of the GMM interval is far more reasonable when β is large, even in the presence of large amounts of mis-classification. It is important to stress that the source of this excess width is *not* Bonferroni: the same behavior emerges if one projects a joint GMS confidence set for $(\alpha_0, \alpha_1, \beta)$ to yield marginal inference for β. Rather, it is the inevitable cost of applying a robust inference procedure in a region of the parameter space where standard inference

---

[19]As expected, median widths decrease with sample size: see the results for $n = 2000$ in Appendix E.

α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
					β				
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 7: Coverage probabilities (%) a hybrid confidence interval constructed from the nominal 95% standard GMM interval and the $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. The hybrid interval reports Bonferroni unless the GMM interval exists and is contained within the Bonferroni interval. Calculations are based on 2000 replications of the DGP from Section 4.1.

performs well. While a detailed theoretical investigation of this problem is beyond the scope of the present paper, we now explore the performance of a "hybrid" confidence interval that uses a simple heuristic to transition between robust and standard inference.[20] The procedure for constructing the hybrid interval is as follows. First compute the robust confidence interval based on Algorithm 3.2. Next, determine whether the GMM interval is well-defined: if so, determine whether it is contained within the robust interval. If the GMM interval exists and lies within the robust interval, report GMM; otherwise report the robust interval. Table 7 presents coverage probabilities (in percentage points) and Table 8 median widths for the resulting hybrid confidence interval. Coverage plots for $\beta = 1, 2, 3$ appear in Figures 4–6. Plots for additional values of β and n appear in Appendix E. The conventions of these figures are identical to those of Figures 1–3 with one exception: in Figures 4–6 the dashed curves correspond to the hybrid confidence interval.

The hybrid interval performs extremely well: with the exception of a slight size distortion

---

[20]Although different in its details, this idea bears some similarity to Andrews (2017).

cite!

| | | | | | | β | | | |
α_0	α_1	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 8: Median width of a hybrid confidence interval constructed from the nominal 95% standard GMM interval and the $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000$, $R = 5000$ and $\delta_1 = \delta_2 = 0.025$. The hybrid interval reports Bonferroni unless the GMM interval exists and is contained within the Bonferroni interval. Calculations are based on 2000 replications of the DGP from Section 4.1.

at $(\alpha_0 = \alpha_1 = 0.3, \beta = 1)$ and $(\alpha_0 = \alpha_1 = 0, \beta = 3)$, it is effectively a free lunch.[21] Note in particular that the coverage curves for the hybrid interval from Figures 4–6 (dashed curves) lie uniformly below those of the Bonferroni interval (solid curves) while still maintaining correct coverage at the true value of β. It could be interesting to investigate this idea further in future work.

