IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15427

Identifying Program Benefits When Participation Is Misreported

Denni Tommasi
Lina Zhang

JULY 2022

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15427

Identifying Program Benefits When Participation Is Misreported

Denni Tommasi
University of Bologna, IZA and CDES

Lina Zhang
University of Amsterdam and Tinbergen Institute

JULY 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Identifying Program Benefits When Participation Is Misreported*

In cases of non-compliance with a prescribed treatment, estimates of causal effects typically rely on instrumental variables. However, when participation is also misreported, this approach can be severely biased. We provide an instrumental variable method that researchers can use to identify the true heterogeneous treatment effects in data that include both non-compliance and misclassification of treatment status. Our method can be used regardless of whether the treatment is misclassified because it is missing at random, missing not at random, or was generally mismeasured. We conclude with the use of a dedicated Stata command, ivreg2m, to assess the return on education in the United Kingdom.

JEL Classification: C14, C21, C26, C35, C51

Keywords: treatment effect, causality, non-differential misclassification, weighted average of LATEs, endogeneity, program evaluation

Corresponding author:
Denni Tommasi
University of Bologna
Piazza Scaravilli 2
Bologna, 40126
Italy
E-mail: denni.tommasi@unibo.it

* We would like to thank Giuseppe Cavaliere, Luca Fanelli, Arthur Lewbel, and Whitney Newey for valuable comments and suggestions and David Slichter for inspiring conversations. The ivreg2m Stata command is available from the SSC repository. All remaining errors are ours.

1 Introduction

In cases of non-compliance with a prescribed treatment, estimates of causal effects typically rely on instrumental variables (Athey and Imbens, 2017). However, this approach requires participation in a program to be correctly measured, which is not the case in many applications. We review articles published between 1996 and 2022 and find 54 reportings of treatment misclassification in most fields of applied economics.[1] Strikingly, studies evaluating the benefits of a program find or estimate a median misclassification rate of 27%. Misreporting is common in research on educational attainment, participation in specific training courses, and insurance possession.[2] Because the misclassification of a binary treatment is nonclassical, evaluating the benefits of a program using standard techniques can be severely biased (Kreider, 2010; Millimet, 2011).

We study the identification and estimation of program benefits when both non-compliance and misclassification of a binary treatment variable are present. We focus on non-differential misclassification errors that (after conditioning on the true treatment status) are uncorrelated with the outcomes. Our method identifies program benefits in cases of random and nonrandom missing treatment observations, as well as generally mismeasured treatment, whether these cases are due to recording mistakes, imperfect compliance, poor recollection, or incomplete awareness of the treatment(s) received. Recent scholarship addresses the problem of non-differential misclassification of binary treatment variables. Calvi, Lewbel, and Tommasi (2021) propose a novel instrumental variables (IV) approach to identify the average causal effect on compliers. Their estimator, the mismeasured robust local average treatment effect (MR-LATE), requires two binary proxies for the same treatment, which can be constructed using estimated treatments, different sources of treatment status, or multiple or repeated treatment measures. The flexibility and simplicity of MR-LATE makes it potentially easier for practitioners to adopt than alternative solutions that have been proposed in the literature (e.g., Battistin et al., 2014; Yanagi, 2018; DiTraglia and García-Jimeno, 2019; Kasahara and Shimotsu, 2021).[3]

We focus on the IV estimand that captures the average causal effect on compliers (Imbens and Angrist, 1994). We begin by clarifying the consequences of dropping or imputing observations with missing or unclear treatment statuses, which is common in applied studies with misreported program participation. Second, we develop an IV method that researchers can use to identify the true heterogeneous treatment effects. Through this process, we generalize the MR-LATE approach by incorporating discrete or multiple-discrete IVs and targeting the weighted average of LATEs (WLATE). Such a generalization is important because more than half of the empirical papers using IVs published in top journals in the last 20 years use multiple instrumental variables (Mogstad et al., 2020b).

[1] See Table A1 in the Appendix for the complete list.

[2] Meyer et al. (2015) also supported this qualitative result, showing that the "desire to shorten the time spent on the interview, the stigma of program participation, the sensitivity of income information, or changes in the characteristics of those who receive transfers" all drive misreporting of program participation and that it is an increasing problem for social scientists (p. 219).

[3] More specifically, point identification of our target parameter(s) is achieved assuming resurvey data (Battistin et al., 2014), two instrumental variables (Yanagi, 2018), homogeneous treatment effects (DiTraglia and García-Jimeno, 2019), or a covariate satisfying an exclusion restriction from the misclassification (Kasahara and Shimotsu, 2021). A different line of research addresses the more general case of differential misclassification (e.g., Kreider et al., 2012; Nguimkeu et al., 2018; Ura, 2018; Jiang and Ding, 2020; Tommasi and Zhang, 2020).

We provide sufficient conditions on the misclassification rates under which the MR-LATE ensures point identification or set identification results. As a further extension, we show how to apply MR-LATE in a discrete treatment setting (Angrist and Imbens, 1995). Third, we complement the MR-LATE approach by providing a novel inferential procedure that is suitable for many applications with a general data dependence structure, including heteroskedasticity and clustering. Finally, we develop a new Stata command, `ivreg2m`, that implements the proposed method and use it to re-assess the return on education in the United Kingdom (UK).

Our method has three potential applications in the context of treatment non-compliance and non-differential misclassification. First, it can be used as the primary identification strategy in studies when it is certain that the available treatment measurement(s) are unreliable. In particular, it can be used regardless of whether the misclassification is due to the treatment being missing at random, missing not at random, or generally mismeasured. Second, it can be adopted as the primary robustness check when practitioners doubt the accuracy of the treatment variable. Third, practitioners can use it to assess the sensitivity of a program's benefits to different hypothetical values or external information on the extent of misclassification. Although it is motivated by the program evaluation literature, our method may not necessarily apply to evaluating a particular program; nevertheless, it is useful whenever researchers aim to study the causal effect of an endogenous treatment variable that suffers from measurement errors.

This paper also contributes to a long-standing tradition in the program evaluation literature addressing the problem of misclassification in the treatment variable (e.g., Angrist and Krueger, 1999; Bound et al., 2001; Card, 2001; Black et al., 2003; Hernandez et al., 2007; Molinari, 2008, 2010)). In the context of homogeneous treatment effects, papers studying the effects of an exogenous and mismeasured binary treatment using IV techniques include Aigner (1973), Bollinger (1996), Kane et al. (1999) and Black et al. (2000). Under more general conditions, Klepper (1988) provides bounds on average treatment effects with multiple misclassified treatments. In the framework of heterogeneous treatment effects, studies addressing treatment misclassification include Mahajan (2006), Lewbel (2007), and Hu (2008). They provide point-identification results for average treatment effects in instances where the treatment is assumed to be exogenous and instruments are adopted to deal with treatment misclassification.

With respect to the above literature, we develop our approach using the standard LATE structure. This is meaningful because, as Imbens (2014) explains, "under weak conditions" the LATE (or the WLATE) "may be the only relevant information that is credibly identifiable." Similarly, Mogstad et al. (2018) note that it "is of intrinsic interest when the instrument itself represents an intervention, like a policy change or a randomized control trial". In situations where the causal effect among those who comply might not be the effect of interest, the LATE (or WLATE) could be used to extrapolate to more general causal effects (e.g., Heckman et al., 2006). Our method can also be viewed as an alternative to Acerenza et al. (2021) and Possebom (2021) who study treatment misclassification via the marginal treatment effect approach.

The remainder of this paper proceeds as follows. Section 2 shows the limits of some common practices in applied works and reviews the MR-LATE approach for a binary instrument. Section 3 generalizes the MR-LATE to cases with discrete or multiple discrete instrument(s) and provides the asymptotic results. Whereas, simulations, guidance for practitioners and an empirical illustration using our new `ivreg2m` command appear in Section 4. Section 5 concludes. Proofs and additional materials are in the Appendix.

2 Treatment Misclassification and the MR-LATE Approach

We start motivating the paper by looking at a simple setting of a binary treatment and a binary instrument. We describe the consequences of some common cases of treatment misclassification. Then, we review the key idea of the MR-LATE approach introduced by Calvi, Lewbel, and Tommasi (2021), and applied in Tommasi (2019), which addresses the problem of treatment misclassification in this simple setting. This is a natural starting point because later we generalize the approach to other settings and develop a novel inferential procedure.

Let D be a binary treatment variable and Z be a binary instrument. Denote by D_0 and D_1 the two binary potential treatments. Then we have $D = ZD_1 + (1-Z)D_0$ and the outcome can be defined as $Y = DY_1 + (1-D)Y_0$, where Y_1 and Y_0 are the two potential outcomes. In case of non-compliance with prescribed treatment, if there is no treatment misclassification, the Imbens and Angrist (1994)'s local average treatment effect (LATE) is identified by the standard IV estimand:

$$\alpha^{IV} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(D,Z)} = \mathbb{E}[Y_1 - Y_0 | C] \tag{1}$$

where $C = \{D_0 = 0, D_1 = 1\}$ denotes the sub-population of compliers. To identify program benefits in case of non-compliance, we just need to observe Y, D, and Z, and apply 2SLS. However, in many studies of program evaluation, we cannot observe the true treatment D for some observations. Suppose the self-reported answer P to a survey question of program participation is available, which reveals incomplete information about the true treatment status. For example, $P = 1$ if an individual reports "treated", $P = -1$ if "not treated", and $P = 0$ if the record is missing (equivalently, the individual reports "do not know").

Case 1: Missing Treatment Observations at Random. Consider the scenario where the self-reported treatment statuses are correct for $P = 1$ or $P = -1$, but the treatment observations with $P = 0$ are missing at random. This is the case of recording mistakes where $(P = 0) \perp (Y, D, Z)$. We can set the observable treatment indicator $T = 1$ if $P = 1$, and $T = 0$ if $P = -1$. Then, the LATE estimand can be identified if one simply discards all the observations missing at random and conduct the standard 2SLS using the sub-sample with $P \neq 0$:

$$\alpha_1 = \frac{\mathrm{Cov}(Y,Z|P \neq 0)}{\mathrm{Cov}(T,Z|P \neq 0)} = \frac{\mathrm{Cov}(Y,Z|P \neq 0)}{\mathrm{Cov}(D,Z|P \neq 0)} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(D,Z)} = \mathbb{E}[Y_1 - Y_0|C]. \tag{2}$$

This is because, dropping the observations with $P = 0$, neither changes the underlying distribution of the population, nor introduces measurement errors in the observed treatment indicator.

Case 2: Missing Treatment Observations Not at Random. In practice, missing treatment observations are often not at random: $(P=0) \not\perp (Y, D, Z)$. For example, when individuals are asked to report their educational qualifications, the missing or uncertain responses ("do not know") may occur with a higher probability for those who have undertaken some schooling but they are not sure if it counts as a qualification, than those who have never done so. In this case, dropping observations with missing treatment would bias the estimation. When imputing the treatment is not possible, one could simply regard the missing or uncertain responses as untreated and construct a treatment indicator $T=1$ if $P=1$ and $T=0$ if $P=\{-1,0\}$. Then, one would implement the 2SLS estimation using T:

$$\alpha_2 = \frac{\text{Cov}(Y,Z)}{\text{Cov}(T,Z)} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1]-\mathbb{E}[T|Z=0]} = \frac{\mathbb{E}[Y_1-Y_0|C]}{p_1-p_0}, \tag{3}$$

where $p_1=\mathbb{E}[T|D=1,C]$ and $p_0=\mathbb{E}[T|D=0,C]$ describe the measurement accuracy. However, since $0 \le |p_1-p_0| \le 1$, the estimand α_2 would be inflated (in absolute value) by the misclassification error, resulting in a biased approach.