---

[21]The distortion at $(\alpha_0 = \alpha_1 = 0.3, \beta = 1)$ disappears when n increases to 2000: see Appendix E.
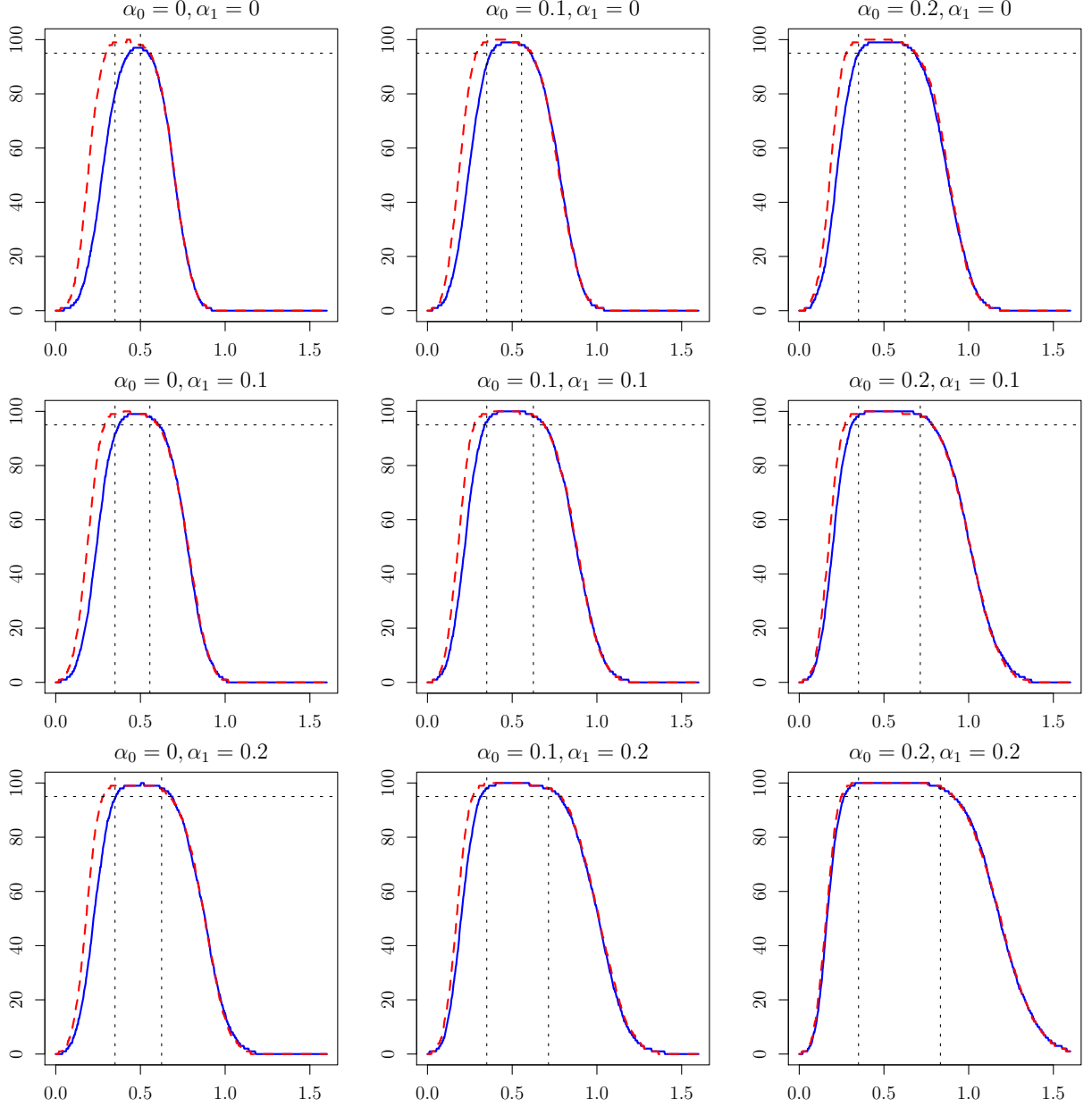
Figure 1: Coverage curves (1 - power) for β when the truth is $\beta = 0.5$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

Figure 2: Coverage curves (1 - power) for β when the truth is $\beta = 1$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

Figure 3: Coverage curves (1 - power) for β when the truth is $\beta = 3$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

Figure 4: Hybrid vs Bonferroni $\beta = 1, n = 1000$

Figure 5: Hybrid vs Bonferroni $\beta = 2, n = 1000$

Figure 6: Two-step vs Bonferroni $\beta = 3, n = 1000$

5   Conclusion

A   Proofs

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

$$P(T^* = 1 | T = 1, Z = k) = P(T = 1 | T^* = 1)\left(\frac{p_k^*}{p_k}\right) = (1 - \alpha_1)\left(\frac{p_k^*}{p_k}\right)$$

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

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

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

Proof of Lemma 2.1. The result follows from a simple calculation using the law of total probability. $\qquad\square$

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

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

$$r_{tk} \equiv \mathbb{P}(T^* = 1 | T = t, z = k) \qquad F_t(\tau) \equiv \mathbb{P}(y \leq \tau | z = k)$$

$$F_{tk}(\tau) \equiv \mathbb{P}(y \leq \tau | T = t, z = k) \qquad F_{tk}^{t^*}(\tau) \equiv \mathbb{P}(y \leq \tau | T^* = t^*, T = t, z = k)$$

$$G_k(\tau) \equiv \mathbb{P}(\varepsilon \leq \tau | z = k) \qquad G_{tk}^{t^*}(\tau) \equiv \mathbb{P}(\varepsilon \leq \tau | T^* = t^*, T = t, z = k).$$

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

$$G_{tk}^0(\tau) = F_{tk}^0(\tau + c), \quad G_{tk}^1(\tau) = F_{tk}^1(\tau + c + \beta) \tag{35}$$

and thus we see that

$$G_k(\tau) = r_{1k}p_k F_{1k}^1(\tau + c + \beta) + r_{0k}(1 - p_k)F_{0k}^1(\tau + c + \beta)$$
$$+ (1 - r_{1k})p_k F_{1k}^0(\tau + c) + (1 - r_{0k})(1 - p_k)F_{0k}^0(\tau + c) \tag{36}$$

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

$$F_{tk}(\tau) = r_{tk}F_{tk}^1(\tau) + (1 - r_{tk})F_{tk}^0(\tau) \tag{37}$$

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

$$r_{1k} = \frac{(1 - \alpha_1)p_k^*}{p_k}, \quad r_{0k} = \frac{\alpha_1 p_k^*}{1 - p_k}. \tag{38}$$

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

Case I: $r_{1k} = 0, r_{0k} \neq 0$ By Equation 38, this requires $\alpha_1 = 1$ which is ruled out by Assumption 2.2 (ii).

Case II: $r_{0k} = r_{1k} = 0$ By Equation 38, this requires $p_k^* = 0$ which in turn requires $p_k = \alpha_0$. Moreover, by Equation 37 we have $F_{tk}^0 = F_{tk}$, while F_{tk}^1 is undefined. Substituting into Equation 36,

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

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 38 this requires $\alpha_1 = 0$ and $p_k^* \neq 0$. By Equation 37 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 36, we obtain

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

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

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

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 37, 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 37 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 36, we have

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

using the fact that $F_k(\tau) = p_k F_{1k}(\tau) + (1-p_k)F_{0k}(\tau)$. 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

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

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 37. For example we could take $F_{tk}^0 = F_{tk}^1 = F_{tk}$. □

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 $\mathbb{E}[y|z = k]$ □

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

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

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

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

41

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

$$\frac{F_{tk}(\tau) - (1 - r_{tk})}{r_{tk}} \leq F_{tk}^1(\tau) \leq \frac{F_{tk}(\tau)}{r_{tk}} \tag{39}$$

Now define

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

Combining Equation 39 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

42

$\underline{F}^1_{tk}$. It follows that

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

But notice that

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

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

Now suppose that $\mu^1_k \in [\underline{m}_{tk}, \overline{m}_{tk}]$. 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 $m_{tk}(\xi^*) = \mu^1_k$. 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

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

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

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

$$\int_{\mathbb{R}} f^0_{tk}(\tau)\, d\tau = \frac{1}{1 - r_{tk}} \int_{I^C_{tk}(\xi^*)} f_{tk}(\tau)\, d\tau = 1$$

where I^C_{tk} is the complement of I_{tk}. And, by construction

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

for any set A. Finally,

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

> 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 $\alpha_0 = \alpha_1 = 0$.

$\square$

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,

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

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

$\alpha_1)\text{Cov}(T^*, z)$ and $\text{Cov}(T^*, z) = \pi/(1 - \alpha_0 - \alpha_1)$. Hence,

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

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)

$$\text{Cov}(T^*\varepsilon, z) = q(1 - q)\left(p_1^* m_{11}^* - p_0^* m_{10}^*\right)$$
$$\text{Cov}(T\varepsilon, z) = q(1 - q)\left\{(1 - \alpha_1)\left[p_1^* m_{11}^* - p_0^* m_{10}^*\right] + \alpha_0\left[(1 - p_1^*)m_{01}^* - (1 - p_0^*)m_{00}^*\right]\right\}$$

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. $\square$

Proof of Lemma 2.4. Throughout this argument we suppress dependence on **x** for simplicity. Since T^* is binary, if follows from the basic properties of covariance that,

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

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

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

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),

$$\text{Cov}(T^*\varepsilon^2, z) = q(1 - q)(p_1^* s_{11}^* - p_0^* s_{10}^*)$$
$$\text{Cov}(T\varepsilon^2, z) = q(1 - q)\left\{(1 - \alpha_1)\left[p_1^* s_{11}^* - p_0^* s_{10}^*\right] + \alpha_0\left[(1 - p_1^*)s_{01}^* - (1 - p_0^*)s_{00}^*\right]\right\}$$

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

$$\text{Cov}(T\varepsilon^2, z) = (1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon^2, z). \tag{40}$$

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

$$\text{Cov}(TT^*\varepsilon, z) = q(1 - q)(1 - \alpha_1)(p_1^* m_{1k}^* - p_0^* m_{10}^*) = (1 - \alpha_1)\text{Cov}(T^*\varepsilon, z) \tag{41}$$

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

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

and combining these equalities with Equations 40 and 41, it follows that

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

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

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

where

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

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

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

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

$$A = \theta_2/\theta_1^2 = 1 + (\alpha_0 - \alpha_1) \tag{42}$$
$$B = \theta_3/\theta_1^3 = (1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1) \tag{43}$$