Suppose the self-reported treatment statuses are correct when $P=1$ or $P=-1$. In such cases, the MR-LATE approach can be used to overcome the problem of missing treatment observations when $P=0$ are missing not at random. Construct two binary variables $T^a=1[P=1]$, and 0 otherwise, and $T^b=1[P=-1]$, and 0 otherwise. T^a never mistakes the true untreated as treated and $1-T^b$ never mistakes the true treated as untreated. Under the key assumption of non-differential misclassification, $\mathbb{E}[Y|D,P,C]=\mathbb{E}[Y|D,C]$, it can be shown that the MR-LATE estimand ρ point identifies the LATE:

$$\rho = \frac{\text{Cov}(T^aY,Z)}{\text{Cov}(T^a,Z)} - \frac{\text{Cov}\left(T^bY,Z\right)}{\text{Cov}(T^b,Z)} = \mathbb{E}[Y_1-Y_0|C]. \tag{4}$$

Intuitively, the MR-LATE estimand is the difference between two IV estimands of YT^a on T^a and of YT^b on T^b, both using Z as instrument. By the construction of T^a and T^b, ρ is closely linked to the local average response function (Abadie, 2003). See Appendix A.1 for a more detailed discussion.

Case 3: Generally Mismeasured Treatment. Consider the more general scenario where also the self-reported treatment statuses $P=1$ and $P=-1$ might suffer from misclassification. This is the case, for example, when the binary treatment, such as educational qualification, is constructed by researchers based on years of schooling, but the years are potentially reported with errors. Implementing the naïve 2SLS estimation ignoring the measurement errors as in (3) would be again a biased approach.

Fortunately, in many applications, the treatment is usually either completely unknown or middling for some people, whereas for others the treatment is highly likely or highly unlikely. That is, the observable treatment is still informative about the actual treatment. If so, then the MR-LATE approach is useful to reduce the estimation bias. Set $T^a = 1$ only for everyone with high probability of being treated, and set $T^b = 1$ only for those with low probability of being treated. For example, we can define $T^a = 1[P = 1]$ and $T^b = 1[P = -1]$. If, with relatively small chances, T^a mistakes the true untreated as treated and $1 - T^b$ mistakes the true treated as untreated, the MR-LATE estimand ρ can be shown to be less biased than the naïve estimand in (3):

$$|\rho - \alpha^{IV}| < |\alpha_2 - \alpha^{IV}|.$$

The three cases above nest each other, Case 1 $\subseteq$ Case 2 $\subseteq$ Case 3, therefore the MR-LATE estimand can also be used to point identify the LATE in the first case of missing treatment observations at random. Importantly, the MR-LATE is applicable to reduce estimation bias in other settings whenever individuals' observed information can help sorting them into groups with high ($T^a = 1$) and low ($T^b = 1$) probabilities of being actually treated. For instance, when multiple treatment proxies agree with each other, or when one discrete/continuous treatment proxy takes its largest or smallest values. For those with unknown or middling probabilities, the practictioner can simply set $T^a = T^b = 0$.

3 Generalizing the MR-LATE Approach

This section proceeds in four acts. First, we introduce the target estimand and assumptions. Second, we generalize the MR-LATE approach to allow for general cases where the instrument(s) can be discrete or multiple-discrete. Third, we develop a novel inferential procedure. Finally, we extend the MR-LATE to accommodate covariates and a discrete treatment.

3.1 Setup

Consider a general model setup where D denotes the true binary treatment variable and $Z \in \Omega_Z = \{z_0, z_1, ..., z_K\}$ is a $h \times 1$ vector of discrete instruments with $z_k \in \mathbb{R}^h$. Let $\pi_k = \Pr(Z = z_k)$. Denote D_k, for $k = 0, 1, ..., K$, as potential treatments for $Z = z_k$. By definition,

$$D = \sum_{k=0}^{K} 1[Z = z_k] D_k,$$

where $1[\cdot]$ stands for the indicator function. Let $Pr(z_k) = \mathbb{E}(D|Z = z_k)$ be the propensity score. Then, the outcome of interest is

$$Y = DY_1 + (1 - D)Y_0,$$

with Y_d the potential outcome for possible realization d of D. A known function $g : \Omega_Z \mapsto \mathbb{R}$ can be used to exploit the multiple instruments. We may simply set $g(z) = \Pr(z)$. Let us introduce some basic assumptions.

Assumption 3.1. *Y, D and Z satisfy the standard Imbens and Angrist (1994) LATE assumptions:*

(i) *$(Y_1, Y_0, \{D_k\}_{k=0}^K, Z)$ are i.i.d. and have finite first and second moments;*

(ii) *(Unconfoundedness) $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K)$ and $\mathbb{E}(D|Z=z)$ is a nontrivial function of z;*

(iii) *(First stage) $Cov(D, g(Z)) \neq 0$;*

(iv) *(Monotonicity) For any $z_l, z_w \in \Omega_Z$, with probability one, either $D_l \geq D_w$ for all individuals, or $D_l \leq D_w$ for all individuals. Furthermore, for all $z_l, z_w \in \Omega_Z$, either $Pr(z_l) \leq Pr(z_w)$ implies $g(z_l) \leq g(z_w)$, or $Pr(z_l) \leq Pr(z_w)$ implies $g(z_l) \geq g(z_w)$.*

Note that when there are more than one instrument, the monotonicity condition in (iv) requires the homogeneity of treatment choice behavior (Mogstad et al., 2020a,b). Given Assumption 3.1, the instrumental variable estimand can be expressed as a weighted average of LATEs:

$$\alpha^{IV} = \frac{\mathrm{Cov}(Y, g(Z))}{\mathrm{Cov}(D, g(Z))} = \frac{\mathbb{E}[(Y-\mathbb{E}(Y))(g(Z)-\mathbb{E}[g(Z)])]}{\mathbb{E}[(D-\mathbb{E}(D))(g(Z)-\mathbb{E}[g(Z)])]} = \sum_{k=1}^{K} \gamma_k^{IV} \alpha_{k,k-1},$$

where $\alpha_{k,k-1} = \mathbb{E}[Y_1 - Y_0|C_k]$ is the LATE for compliers $C_k = \{D_k = 1, D_{k-1} = 0\}$ whose treatment status is changed due to the change in the instrument from z_{k-1} to z_k, the weight $\gamma_k^{IV} = \frac{\Pr(C_k)\sum_{l=k}^K (g(z_l)-\mathbb{E}[g(Z)])\pi_l}{\sum_{m=1}^K \Pr(C_m)\sum_{l=m}^K (g(z_l)-\mathbb{E}[g(Z)])\pi_l}$ are nonnegative and $\sum_{k=1}^K \gamma_k^{IV} = 1$.

Suppose the actual treatment D is unobservable and a potentially misclassified treatment indicator T is available, which could come from a treatment proxy or self-reported treatment status. By definition, we have

$$T = DT_1 + (1-D)T_0,$$

where T_0 and T_1 can be interpreted as indicators of false positive and false negative misclassification. If $T_0 = 1$, then a true untreated $D = 0$ is misclassified as treated (false positive), and if $T_1 = 0$, then a true treated $D = 1$ is misclassified as untreated (false negative). In a extreme case, where $T_0 = 0$ and $T_1 = 1$, the true treatment $D = T$ is not misclassified. Denote

$$p_{1,k} = \mathbb{E}[T_1|C_k] \ \text{ and } \ p_{0,k} = \mathbb{E}[T_0|C_k],$$

where $p_{1,k}$ is the probability that treated compliers C_k would have their treatment correctly indicated by T, while $p_{0,k}$ is the probability that untreated compliers C_k would be mistakenly indicated as treated by T.

Assumption 3.2. *The treatment indicator T is such that the following conditions are satisfied:*

(i) *(Extended unconfoundedness) $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, T_1, T_0)$;*

(ii) *(Extended first stage) $Cov(T, g(Z)) \neq 0$.*

Assumption 3.2 extends the unconfoundedness of the instrument Z in the presence of treatment misclassification. In addition, it also ensures that the observed treatment indicator contains relevant information regarding the true treatment status which is a minimal relevance condition to guarantee meaningful estimation.

Under Assumptions 3.1 and 3.2, Tommasi and Zhang (2020) show that naïvely computing the IV estimand using the proxy T in place of D leads to a new parameter:

$$\alpha^{Mis} = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(T, g(Z))} = \frac{\mathbb{E}[(Y-\mathbb{E}(Y))(g(Z)-\mathbb{E}[g(Z)])]}{\mathbb{E}[(T-\mathbb{E}(T))(g(Z)-\mathbb{E}[g(Z)])]} = \sum_{k=1}^{K} \gamma_k^{Mis} \alpha_{k,k-1},$$

where $\gamma_k^{Mis} = \frac{\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}{\sum_{m=1}^{K}(p_{1,m}-p_{0,m})\Pr(C_m)\sum_{l=m}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}$ could be positive or negative depending on the degree of misclassification. The relationship between α^{Mis} and α^{IV} can be characterized as

$$\alpha^{IV} = \xi\alpha^{Mis}, \quad \text{with} \quad \xi = \sum_{k=1}^{K} \gamma_k^{IV}(p_{1,k}-p_{0,k}) \quad \text{and} \quad 0 < |\xi| \leq 1. \tag{5}$$

If the treatment is misclassified, $\xi \neq 1$, hence $|\alpha^{IV}| < |\alpha^{Mis}|$ and α^{IV} is not identifiable by the standard IV approach.

In empirical settings, the parameter ξ is useful for assessing the sensitivity of the standard IV approach with respect to the treatment misclassification. Intuitively, ξ can be rewritten as

$$\xi = 1 - w^n - w^p, \tag{6}$$

where $w^n = \sum_{l=k}^{K}\gamma_k^{IV}(1-p_{1,k})$ is the average probability of treated individuals misclassified as untreated (false negative) and $w^p = \sum_{l=k}^{K}\gamma_k^{IV}p_{0,k}$ is the average probability of untreated individuals misclassified as treated (false positive). The bias in α^{Mis} is $\alpha^{Mis} - \alpha^{IV} = (1/\xi - 1) \times \alpha^{IV}$. If there is no misclassification ($w^n = w^p = 0$), $\xi = 1$ and α^{Mis} collapses to α^{IV}. If misclassification worsens ($w^n > 0, w^p > 0$), the value of ξ falls and the bias in α^{Mis} becomes severe. When $0 < \xi < 1$, α^{Mis} and α^{IV} are of the same sign. The breakdown point at which α^{Mis} and α^{IV} starts to have opposite signs is when ξ turns negative. In practice, if possible values of the average probabilities of false positive and/or false negative are available, for example, from validation studies or external information, then they can be used to compute the bias of the naïve estimation ignoring the treatment misclassification.