Equation 42 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 43 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,

$$r = \frac{A}{2} \pm \sqrt{3A^2 - 2B} = \frac{1}{\theta_1^2}\left(\frac{\theta_2}{2} \pm \sqrt{3\theta_2^2 - 2\theta_1\theta_3}\right). \tag{44}$$

By Equations 12 and 13,

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

Expanding the first term we find that

$$3(1 + \alpha_0 - \alpha_1)^2 = 3\left[1 + 2(\alpha_0 - \alpha_1) + (\alpha_0 - \alpha_1)^2\right]$$
$$= 3 + 6\alpha_0 - 6\alpha_1 + 3\alpha_0^2 + 3\alpha_1^2 - 6\alpha_0\alpha_1$$

and expanding the second

$$2\left[(1 - \alpha_0 - \alpha_1)^2 + 6\alpha_0(1 - \alpha_1)\right] = 2\left[1 - 2(\alpha_0 + \alpha_1) + (\alpha_0 + \alpha_1)^2 + 6\alpha_0 - 6\alpha_0\alpha_1\right]$$
$$= 2 + 8\alpha_0 - 4\alpha_1 + 2\alpha_0^2 + 2\alpha_1^2 - 8\alpha_0\alpha_1.$$

Therefore

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

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 $\mathrm{sign}(\beta) = \mathrm{sign}(\theta_1)$ 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. $\qquad\square$

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.[22]

---

[22]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) $\mathbb{E}[\nu|T^*, T, z] = 0$.[23]

(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.[24] 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.[25]

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 ???.*

---

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

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

[25]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,

$$\mathbb{E}[\varepsilon|T^*, z] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*, T, z\right)\right] = \mathbb{E}_{T|T^*,z}\left[\mathbb{E}\left(\varepsilon|T^*\right)\right] = \mathbb{E}\left[\varepsilon|T^*\right] \tag{45}$$

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 $\mathbb{E}[\varepsilon|T^*] \neq 0$. $\qquad\square$

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 $\mathbb{E}[u|T^* = t, z_1] = \mathbb{E}[u|T^* = t, z_2]$ 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.

C 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 $\mathbf{x}$.

C.0.1 Derivations from the Notes

Is there a LATE interpretation of our results? Let $J \in \{a, c, d, n\}$ index an individual's *type*: always-taker, complier, defier, or never-taker. Let $\pi_a, \pi_c, \pi_d, \pi_n$ denote the population proportions of always-takers, compliers, defiers, and never-takers. The unconfounded type assumption is $P(J = j|z = 1) = P(J = j|z = 0)$. Combined with the law of total probability, this gives

$$p_1^* = P(T^* = 1|z = 1) = \pi_a + \pi_c$$
$$1 - p_1^* = P(T^* = 0|z = 1) = \pi_d + \pi_n$$
$$p_0^* = P(T^* = 1|z = 0) = \pi_d + \pi_a$$
$$1 - p_0^* = P(T^* = 0|z = 0) = \pi_n + \pi_c$$

Imposing no-defiers, $\pi_d = 0$, these expressions simplify to

$$p_1^* = \pi_a + \pi_c$$
$$1 - p_1^* = \pi_n$$
$$p_0^* = \pi_a$$
$$1 - p_0^* = \pi_n + \pi_c$$

Solving for π_c, we see that

$$\pi_c = p_1^* - p_0^*$$
$$\pi_a = p_0^*$$
$$\pi_n = 1 - p_1^*$$

Now, let $Y(1)$ indicate the potential outcome when $T^* = 1$ and $Y(0)$ indicate the potential outcome when $T^* = 0$. The standard LATE assumptions (no defiers, mean exclusion, unconfounded type) imply

$$\mathbb{E}\left(Y|T^* = 1, z = 1\right) = \left(\frac{p_0^*}{p_1^*}\right)\mathbb{E}\left[Y(1)|J = a\right] + \left(\frac{p_1^* - p_0^*}{p_1^*}\right)\mathbb{E}\left[Y(1)|J = c\right]$$

$$\mathbb{E}\left(Y|T^* = 0, z = 0\right) = \left(\frac{p_1^* - p_0^*}{1 - p_0^*}\right)\mathbb{E}\left[Y(0)|J = c\right] + \left(\frac{1 - p_1^*}{1 - p_0^*}\right)\mathbb{E}\left[Y(0)|J = n\right]$$

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

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

LATE Version of Theorem 2 from original Draft

$$\Delta \overline{yT} = \mathbb{E}\left(yT|z=1\right) - \mathbb{E}\left(yT|z=0\right)$$
$$= (1-\alpha_1)\left[p_1^*\mathbb{E}\left(y|T^*=1,z=1\right) - p_0^*\mathbb{E}\left(y|T^*=1,z=0\right)\right]$$
$$+ \alpha_0\left[(1-p_1^*)\mathbb{E}\left(y|T^*=0,z=1\right) - (1-p_0^*)\mathbb{E}\left(y|T^*,z=0\right)\right]$$

So we find that

$$\Delta \overline{yT} = (p_1^* - p_0^*)\left\{(1-\alpha_1)\mathbb{E}\left[Y(1)|J=c\right] - \alpha_0\mathbb{E}\left[Y(0)|J=c\right]\right\}$$
$$= (1-\alpha_1)\left\{\frac{\mathbb{E}\left[Y(1)-Y(0)|J=c\right]}{1-\alpha_0-\alpha_1}(p_1-p_0)\right\} + (p_1-p_0)\mathbb{E}\left[Y(0)|J=c\right]$$

Recall that the analogous expression in the homogeneous treatment effect case is

$$\Delta \overline{yT} = (1-\alpha_1)\mathcal{W}(p_1-p_0) + \mu_{10}^*$$
$$= (1-\alpha_1)\left(\frac{\beta}{1-\alpha_0-\alpha_1}\right)(p_1-p_0) + (p_1-\alpha_0)m_{11}^* - (p_0-\alpha_0)m_{10}^*$$

while the expression for the difference of variances is

$$\Delta \overline{y^2} = \beta\mathcal{W}(p_1-p_0) + 2\mathcal{W}\mu_{10}^*$$

From above we see that the analogue of μ_{10}^* in the heterogeneous treatment effects setting is $(p_1-p_0)E\left[Y(0)|J=c\right]$ and since the LATE is $\mathbb{E}\left[Y(1)-Y(0)|J=c\right]$, the analogue of $\mathcal{W}$ is

$$\frac{\mathbb{E}\left[Y(1)-Y(0)|J=c\right]}{1-\alpha_0-\alpha_1}$$

so *if* we could establish that

$$\Delta \overline{y^2} = \left(\frac{p_1-p_0}{1-\alpha_0-\alpha_1}\right)\mathbb{E}\left[Y(1)-Y(0)|J=c\right] \cdot \mathbb{E}\left[Y(1)+Y(0)|J=c\right]$$

in the heterogeneous treatment effects case, the proof of Theorem 2 would go through immediately. Now, if we assume an exclusion restriction on the *second* moment of y an argument almost identical to the standard LATE derivation gives

$$\Delta \overline{y^2} = \frac{\mathbb{E}\left[Y^2(1)-Y^2(0)|J=c\right]}{p_1^*-p_0^*} = \left(\frac{p_1-p_0}{1-\alpha_0-\alpha_1}\right)\mathbb{E}\left[Y^2(1)-Y^2(0)|J=c\right]$$

so we see that the necessary and sufficient condition for our proof to go through is

$$\mathbb{E}\left[Y^2(1)-Y^2(0)|J=c\right] = \mathbb{E}\left[Y(1)-Y(0)|J=c\right] \cdot \mathbb{E}\left[Y(1)+Y(0)|J=c\right]$$

Rearranging, this in turn is equivalent to

$$\text{Var}\left[Y(1)|J=c\right] = \text{Var}\left[Y(0)|J=c\right]$$

D   Moment Conditions for Mahajan and FL

E   Supplementary Simulation Results: Online Only

Give an overview of the results and refer to the sections of the paper above. Also update the captions and add the full complement of figures.

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

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

Table 10: Coverage (1 - size) of 90% GMS joint test for α_0 and α_1: $n = 2000$

α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
					β				
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 11: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 1000$