3.2 Main Results

Assume we observe *two* different indicators T^a and T^b for each latent treatment status, $D = 1$ and $D = 0$, respectively. For the moment, one can think of these two variables being built upon two different (binary) proxies, two repeated measurements, or an estimated treatment propensity, of the same underlying true treatment variable. Later we provide some examples to precisely illustrate the construction of T^a and T^b. Let

$$T^a = DT_1^a + (1-D)T_0^a, \quad T^b = DT_1^b + (1-D)T_0^b,$$

where T_d^a and T_d^b with $d \in \{0,1\}$ are the misclassification indicators associated with T^a and T^b, respectively. For $j = a, b$, denote

$$\lambda^j = \frac{\text{Cov}(YT^j, g(Z))}{\text{Cov}(T^j, g(Z))}, \quad \lambda_k^j = \frac{\mathbb{E}[YT^j|Z = z_k] - \mathbb{E}[YT^j|Z = z_{k-1}]}{\mathbb{E}[T^j|Z = z_k] - \mathbb{E}[T^j|Z = z_{k-1}]}.$$

Note that the idea of the MR-LATE method is to employ λ_k^a and λ_k^b to approximate the local average response functions $\mathbb{E}[Y_1|C_k]$ and $\mathbb{E}[Y_0|C_k]$ (Abadie, 2003), respectively:

$$\mathbb{E}[Y_1|C_k] = \frac{\mathbb{E}[YD|Z = z_k] - \mathbb{E}[YD|Z = 0]}{\mathbb{E}[D|Z = z_k] - \mathbb{E}[D|Z = z_{k-1}]},$$
$$\mathbb{E}[Y_0|C_k] = \frac{\mathbb{E}[Y(1-D)|Z = z_k] - \mathbb{E}[Y(1-D)|Z = z_{k-1}]}{\mathbb{E}[(1-D)|Z = z_k] - \mathbb{E}[(1-D)|Z = z_{k-1}]}.$$

In addition, we aim to use λ^a and λ^b for the approximation of

$$\frac{\text{Cov}(YD, g(Z))}{\text{Cov}(D, g(Z))} \text{ and } \frac{\text{Cov}(Y(1-D), g(Z))}{\text{Cov}(1-D, g(Z))}, \text{ respectively,}$$

so that the difference between them gives the IV estimand α^{IV}. The estimand λ^j stands for IV estimand of YT^j on T^j using $g(Z)$ as an instrument. Besides, λ_k^j is similarly defined for subgroup C_k. We introduce the following key assumptions on T^j, with $j = a, b$, for the identification of α^{IV} in our general setting.

Assumption 3.3. *For $j = a, b$ and $k = 1, 2, ..., K$, we have*

(i) (Non-differential misclassification) $\mathbb{E}[Y|T^j, D, C_k] = \mathbb{E}[Y|D, C_k]$;

(ii) (Homogeneous misclassification) $\mathbb{E}\left[T^j|D, C_k\right] = \mathbb{E}[T^j|D]$.

Assumption 3.3-(i) is similar to the non-differential misclassification assumption of, e.g., Lewbel (2007), Hu (2008) and Battistin and Sianesi (2011), and is weaker than Assumption 2-(ii) of Calvi, Lewbel, and Tommasi (2021). It says that, for compliers C_k, given the actual treatment status D, the proxies T^a and T^b contain no extra information about the mean of the outcome. Whereas, Assumption 3.3-(ii) requires that, conditional on the actual treatment D, the potential treatments (D_k, D_{k-1}) do not contain information that may affect the proxy T^j. Put it differently, this assumption is satisfied if, for any individuals, e.g., a complier, an always taker, and a never taker, their treatment misclassification probabilities are the same if they are both treated or untreated.

The non-differential and homogeneous misclassification in Assumption 3.3 make it clear that two types of measurement errors are allowed. The first type is *missing (or misreporting) at random*. That is, errors that are independent of observed and unobserved variables, especially independent of the true treatment status. For example, suppose for $j = a, b$,

$$T^j = D(1-\varepsilon^j) + (1-D)\varepsilon^j, \quad \text{with} \quad \varepsilon^j \in \{0,1\} \quad \text{and} \quad \varepsilon^j \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, Z),$$

where, as an example, if T^a is a proxy for $D=1$, $\varepsilon^a=1$ simultaneously indicates the false negative misclassification for a true treated and also the false positive misclassification for a true untreated. In this example, the misclassification error ε^j is a random error, and the misclassification rate of a true treated, $\Pr(T^j=0|D=1)$, is the same as the misclassification rate of a true untreated, $\Pr(T^j=1|D=0)$, which are both equal to $\Pr(\varepsilon^j=1)$.

The second type of error includes those *missing (or misreporting) not at random*. That is, the misreporting behavior may depend on the actual treatment status. For example, individuals who have undertaken some job-related training may be more likely to under-report than those who have never done so, due to incomplete awareness. In this case, suppose, for $j=a,b$,

$$T^j = DT_1^j + (1-D)T_0^j, \quad \text{with} \quad T_1^j, T_0^j \in \{0,1\} \quad \text{and} \quad (T_1^j, T_0^j) \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, Z),$$

where, if, as an example, T^a is a proxy for $D=1$, then $1-T_1^a$ and T_0^a separately indicate the false negative and false positive misclassification, and do not necessarily follow the same distribution: the probability of false negative, $\Pr(T_1^a=0|D=1)$, and the probability of false positive, $\Pr(T_0^a=1|D=0)$, can be different. Other examples for the second type of error may include recording mistakes, imperfect compliance or information poorly recalled.

For $j=a,b$ and $k=1,2,...,K$, let $p_{1,k}^j = \mathbb{E}[T_1^j|C_k]$ and $p_{0,k}^j = \mathbb{E}[T_0^j|C_k]$. Lemma 3.1 below shows that the homogeneous misclassification assumption leads to invariant misclassification probabilities for all complier groups.

Lemma 3.1. *Under Assumption 3.3-(ii), there exist two constants $0 \le p_1^j, p_0^j \le 1$, such that $p_{1,k}^j = p_1^j$ and $p_{0,k}^j = p_0^j$ for $j=a,b$ and all $k=1,2,...,K$.*

Proof of Lemma 3.1. See Appendix A.2.1. □

Denote $q_k^j = \frac{p_{1,k}^j}{p_{1,k}^j - p_{0,k}^j}$ for $k=1,2,...,K$ and $q^a = p_1^a/(p_1^a - p_0^a)$ and $q^b = p_1^b/(p_1^b - p_0^b)$, which are defined only if $p_{1,k}^j - p_{0,k}^j \neq 0$ and $p_1^j - p_0^j \neq 0$, respectively.

Theorem 3.1. *Let Assumptions 3.1, 3.2 and 3.3-(i) hold. We have that for $k=1,2,...,K$,*

$$\lambda_k^a - \lambda_k^b = (q_k^a - q_k^b)\mathbb{E}[Y_1 - Y_0|C_k] = (q_k^a - q_k^b)\alpha_{k,k-1}.$$

If we further assume Assumption 3.3-(ii), then

$$\rho = \lambda^a - \lambda^b = (q^a - q^b)\alpha^{IV}.$$

Proof of Theorem 3.1. See Appendix A.2.2. □

Theorem 3.1 generalizes the MR-LATE method of Calvi, Lewbel, and Tommasi (2021) to cases with discrete and multiple-discrete instrument(s). It shows that each LATE is a linear combination of $(\lambda_k^a, \lambda_k^b)$ of the same complier group, and the IV estimand α^{IV} is a linear combination of (λ^a, λ^b).

3.3 Treatment Misclassification and the MR-LATE in General Settings

Consider the general setting with discrete or multiple-discrete instrument(s). Similarly to the discussion in Section 2, the MR-LATE approach can be used either to point identify the target parameter, α^{IV}, or as a bias reduction method. We establish its usefulness for both the case of missing treatment observations and generally mismeasured treatment.

Missing Treatment Observations (Case 1 and 2). Suppose some individuals' treatment statuses are missing, either at random or not at random, while others can be correctly observed. In this case we can construct two indicators, $T^a = D$ and $T^b = 1 - D$, if the treatment is observed and we set them to be zero if we are missing treatment information. Hence, both indicators contain one-sided misclassification error. T^a may mistake a true treated as untreated, but it would not mistake true untreated as treated. Conversely, $1 - T^b$ only has the opposite kind of measurement error, as it may mistake true untreated as treated, but it would not mistake true treated as untreated. The following assumption summarizes the above paragraph.

Assumption 3.4. *(One Type of Misclassification Error) Under Assumption 3.3-(ii), further assume that $p_0^a = p_1^b = 0$, and $p_1^a > 0$, $p_0^b > 0$.*

Assumption 3.5 is an analog of the standard no defiers assumption in the treatment effect literature. The latter indicates that certain combinations of D and Z never occur. The former requires a zero probability of certain combinations of D and T^a, and of D and T^b. If a data set justifies Assumption 3.5, Theorem 3.1 has some straightforward implications.

Corollary 3.1 (**Point Identification**). *Let Assumptions 3.1 to 3.5 hold for T^a and T^b. Then, we have*

$$\lambda_k^a - \lambda_k^b = \alpha_{k,k-1}, \quad \text{and} \quad \rho = \lambda^a - \lambda^b = \alpha^{IV}.$$

The Corollary above presents the point identification result for LATEs and IV estimand α^{IV} using the MR-LATE approach in the presence of discrete or multiple discrete instrument(s). Note that missing at random may not be easy to verify in practice. Since missing at random is a special case of missing not at random, the generalized MR-LATE can be a reliable alternative to the method of dropping or imputing missing treatment observations.

Generally Mismeasured Treatment (Case 3). The one type of misclassification error assumption is powerful for point identifying the parameter of interest. However, it may not hold in some applications with generally mismeasured treatment. Fortunately, even if available information cannot guarantee Assumption 3.5, it is still possible to use MR-LATE to set identify α^{IV}.

Corollary 3.2 (**Set Identification**). *Let Assumptions 3.1 to 3.3 hold for T^a and T^b. If $q^a - q^b > 0$, then ρ signs α^{IV}. If $q^a - q^b \geq 1$, then α^{IV} lies between zero and ρ.*

A sufficient condition for $q^a - q^b \geq 1$ is $p_1^a > p_0^a$ and $p_0^b > p_1^b$, which relaxes the one type of misclassification error $p_0^a = p_1^b = 0$ but still requires T^a and T^b to be informative about the actual treatment status. Specifically, $p_1^a > p_0^a$ states that the share of true treated in T^a, $\Pr(T^a = 1|D = 1)$, is larger than the share of misclassified true untreated, $\Pr(T^a = 1|D = 0)$. Analogously, $p_0^b > p_1^b$ says that the share of true untreated in T^b, $\Pr(T^b = 1|D = 0)$, is larger than the share of misclassified true treated, $\Pr(T^a = 1|D = 1)$. Note that the same arguments can be applied to describe the relationship of q_k^a, q_k^b with the LATE, $\alpha_{k,k-1}$.