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

Table 12: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 2000$

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

α_0	α_1	β							
		0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.6	97.6	97.5	97.4	97.5
	0.1	98.0	98.1	98.4	98.3	98.8	98.6	97.8	97.0
	0.2	98.1	98.2	98.8	98.6	98.9	98.6	98.3	97.3
	0.3	98.2	98.5	98.6	98.6	98.8	98.4	98.2	97.4
0.1	0.0	97.4	98.1	98.3	98.8	98.5	98.5	97.9	96.9
	0.1	98.0	98.6	99.1	99.4	99.5	99.3	98.4	96.8
	0.2	98.2	98.9	99.4	99.6	99.7	99.3	98.8	96.8
	0.3	98.6	99.1	99.6	99.8	99.6	99.2	98.4	97.0
0.2	0.0	97.8	98.1	98.5	98.6	98.5	98.4	98.0	97.6
	0.1	98.3	98.9	99.2	99.6	99.5	99.1	98.6	97.0
	0.2	98.7	99.4	99.7	99.6	99.5	99.0	98.4	96.9
	0.3	99.1	99.6	99.7	99.7	99.5	99.0	98.2	97.0
0.3	0.0	98.2	98.3	98.7	98.5	98.6	98.5	98.0	97.7
	0.1	98.6	99.3	99.4	99.6	99.5	99.2	98.1	97.0
	0.2	99.2	99.7	99.7	99.6	99.4	98.8	98.4	97.4
	0.3	99.6	99.8	99.8	99.7	99.4	99.1	98.8	98.2

Table 14: Coverage (1 - size) of 97.5% GMS joint test for α_0 and α_1: $n = 2000$

α_0	α_1	β							
		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 15: 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 ($n = 1000$)

α_0	α_1	β							
		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 16: 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 ($n = 2000$)

α_0	α_1	β							
		0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	74	54	63	87	95	94	96	95
	0.1	72	54	62	86	94	95	95	96
	0.2	72	53	64	85	94	95	95	94
	0.3	73	54	64	81	94	95	95	94
0.1	0.0	73	54	65	83	94	95	94	96
	0.1	74	55	64	84	93	95	95	95
	0.2	72	52	63	80	93	96	95	95
	0.3	75	53	59	77	90	95	95	95
0.2	0.0	74	54	61	84	93	96	95	94
	0.1	74	54	63	81	92	96	95	96
	0.2	73	52	60	75	90	96	96	95
	0.3	74	50	57	72	86	95	96	96
0.3	0.0	74	53	61	83	92	97	95	95
	0.1	75	52	60	78	90	95	96	96
	0.2	73	52	57	73	85	95	96	96
	0.3	73	53	54	69	80	93	96	96

Table 17: Coverage 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	17.4	2.42	1.47	1	0.62	0.33	0.27	0.24
	0.1	16.56	2.51	1.49	1.06	0.7	0.43	0.36	0.33
	0.2	16.33	2.4	1.53	1.13	0.81	0.53	0.46	0.41
	0.3	17.06	2.52	1.57	1.19	0.91	0.65	0.56	0.5
0.1	0.0	17.2	2.5	1.53	1.05	0.71	0.43	0.36	0.33
	0.1	17.48	2.5	1.53	1.15	0.83	0.56	0.48	0.43
	0.2	16.32	2.45	1.57	1.2	0.97	0.69	0.6	0.53
	0.3	18.37	2.43	1.51	1.3	1.1	0.84	0.73	0.65
0.2	0.0	17.64	2.5	1.49	1.13	0.8	0.54	0.46	0.41
	0.1	18.25	2.47	1.58	1.22	0.96	0.69	0.6	0.54
	0.2	17.02	2.4	1.57	1.31	1.13	0.86	0.76	0.67
	0.3	18.05	2.39	1.61	1.43	1.33	1.09	0.95	0.85
0.3	0.0	17.72	2.43	1.53	1.19	0.91	0.65	0.56	0.5
	0.1	18.8	2.46	1.55	1.32	1.11	0.84	0.74	0.65
	0.2	18.24	2.45	1.61	1.45	1.3	1.08	0.96	0.85
	0.3	17.43	2.55	1.67	1.62	1.57	1.4	1.24	1.1

Table 18: 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 19: 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 20: 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 21: 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 22: Median Width of nominal $> 95\%$ Bonferroni Intervals with $n = 2000$

α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
					β				
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 23: Coverage of two-step CI constructed from nominal 95% GMM and nominal $> 95\%$ Bonferroni intervals: $n = 1000$

α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
					β				
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 24: 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 25: 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 26: Median width of two-step CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 2000$

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.

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.

Bugni, F. A., Canay, I. A., Shi, X., 2017. Inference for subvectors and other functions of partially identified parameters in moment inequality models. Quantitative Economics 8 (1), 1–38.

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).

Kaido, H., Molinari, F., Stoye, J., 2016. Confidence intervals for projections of partially identified parameters. arXiv preprint arXiv:1601.00934.

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.