Given Corollary 3.3, we can show that under mild conditions, the MR-LATE estimand ρ is less biased (in absolute value) than α^{Mis}. Since two treatment indicators T^a and T^b are available, when comparing the bias, we consider $T = rT^a + (1-r)T^b$ with $r \in \{0,1\}$. Then, we have

$$\alpha^{Mis} - \alpha^{IV} = \left[\frac{1}{r(p_1^a - p_0^a) + (1-r)(p_0^b - p_1^b)} - 1\right]\alpha^{IV}, \text{ and } \rho - \alpha^{IV} = \left[\frac{p_0^a}{p_1^a - p_0^a} + \frac{p_1^b}{p_0^b - p_1^b}\right]\alpha^{IV}.$$

Corollary 3.3 (**Bias Reduction**). *Let Assumptions 3.1 to 3.3 hold for T^a and T^b. Assume $p_1^a > p_0^a$ and $p_0^b > p_1^b$. If*

$$p_0^a + p_1^b < \left(\frac{1}{\max\{p_1^a - p_0^a,\, p_0^b - p_1^b\}} - 1\right)\min\{p_1^a - p_0^a,\, p_0^b - p_1^b\},$$

then $|\rho - \alpha^{IV}| < |\alpha^{Mis} - \alpha^{IV}|$ where α^{Mis} can be based on either T^a or T^b.

Proof of Corollary 3.3. See Corollary 4 in Calvi, Lewbel, and Tommasi (2021). □

Essentially, Corollary 3.3 provides an upper bound for the summation of misclassification probabilities in the two indicators, $p_0^a + p_1^b$, so that the bias of ρ is smaller than that of the naïve estimand. Note that T^a and T^b are constructed in such a way to ensure small chances of the undesirable misclassification. Therefore, such a condition is easily satisfied and MR-LATE can be applied in many empirical studies as a powerful bias reduction approach.

3.4 Inference

Denote by $W_i = \{Y_i, T_i^a, T_i^b, Z_i\}_{i=1}^n$ the observations for individuals $i = 1, \ldots, n$. To estimate ρ, we need to conduct two IV regressions. First, the regression of $Y_i T_i^a$ on T_i^a and a constant using $g(Z_i)$ as an instrument gives us $\hat{\lambda}_n^a$. Second, following the same procedure using $Y_i T_i^b$ and T_i^b, we obtain $\hat{\lambda}_n^b$. We assume the function $g(Z) = g(Z;\theta)$ is known up to an unknown parameter $\theta \in \mathbb{R}^{d_\theta}$, and $g(Z)$ can be estimated as $\hat{g}(Z_i) := g(Z;\hat{\theta}_n)$. Formally, let $\hat{\rho}_n = \hat{\lambda}_n^a - \hat{\lambda}_n^b$, and for $j \in \{a,b\}$, where

$$\hat{\lambda}_n^j = \frac{\sum_{i=1}^n \hat{g}(Z_i)(Y_i T_i^j - \overline{YT}_n^j)}{\sum_{i=1}^n \hat{g}(Z_i)(T_i^j - \overline{T}_n^j)}, \quad \overline{YT}_n^j = \frac{1}{n}\sum_{i=1}^n Y_i T_i^j \quad \text{and} \quad \overline{T}_n^j = \frac{1}{n}\sum_{i=1}^n T_i^j.$$

Define $\gamma^j = \mathbb{E}[Y_i T_i^j] - \lambda^j \mathbb{E}[T_i^j]$ with $j = a, b$. Denote $\eta = (\theta', \gamma^a, \gamma^b, \lambda^a, \lambda^b)' \in \mathbb{R}^{d_\eta}$ and let ρ^0 be the true value of ρ. Let $\mathbf{0}_{k \times k'}$ be a $k \times k'$ matrix of zeros.

Theorem 3.2. *Under Assumptions 3.1 and 3.2, we have*

$$\sqrt{n}(\hat{\rho}_n - \rho^0) \xrightarrow{d} \mathcal{N}\left(0, \delta H^{-1}\Sigma H^{-1'}\delta'\right),$$

where denote $\delta = (\mathbf{0}_{1\times d_\theta}, 0, 0, 1, -1)$, $H = \mathbb{E}\left[\frac{\partial h(W_i;\eta^0)}{\partial \eta'}\right]$ *a* $d_\eta \times d_\eta$ *matrix,* $\Sigma = Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n h(W_i;\eta^0)\right]$, *and* $h(W_i;\eta)$ *is a* $d_\eta \times 1$ *vector defined in the proof of this theorem.*

Proof of Theorem 3.2. See Appendix A.2.3. □

A consistent estimator of the matrix H is $\hat{H}_n = \frac{1}{n}\sum_{i=1}^n \frac{\partial h(W_i;\hat{\eta}_n)}{\partial \eta'}$. Denote a $n \times d_\eta$ matrix $\mathbf{h}_n(\eta) = (h(W_1;\eta), ..., h(W_n;\eta))'$ and let G be an $n \times n$ matrix capturing the dependence structure of all the observations $\{W_i\}_{i=1}^n$. Then a consistent estimator of the matrix Σ can be expressed as

$$\hat{\Sigma}_n = \frac{1}{n}\mathbf{h}_n(\hat{\eta}_n)' G \mathbf{h}_n(\hat{\eta}_n). \tag{7}$$

For i.i.d. samples, it is clear that G is an identity matrix and $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n h(W_i;\hat{\eta}_n)h(W_i;\hat{\eta}_n)'$. By using the expression in (7), it is also straightforward to generalize the i.i.d. setting to allow for more general dependence structure of samples and to compute robust standard errors of $\hat{\rho}_n$. For example, suppose there are C clusters $g = 1, ..., C$, and samples in different clusters are independent but they are dependent of each other within clusters. Then, G can be set as a block-diagonal where the blocks are matrices of ones and the off-diagonal entries are zeros. By construction, $\hat{\Sigma}_n$ in (7) gives an consistent estimator for the clustered covariance matrix of $\frac{1}{\sqrt{n}}\sum_{i=1}^n h(W_i;\eta^0)$. To deal with more general data correlation and heteroskedasticity, we can estimate Σ by adopting Newey and West (1987) approach and the structure of G need to be adjusted accordingly.

3.5 Adding Covariates

Recall that the MR-LATE's estimand, ρ, is a difference between two IV estimands. Hence, if covariates affect the model linearly, ρ can be equivalently computed via including X as additional regressors in the IV regression. Much as in Hu (2008), once explanatory variables are taken into account, we allow the measurement error to be correlated with covariates. Formally, we require Assumption 3.3 to change as follows:

Assumption 3.5. *For* $j = a, b$ *and* $k = 1, 2, ..., K$, *we have*

(i) (Non-differential misclassification with covariates) $\mathbb{E}[Y|T^j, D, C_k, X] = \mathbb{E}[Y|D, C_k, X]$;

(ii) (Homogeneous misclassification with covariates) $\mathbb{E}\left[T^j|D, C_k, X\right] = \mathbb{E}[T^j|D, X]$.

Under Assumption 3.1, 3.2 and 3.5, all our main results still hold by conditioning on set of covariates X.

3.6 Discrete Treatment Variable

Thus far, we have considered a binary treatment framework. The key insight is that, when the true treatment is binary, we only need two binary proxies to compute ρ (one for $D = 1$ and the other for $D = 0$). In principle, the same idea can be applied to study the incremental returns to a discrete treatment variable (Angrist and Imbens, 1995). Nevertheless, the analysis needs to be adjusted, as it requires taking into account the potential misclassification in all the treatment levels. Based on the full set of results presented in Appendix A.3, we obtain that, when the true treatment is discrete, one would require one binary proxy for each realization of D. Hence, it is easy to recognize that, in a discrete treatment framework with misclassification, the sufficient conditions for point identification are much stricter than those required in the binary treatment case. This is why we prefer to leave it out of the main text.

4 Simulations and Applications

This section is organized in three parts. First, we illustrate our identification strategy and its finite sample performance by Monte Carlo simulations. Second, we provide some practical guidance on how to implement our method. Third, we reassess the returns to education in the U.K.

4.1 Monte Carlo Simulations

Consider the following data generating process (DGP):

$$\begin{aligned} Y_0 &= 0.5 + X + O + V_0, \\ Y_1 &= 1.5 + X + O + V_1, \\ Y &= DY_1 + (1-D)Y_0, \end{aligned}$$

where $X \sim \mathcal{N}(0, 0.5)$ is a random covariate, O is an unobservable omitted variable, and V_0 and V_1 are standard normal random errors. The true effect $\alpha^{IV} = 1$. The unobserved true treatment D is generated by

$$D = 1[\gamma_0 + \gamma_1 Z + \gamma_2 X + V_D \geq 0].$$

The error terms V_0 and V_1 are mutually independent and $(V_0, V_1)' \perp (O, V_D, Z)$. We set $\gamma_0 = -2$, $\gamma_1 = \{1, 1.5\}$ (instrument strength) and $\gamma_2 = 1$. The randomly generated discrete instrument Z takes values in a finite set $\Omega_Z = \{0, 1, 2\}$ with probabilities $\pi_0 = 0.4, \pi_1 = 0.4, \pi_2 = 0.2$, and V_D is the error term of D that is correlated with the omitted variable O:

$$\begin{bmatrix} O \\ V_D \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \right).$$

Suppose that (Y, Z, P, X) is observable, where $P \in \{1, 0, -1\}$ can be thought of as a self-reported treatment status and reveals partial information about the true treatment D. For example, $P=1$ if "treated", $P=0$ if "do not remember" or missing value, and $P=-1$ if "untreated". Generate

$$P = DP_1 - (1-D)P_0, \tag{8}$$

with $P_d \in \{-1,0,1\}$ the unobserved potential reporting quality associated with treatment status $D=d$. Then, P_0 and P_1 take value 1 for correct reporting, 0 for unclear reporting, and -1 for opposite reporting. Denote by $\Phi(\cdot)$ the CDF of the standard normal distribution. We generate P_1 and P_0 as follows:

$$P_1 = \begin{cases} 1, & \text{if } 1-p_1^a \le \Phi(U_1) \\ 0, & \text{if } p_1^b \le \Phi(U_1) < 1-p_1^a \\ -1, & \text{if } \Phi(U_1) < p_1^b \end{cases}, \text{ and } P_0 = \begin{cases} 1, & \text{if } 1-p_0^b \le \Phi(U_0) \\ 0, & \text{if } p_0^a \le \Phi(U_0) < 1-p_0^b \\ -1, & \text{if } \Phi(U_0) < p_0^a \end{cases},$$

with $p_0^a \le 1-p_0^b$, $p_1^b \le 1-p_1^a$, and unobservables (U_1,U_0) are jointly normal:

$$\begin{bmatrix} U_1 \\ U_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right).$$

The DGP for P_0 and P_1 indicate that the unclear or missing responses ($P=0$) are not at random, if $\Pr(P=0|D=1) \neq \Pr(P=0|D=0)$. Figure 1 provides a graphical illustration of the data generating process for P_1 and P_0. Based on Equation (8), $P=1$ either because $D=1$ and the individual correctly reports ($P_1=1$), or because $D=0$ and the individual misreports ($P_0=-1$). Furthermore, $P=0$ either because $D=1$ and the individual misreports being treated ($P_1=0$), or because $D=0$ and the individual misreports being untreated ($P_0=0$). Finally, $P=-1$ either because $D=1$ and the individual misreports her treatment status ($P_1=-1$), or because $D=0$ and the individual correctly reports her treatment status ($P_0=1$).

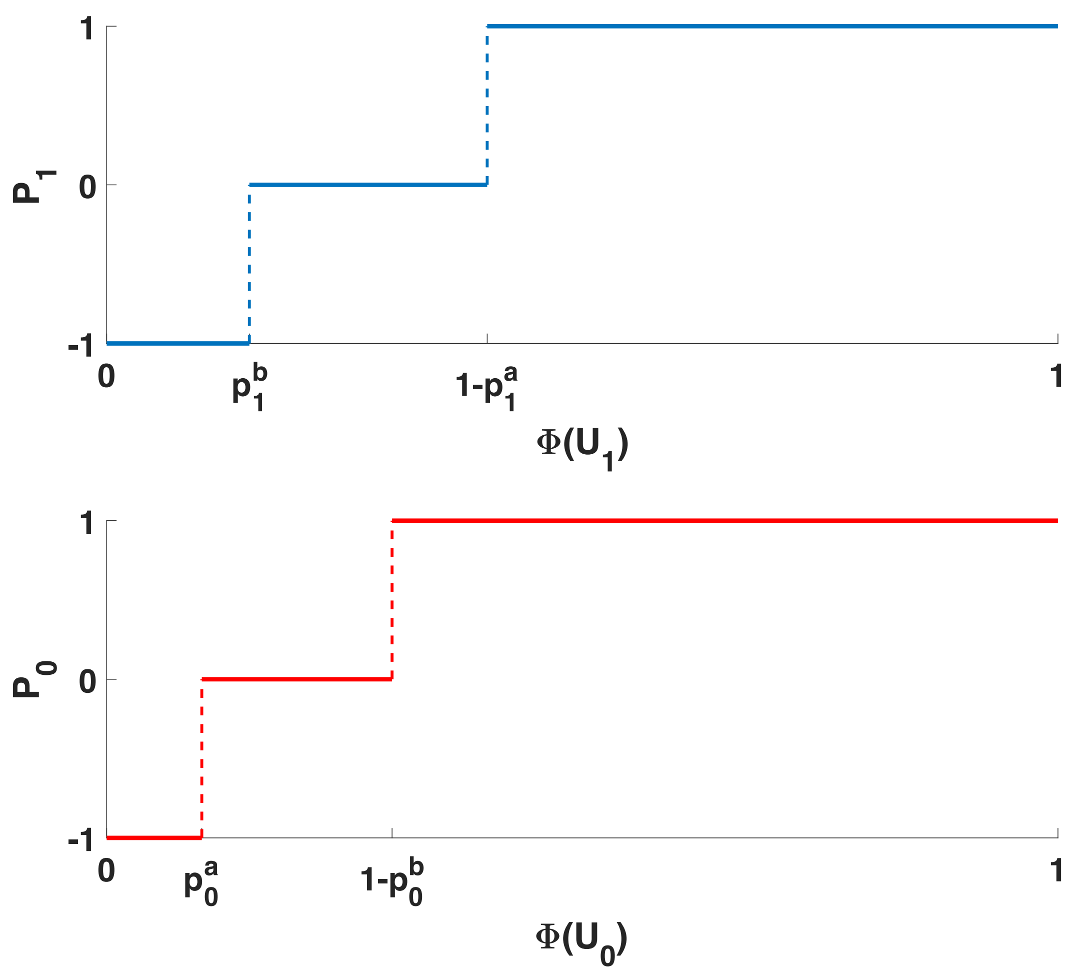
We generate two treatment proxies $T^a = 1[P=1]$ and $T^b = 1[P=-1]$, which can be generated equivalently as follows:

$$T^a = DT_1^a + (1-D)T_0^a, \text{ where } T_0^a = 1[\Phi(U_0) < p_0^a], \ \ T_1^a = 1[\Phi(U_1) \ge 1-p_1^a],$$
$$T^b = DT_1^b + (1-D)T_0^b, \text{ where } T_0^b = 1[\Phi(U_0) \ge 1-p_0^b], \ \ T_1^b = 1[\Phi(U_1) < p_1^b].$$

We set $p_1^a = 0.7$, $p_0^b = 0.9$, which mimic the situation where treated individuals are more likely to be unclear about and misreport their treatment status compared to the untreated.[4] We generate random samples of size $n = \{500, \ 1,000, \ 2,000\}$ and replications $M = 5,000$ times. We compare the performance of several methods using ordinary least squares (OLS) and two-stage least squares (2SLS): (1) Infeasible OLS, which is the OLS of Y on the true treatment D; (2) Infeasible 2SLS, which is the 2SLS of Y on D using Z as an instrument; (3) Feasible OLS, which is the OLS of

[4]The choice of p_1^a and p_0^b aims to be realistic. The qualitative results of the analysis do not change for different choices of these parameters. Specifically, in this example, 70% of true treated respond to be treated and 90% of true untreated respond to be untreated.

Figure 1: Graphic Illustration DGP for P_0 and P_1



Notes: P_1 and P_0 stand for unobserved potential reporting quality of true treatment status and true control status, respectively. As the figure shows, true treated individuals correctly report to be treated ($P_1 = 1$) if $1 - p_1^a \leq \Phi(U_1)$; they incorrectly report to be untreated ($P_1 = -1$) if $\Phi(U_1) < p_1^b$; and they report an ambiguous treatment status ($P_1 = 0$) if $p_1^b \leq \Phi(U_1) < 1 - p_1^a$. The same logic applies for P_0.

Y on the observable proxy T^a; (4) Feasible 2SLS (replace), which is the 2SLS of Y on T^a using Z as an instrument and replacing unclear or missing treatment observations to zero; (5) Feasible 2SLS (drop), which is the 2SLS of Y on T^a using Z as an instrument by dropping samples with $T^a = T^b = 0$ (or equivalently dropping samples with $P = 0$); and (6) $\rho = \lambda^a - \lambda^b$, which is the MR-LATE approach.

Let us consider three DGP designs for p_0^a, p_1^b and (U_1, U_0, O). In the first DGP design, we set $p_0^a = 0$ and $p_1^b = 0$ to ensure that there is only one type of misclassification error. In addition, we generate random (U_1, U_0) independent of all other variables, indicating non-differential and homogeneous misreporting. In this case, samples with $P = 0$ are not missing at random because $\Pr(P = 0|D = 1) = 1 - p_1^a - p_1^b$ is not the same as $\Pr(P = 0|D = 0) = 1 - p_0^a - p_0^b$. Under this scenario, $\rho = \lambda^a - \lambda^b$ point identifies α^{IV}.

Panel (a) of Table 1 reports the simulation results for the scenario with $\gamma_1 = 1$ when all the sufficient conditions required by MR-LATE are satisfied (first design). As one can see, MR-LATE outperforms the other three feasible methods, as the bias and standard deviation of ρ are much closer to those of the infeasible 2SLS. Panel (b) reports the same set of results using a stronger instrumental variable ($\gamma_1 = 1.5$). Comparing the results between Panel (a) and (b), one can see that, as expected, a stronger instrument improves further the finite sample performance of MR-LATE. For the Feasible 2SLS (drop) method, its bias is larger than that of our MR-LATE method, and

Table 1: Binary Treatment with Non-differential Misclassification (True $\alpha^{IV} = 1$)

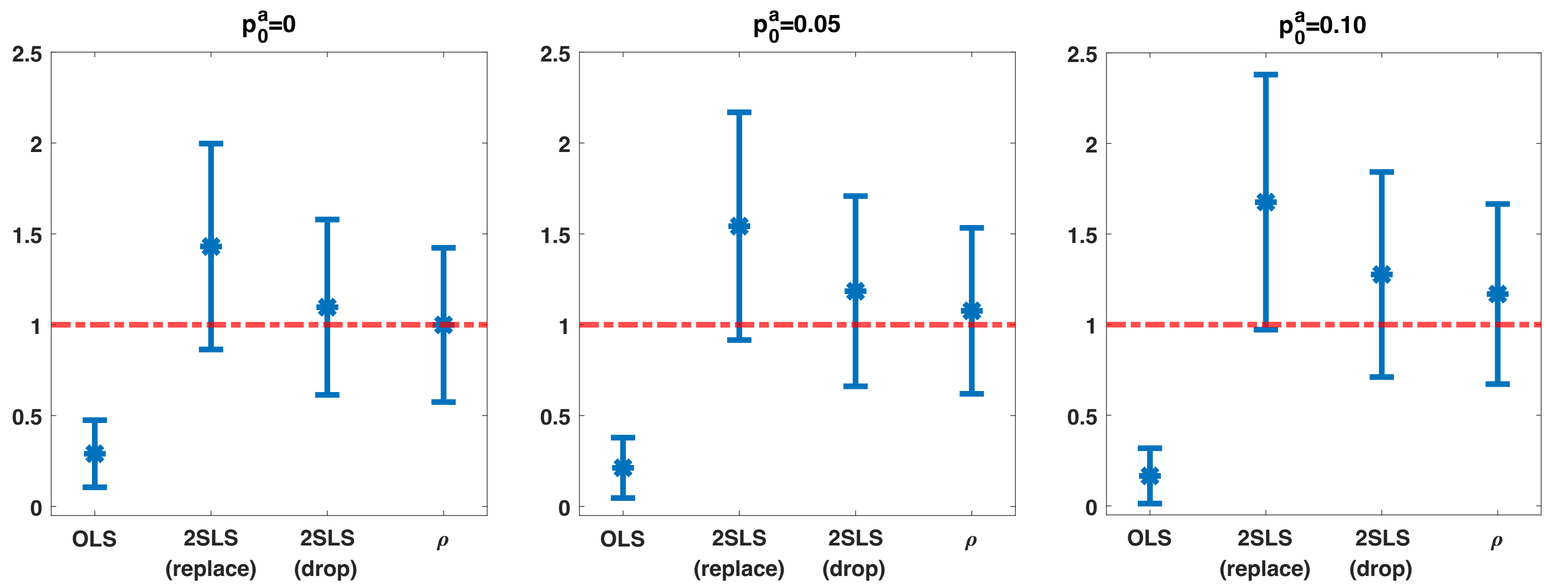
		Infeasible		Feasible			
		(1)	(2)	(3)	(4)	(5)	(5)
		OLS	2SLS	OLS	2SLS (replace)	2SLS (drop)	$\rho = \lambda^a - \lambda^b$ (MR-LATE)
Panel (a): $\gamma_1 = 1$							
n=500	Mean	0.319	1.010	0.292	1.457	1.118	1.016
	Bias	-0.681	0.010	-0.708	0.457	0.118	0.016
	S.D.	0.164	0.391	0.190	0.582	0.493	0.435
	MSE	0.491	0.153	0.538	0.547	0.257	0.190
n=1,000	Mean	0.319	1.003	0.292	1.438	1.103	1.005
	Bias	-0.681	0.003	-0.708	0.438	0.103	0.005
	S.D.	0.115	0.274	0.132	0.407	0.346	0.306
	MSE	0.477	0.075	0.519	0.357	0.130	0.094
n=2,000	Mean	0.320	1.002	0.292	1.436	1.101	1.003
	Bias	-0.680	0.002	-0.708	0.436	0.101	0.003
	S.D.	0.081	0.194	0.092	0.287	0.242	0.215
	MSE	0.469	0.038	0.510	0.273	0.069	0.046
Panel (b): $\gamma_1 = 1.5$							
n=500	Mean	0.494	1.007	0.429	1.444	1.054	1.008
	Bias	-0.506	0.007	-0.571	0.444	0.054	0.008
	S.D.	0.140	0.227	0.158	0.343	0.273	0.259
	MSE	0.276	0.052	0.351	0.315	0.077	0.067
n=1000	Mean	0.491	1.000	0.428	1.431	1.048	1.003
	Bias	-0.509	0.000	-0.572	0.431	0.048	0.003
	S.D.	0.098	0.158	0.108	0.236	0.189	0.179
	MSE	0.268	0.025	0.339	0.241	0.038	0.032
n=2000	Mean	0.493	1.000	0.425	1.430	1.044	0.999
	Bias	-0.507	0.000	-0.575	0.430	0.044	-0.001
	S.D.	0.070	0.114	0.078	0.170	0.136	0.130
	MSE	0.262	0.013	0.336	0.214	0.021	0.017

Notes: The table reports the simulation results when all the sufficient conditions required by MR-LATE are satisfied (first design). In each simulation, the true value $\alpha^{IV} = 1$. Results are based on 5,000 replications. In Panel (a), $\gamma_1 = 1$. In Panel (b), the instrument is stronger as $\gamma_1 = 1.5$. We compare the performance of several methods: (1) Infeasible OLS, which is the OLS of Y on the true treatment D; (2) Infeasible 2SLS, which is the 2SLS of Y on D using Z as an instrument; (3) Feasible OLS, which is the OLS of Y on the observable proxy T^a; (4) Feasible 2SLS (replace), which is the 2SLS of Y on T^a using Z as an instrument and replacing unclear or missing treatment observations to zero; (5) Feasible 2SLS (drop), which is the 2SLS of Y on T^a using Z as an instrument by dropping samples with $T^a = T^b = 0$ (or equivalently dropping samples with $P = 0$); and (6) $\rho = \lambda^a - \lambda^b$, which is the MR-LATE approach.

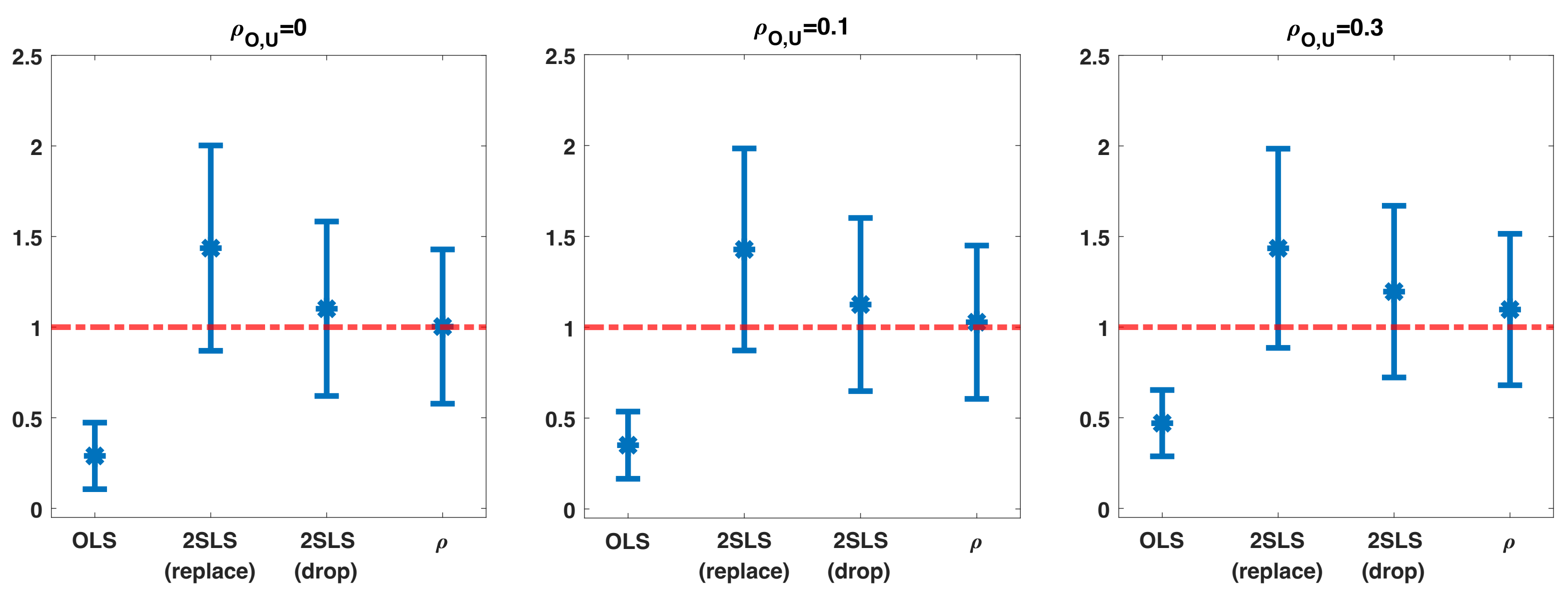
the decrease in the bias is negligible as the sample size increases. This is because the unclear or missing treatment observations are not missing at random. Therefore, our method is preferable to the Feasible 2SLS (drop) in this case.

Figure 2: Feasible Estimators under the Violation of Assumptions ($\gamma_1 = 1, n = 2,000$)

(a) Violation of No False Positive



(b) Violation of Non-differential and Homogeneous Misclassification

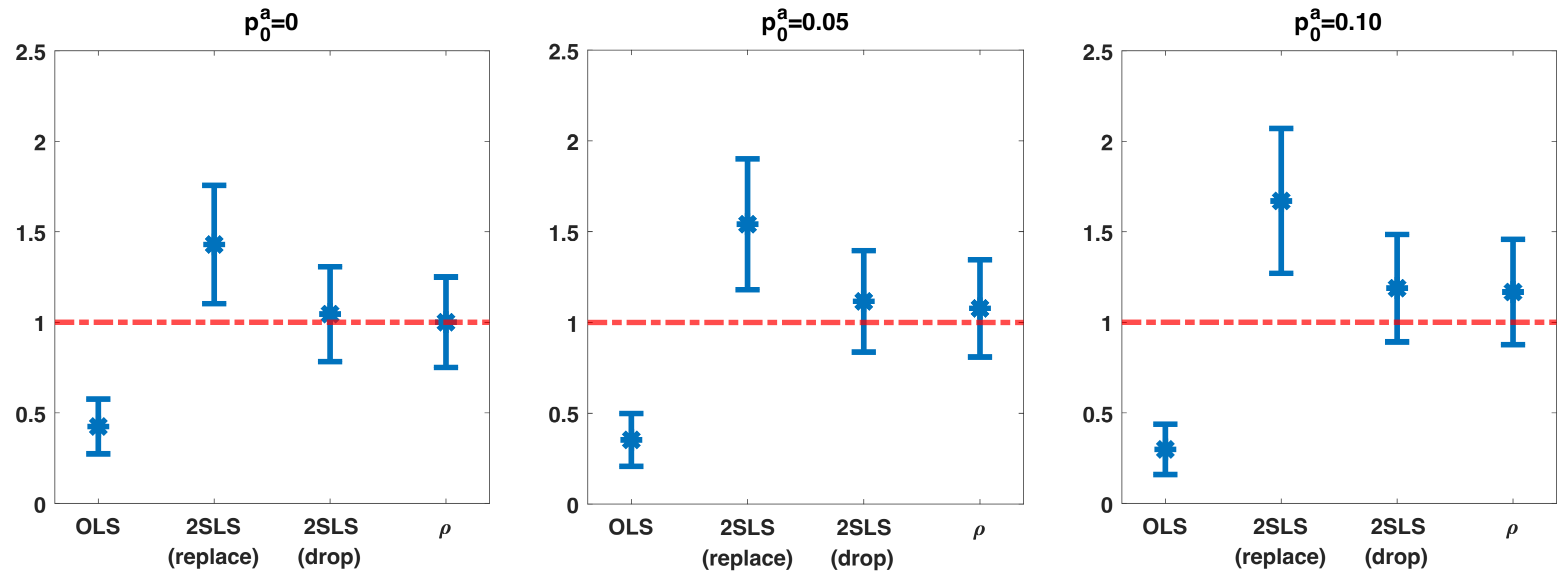


Notes: This figure plots the results of the feasible estimators for α^{IV} under instrument strength $\gamma_1 = 1$ and sample size $n = 2,000$: OLS, 2SLS (replace), 2SLS (drop), and $\rho = \lambda^a - \lambda^b$. The true value of $\alpha^{IV} = 1$ is the red dash-dotted lines, and the average 95% confidence intervals of each method are in blue line with * as the mean of the estimates over 5,000 simulations. Plots in panel (a) are under the violation of the no false positive assumption ($p_0^a \geq 0$), while keeping $p_1^b = 0$ and the non-differential and homogeneous misclassification assumptions. Plots in panel (b) are under the violation of both non-differential and homogeneous misclassification assumptions ($\rho_{O,U} \geq 0$ which is the correlation between (O, U_1) and (O, U_0)), while keeping $p_0^a = p_1^b = 0$.

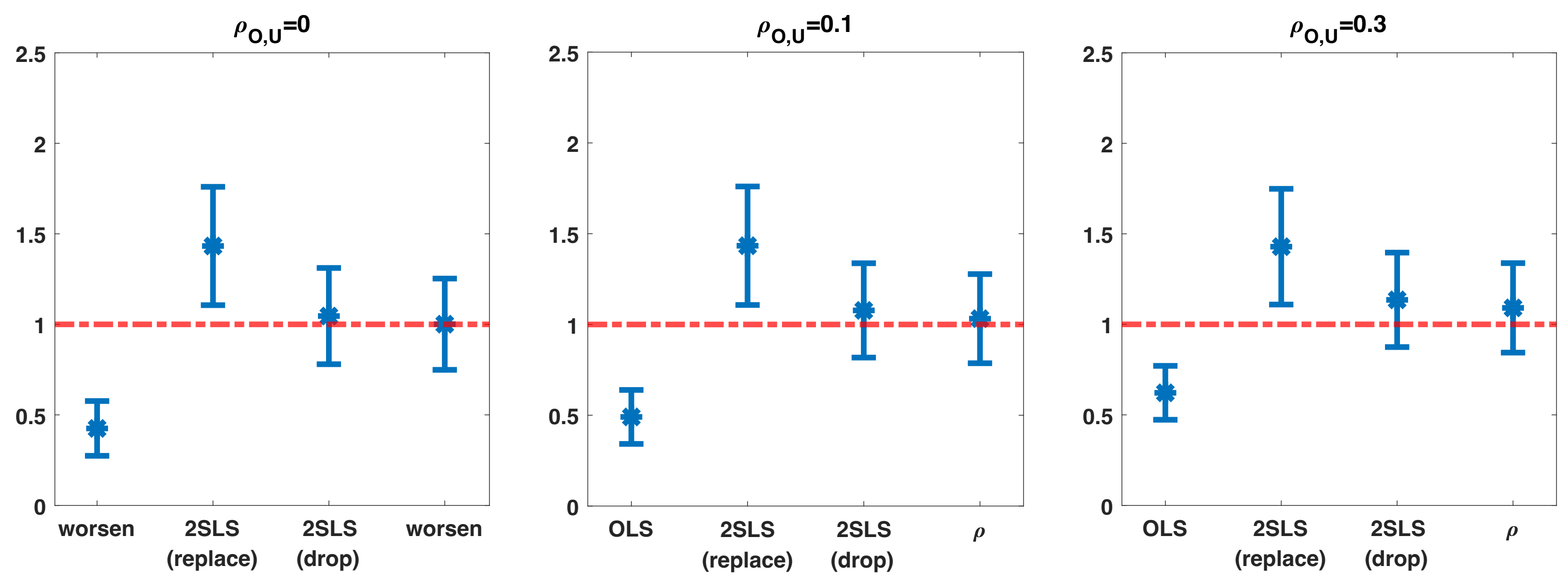
In the next two DGP designs, we wish to check the robustness of our point identification strategy. Hence, we study the performance of the four feasible methods when some of the sufficient conditions for the point identification in Section 3 are violated. First, as aforementioned, the condition $p_0^a = p_1^b = 0$ is quite restrictive and hard to verify in practice. Hence, we proceed by allowing $p_0^a \in \{0.05, 0.1\}$, that is, we relax the no false positive assumption, while keeping $p_1^b = 0$ and the independence $(U_1, U_0)' \perp O$ (second design). Second, we maintain $p_0^a = p_1^b = 0$, but relax both the non-differential and homogeneous misclassification assumptions by generating $(U_1, U_0, O)'$ jointly from a normal distribution with $\rho_{O,U}$=Corr(O, U_1) =Corr$(O, U_0) \in \{0.1, 0.3\}$ (third design). As before, we compare the results for different values of γ_1.

Figure 3: Feasible Estimators under Violation of Assumptions ($\gamma_1 = 1.5, n = 2,000$)

(a) Violation of No False Positive



(b) Violation of Non-differential and Homogeneous Misclassification



Notes: This figure plots the results of the feasible estimators for α^{IV} under instrument strength $\gamma_1 = 1.5$ and sample size $n = 2,000$: OLS, 2SLS (replace), 2SLS (drop), and $\rho = \lambda^a - \lambda^b$. The true value of $\alpha^{IV} = 1$ is the red dash-dotted lines, and the average 95% confidence intervals of each method are in blue line with * as the mean of the estimates over 5,000 simulations. Plots in panel (a) are under the violation of the no false positive assumption ($p_0^a \geq 0$), while keeping $p_1^b = 0$ and the non-differential and homogeneous misclassification assumptions. Plots in panel (b) are under the violation of both non-differential and homogeneous misclassification assumptions ($\rho_{O,U} \geq 0$ which is the correlation between (O, U_1) and (O, U_0)), while keeping $p_0^a = p_1^b = 0$.

Figure 2 displays the results of these two robustness checks. Panel (a) plots the estimation results of the four feasible methods under the violation of the "no false positive" assumption. We can see that the bias (in absolute value) and the standard deviation of ρ are less than those of other feasible methods. In addition, the 95% confidence intervals of ρ contain the true value $\alpha^{IV} = 1$ when the "no false positive" assumption is violated. Moreover, results in Panel (b) demonstrate that the performance of ρ, as a bias reduction method, is also quite robust to the violation of the assumptions of non-differential and homogeneous misclassification error. The estimates of ρ give smaller bias and standard deviation compared to other feasible methods in all the settings.

Figure 3 reports the same results using a stronger instrument ($\gamma_1 = 1.5$). As one can see, MR-LATE with stronger instrument(s) is more robust to the failure of the sufficient conditions for point identification of α^{IV}.

4.2 Practical Guidance

The MR-LATE estimator can be applied to a variety of contexts. Consider estimating the benefits of attaining any academic qualification, compared to leaving school without any formal qualifications, using some available IVs. The actual treatment D takes value one if an individual completes the academic program and zero otherwise. If D was correctly observed, under the conditions listed in Assumption 3.1, the Imbens and Angrist (1994)'s weighted average of local average treatment effects would be identified by the standard IV estimand α^{IV}. However, due to the growing concern regarding the quality of self-reported data, the researcher may suspect that the treatment D is measured with error and the standard IV approach is not appropriate in this case. In order to show how to implement MR-LATE, we consider four examples and illustrate how one should construct the proxies T^a and T^b in each scenario.

Example 1. The first example mimics a context where, instead of D, we can observe one answer P to a survey question, where P takes three values $\{1, 0, -1\}$ (i.e. treated, unclear or missing value, untreated) and reveals some information about the true treatment D. This is the same situation described in Section 2. In this context, one can define $T^a = 1$ if $P = 1$ ($T^a = 0$ otherwise), and $T^b = 1$ if $P = -1$ ($T^b = 0$ otherwise). Given T^a and T^b, if observations having $P = 1$ and $P = -1$ are not misclassified, the MR-LATE will correctly identify and consistently estimate the true parameter. Alternatively, if some of the observations with $P = 1$ or $P = -1$ are also misclassified, MR-LATE is biased but it can still provide a significantly less biased estimate of the target parameter relative to the standard IV estimation.

Example 2. In the second example, consider a context where the practitioner has, for each individual, two binary survey questions, $P^j \in \{0, 1\}$ with $j = \{1, 2\}$, of the same unobserved true treatment status D. Using the example of a study on returns to education, these could be two separate questions in the same questionnaire, or two different sources of information, such as self-reported education levels and transcript records from the schools. For example: a direct question to each individual, such as "Did you finish your O level in high school?", where the answers are yes ($P^1 = 1$) or no ($P^1 = 0$); and, at the same time, a direct question to the school, such as "Did the student complete her (his) O level in high school?", where the answers are again yes ($P^2 = 1$) or no ($P^2 = 0$). In this context, one can define $T^a = 1$ if $P^1 = P^2 = 1$ ($T^a = 0$ otherwise), and $T^b = 1$ if $P_1 = P_2 = 0$ ($T^b = 0$ otherwise). As in Example 1, if observations having $P_1 = P_2 = 1$ are all actually treated and $P_1 = P_2 = 0$ are all actually not treated, the MR-LATE will correctly identify and consistently estimate the true parameter. Alternatively, if some of the observations are also misclassified, the estimator will provide a better approximation of the target parameter.

Example 3. The third example mimics a context where the practitioner has, for each individual, two repeated measurements of qualification that are self-reported by the individual at different time points, denoted by P^j with $j = \{1, 2\}$ (the superscript j now would index time). Then, both P^1 and P^2 are proxies for the unobserved treatment variable D . For example, in different years, we ask the same individuals: "Did you finish your O level in high school?", where the answers are again yes ($P^j = 1$) or no ($P^j = 0$). The practitioner can define T^a and T^b in the same way as in Example 2.

Example 4. In the last example, we discuss a situation where the practitioner has, for each individual, more than two measures of the same unobserved treatment variable D based on different sources, or more than two repeated answers to the same survey question. This example would probably fit better a common health-related application with multiple 0-1 survey responses, indicating the absence or presence of a condition. In this case, the practitioner can define $T^a = 1$ if $\sum_{j=1}^{m} P^j = m$ and $T^b = 1$ if $\sum_{j=1}^{m} P^j = 0$ with m the number of survey questions.

4.3 Returns to Education in the UK

The National Child Development Survey (NCDS) is particularly suited to illustrate the use of MR-LATE. Researchers at the Centre for Longitudinal Studies at the University College London have been following the lives of more than 17,000 people born in England, Scotland, or Wales since they were born in a single week of March 1958. We use a version of the dataset constructed by Battistin and Sianesi (2011) and Battistin et al. (2014). Our main outcome variable is real gross hourly wages at age 33. To avoid issues of selection into employment, we restrict our sample to men employed in the formal workforce in 1991.

Binary Treatment. Application of MR-LATE to study the returns to education requires either multiple measurements of educational attainment (as in Example 2, 3 or 4 in Section 4.2) or one measurement of educational attainment that can have three values (as in Example 1 in Section 4.2). The NCDS data has the former: self-reported attainment collected in 1981, when individuals were 23, and in 1991; and, for academic O-level qualifications achieved by age 20, a report in the official files of the schools they attended when aged 16. Thus, we construct $T^a = 1$ if both the self-reported measure in 1981 and school reports agree that the individual did obtain an O-level qualification, and 0 otherwise, and $T^b = 1$ if both sources agree that they did not obtain the qualification, and 0 otherwise.

Discrete Instrument. NCDS included a combined, dichotomous measure of parental interest in participants' education based on the assessment of the child's teacher when they were 7 along with detailed family variables and the type of school the child was attending at the age of 16 (Blundell et al., 2005). Parental interest was equal to 0 if the teacher judged both parents as having "some" or "little" interest in their child's education and 1 if, according to the child's teacher, the parents were "very" or "overly" interested. We take this measure and combine it with a common proxy of mother's bargaining power in the family. Power equals 0 if the mother is equally or less educated than the father and 1 if she has more education. The interaction of parental interest with mother's bargaining power according to the difference in education between parents yields a discrete instrument and will illustrate the performance of our estimator. If parental interest and power are both 1, $Z = 2$. If parental interest is 1 and power is 0, or parental interest is 0 and power is 1, $Z = 1$. If power is 0 and parental interest is 0, the discrete instrument takes the value $Z = 0$. Appendix A.5 provides detailed summary statistics for the variables we employ.

Sensitivity of program benefits. Before proceeding to the main results, we show how the MR-LATE approach can be used by practitioners to assess the sensitivity of program benefits based on different hypothetical values or external information of the extent of misclassification. Given the available information regarding treatment misclassification probabilities, a researcher can use Equation (6) to approximate the possible level of biases of the benefits of O-level qualifications. In our setting, Battistin et al. (2014) estimate that the extent of correct classification in the education attainment in this sample is between 82-86.8% (Table 3, p.144). Recall, $\alpha^{IV} = \xi \alpha^{Mis}$ and $\xi = 1 - w^p - w^n$, where w^p is the average percentage of false positive and w^n is the average percentage of false negative. We use this information to set $w^n + w^p \approx$ 13.2-18%, which means that, without accounting for treatment misclassification, the estimated treatment effect would be likely biased (upward) by approximately $\frac{w^n + w^p}{1 - w^n - w^p} =$ 15-22%.

Results. For illustration, we use the school reports and the self-reports from 1981. Our estimates of the average wage return to any academic qualification are derived using five different methods and two sets of control variables. They are shown in Table 2. Column (1) reports the OLS estimates using only self-reported qualification as the treatment variable. Here we ignore the problems of both endogeneity and misclassification. Column (2) provides the naïve 2SLS estimates using the same treatment variable. Here we exploit the discrete instrument defined earlier. Results indicate that our discrete instrument is an important determinant of acquiring academic qualifications, even conditional on a rich set of observables (the first-stage F-test statistic is 55). Obtaining an O-level qualification increases wages by roughly 54%, a statistically significant effect. Once we include the full set of covariates, the effect is 44%. While the estimator accounts for treatment effect endogeneity, it ignores the potential misclassification and thus it is biased.

Columns (3) and (4) display the average wage return using the other two (replace and drop) 2SLS approaches described in Section 2. First, we replace with zero the treatment status of all the observations where the measurements of the academic qualification are discordant, which corresponds to the 2SLS (replace) in Monte Carlo simulations. Column (3) reports the 2SLS estimates. Second, we drop all the observations where the measurements of the academic qualification are discordant, which corresponds to the 2SLS (drop) in Monte Carlo simulations. Column (4) reports the 2SLS estimates. Neither result exceeds the naïve 2SLS estimates in column (2). Column (5) reports our MR-LATE estimates using the approach described in Section 3. Since both the school and the self-reports contain measurement errors (Battistin et al., 2014), we cannot rule out the pos-

Table 2: Empirical illustration

	(1) OLS	(2) 2SLS (naïve)	(3) 2SLS (replace)	(4) 2SLS (drop)	(5) MR-LATE
Panel (a): No covariates					
Treatment [0,1]	0.328***	0.541***	0.527***	0.496***	0.487***
	(0.016)	(0.062)	(0.060)	(0.059)	(0.059)
Observations	2,454	2,454	2,454	2,218	2,454
R-squared	0.186	0.127	0.148	0.187	
Controls	No	No	No	No	No
Panel (b): With covariates					
Treatment [0,1]	0.273***	0.437***	0.440***	0.379***	0.369***
	(0.016)	(0.113)	(0.112)	(0.105)	(0.108)
Observations	2,454	2,454	2,454	2,218	2,454
R-squared	0.224	0.194	0.204	0.238	
Controls	Yes	Yes	Yes	Yes	Yes

Notes: Dependent variable: Log of wage in 1991. In each specification in Panel (a), we control only for ethnicity and region. Whereas, in Panel (b) we control also for detailed family background variables when the child was 16 and school type variables. We report only the coefficients on the treatment. Robust standard errors. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

sibility that T^a and T^b are still misclassified even for individuals with concordant reports from the school and themselves. Thus, the one type misclassification Assumption 3.3 may not apply. Rather, this empirical illustration is an example of Case 3: generally mismeasured treatment. This means that, while the point estimate is likely biased, it is closer to the true effect than the values reported in Columns (2), (3), and (4).

5 Conclusion

This paper develops an instrumental variable approach to identify and estimate the weighted average of local average treatment effects (LATE) of Imbens and Angrist (1994) in a context of endogenous and misclassified treatment. We focus on cases of non-differential misclassification, such as recording mistakes, imperfect compliance, poor recalling, or incomplete awareness. Since the misclassifications are nonclassical, standard instrumental variable techniques are not able to eliminate the bias.

This paper has three main results. First, we provide sufficient conditions under which the point identification the effect of the treatment can be achieved. Importantly, we generalize the MR-LATE approach for binary instrument and binary treatment proposed by Calvi, Lewbel, and Tommasi (2021) to incorporate discrete instrument(s) and discrete treatment. Second, we establish the asymptotic properties of the proposed estimator to infer the parameter of interest. Third, we provide Monte Carlo simulation studies to explore the finite sample performance of our estimation method and to illustrate its practical usefulness. Our proposed method can be applied in any settings where the accuracy of observable treatment measurement(s) is questionable and a treatment indicator for each latent treatment level is available. It can be applied as either the leading identification strategy or the leading robustness check.

References

ABADIE, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263. [5], [9], [27], [30]

ACERENZA, S., K. BAN, AND D. KÉDAGNI (2021): "Marginal Treatment Effects with Misclassified Treatment," Tech. rep. [3]

AIGNER, D. J. (1973): "Regression with a binary independent variable subject to errors of observation," *Journal of Econometrics*, 1, 49 – 59. [3]

ANGRIST, J. D. AND G. W. IMBENS (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 90, 431–442. [3], [14], [35], [37], [41], [42]

ANGRIST, J. D. AND A. B. KRUEGER (1999): "Chapter 23 - Empirical Strategies in Labor Economics," Elsevier, vol. 3, Part A of *Handbook of Labor Economics*, 1277 – 1366. [3]

ATHEY, S. AND G. IMBENS (2017): "Chapter 3 - The econometrics of randomized experiments," in *Handbook of Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Economic Field Experiments*, 73 – 140. [2]

BATTISTIN, E., M. D. NADAI, AND B. SIANESI (2014): "Misreported schooling, multiple measures and returns to educational qualifications," *Journal of Econometrics*, 181, 136 – 150. [2], [21], [22], [47]

BATTISTIN, E. AND B. SIANESI (2011): "Misclassified treatment status and treatment effects: An application to returns to education in the United Kingdom," *Review of Economics and Statistics*, 93, 495–509. [9], [21], [47]

BLACK, D., S. SANDERS, AND L. TAYLOR (2003): "Measurement of higher education in the census and current population survey," *Journal of the American Statistical Association*, 98, 545–554. [3]

BLACK, D. A., M. C. BERGER, AND F. A. SCOTT (2000): "Bounding parameter estimates with nonclassical measurement error," *Journal of the American Statistical Association*, 95, 739–748. [3]

BLUNDELL, R., L. DEARDEN, AND B. SIANESI (2005): "Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 473–512. [21]

BOLLINGER, C. R. (1996): "Bounding mean regressions when a binary regressor is mismeasured," *Journal of Econometrics*, 73, 387 – 399. [3]

BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): "Measurement error in survey data," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 5, chap. 59, 3705–3843, 1 ed. [3]

CALVI, R., A. LEWBEL, AND D. TOMMASI (2021): "LATE With Missing or Mismeasured Treatment," *Journal of Business & Economic Statistics*, 0, 1–17. [2], [4], [9], [10], [12], [23] CARD, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. [3]

CHEN, X., C. A. FLORES, AND A. FLORES-LAGUNES (2018): "Going beyond LATE Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*, 53, 1050–1099. [39]

DITRAGLIA, F. J. AND C. GARCÍA-JIMENO (2019): "Identifying the effect of a mis-classified, binary, endogenous regressor," *Journal of Econometrics*, 209, 376–390. [2]

HECKMAN, J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. [3]

HERNANDEZ, M., S. PUDNEY, AND R. HANCOCK (2007): "The welfare cost of means-testing: pensioner participation in income support," *Journal of Applied Econometrics*, 22, 581–598. [3]

HU, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27 – 61. [3], [9], [13]

IMBENS, G. W. (2014): "Instrumental Variables: An Econometrician's Perspective," *Statist. Sci.*, 29, 323–358. [3]

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [2], [4], [6], [20], [23], [30]

JIANG, Z. AND P. DING (2020): "Measurement errors in the binary instrumental variable model," *Biometrika*, 107, 238–245. [2]

KANE, T. J., C. E. ROUSE, AND D. STAIGER (1999): "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper 7235. [3]

KASAHARA, H. AND K. SHIMOTSU (2021): "Identification of Regression Models with a Misclassified and Endogenous Binary Regressor," *Econometric Theory*, 1–23. [2]

KLEPPER, S. (1988): "Bounding the effects of measurement error in regressions involving dichotomous variables," *Journal of Econometrics*, 37, 343 – 359. [3]

KREIDER, B. (2010): "Regression coefficient identification decay in the presence of infrequent classification errors," *The Review of Economics and Statistics*, 92, 1017–1023. [2]

KREIDER, B., J. V. PEPPER, C. GUNDERSEN, AND D. JOLLIFFE (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, 958–975. [2]

LEWBEL, A. (2007): "Estimation of average treatment effects with misclassification," *Econometrica*, 75, 537–551. [3], [9]

MAHAJAN, A. (2006): "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631–665. [3]

MEYER, B. D., W. K. C. MOK, AND J. X. SULLIVAN (2015): "Household Surveys in Crisis," *Journal of Economic Perspectives*, 29, 199–226. [2]

MILLIMET, D. (2011): "The elephant in the corner: a cautionary tale about measurement error in treatment effects models," in *Missing Data Methods: Cross-Sectional Methods and Applications. In: Advances in Econometrics*, Emerald Group Publishing Limited, vol. 27, 1–39, 1 ed. [2]

MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. [3]

MOGSTAD, M., A. TORGOVITSKY, AND C. R. WALTERS (2020a): "Policy evaluation with multiple instrumental variables," Tech. rep., National Bureau of Economic Research. [7]

——— (2020b): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*. [2], [7]

MOLINARI, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144, 81 – 117. [3]

——— (2010): "Missing Treatments," *Journal of Business & Economic Statistics*, 28, 82–95. [3]

NEWEY, W. K. AND K. D. WEST (1987): "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [13], [40]

NGUIMKEU, P., A. DENTEH, AND R. TCHERNIS (2018): "On the estimation of treatment effects with endogenous misreporting," *Journal of Econometrics*. [2]

POSSEBOM, V. (2021): "Crime and Mismeasured Punishment: Marginal Treatment Effect with Misclassification," . [3]

TOMMASI, D. (2019): "Control of resources, bargaining power and the demand of food: Evidence from PROGRESA," *Journal of Economic Behavior & Organization*, 161, 265 – 286. [4]

TOMMASI, D. AND L. ZHANG (2020): "Bounding Program Benefits When Participation Is Misreported," *IZA Discussion Paper*. [2], [8], [36]

URA, T. (2018): "Heterogeneous treatment effects with mismeasured endogenous treatment," *Quantitative Economics*, 9, 1335–1370. [2]

YANAGI, T. (2018): "Inference on local average treatment effects for misclassified treatment," *Econometric Reviews*, 0, 1–23. [2]

A Appendix

This Appendix is organized as follows. Appendix A.1 shows the connection between MR-LATE and LARF (the local average response function) introduced by Abadie (2003). Appendix A.2 contains the proofs of Section 3. Appendix A.3 gives identification and estimation results for the discrete treatment setting. Appendix A.4 contains all the proofs of Section A.3. Appendix A.5 present additional tables from the empirical illustration.

Contents

A.1 MR-LATE and the Local Average Response Function 30
A.2 Proofs of Section 3 32
A.2.1 Proof of Lemma 3.1 32
A.2.2 Proof of Theorem 3.1 32
A.2.3 Proof of Theorem 3.2 34
A.3 MR-LATE with a Discrete Treatment Variable: Details 35
A.3.1 Point Identification with a Binary IV 37
A.3.2 Set Identification with a Binary IV 39
A.3.3 Set Identification in the General Case with a Discrete IV 39
A.3.4 Inference 39
A.4 Proofs of Section A.3 41
A.4.1 Proof of Theorem A.1 41
A.4.2 Proof of Corollary A.1 41
A.4.3 Proof of Theorem A.2 42
A.4.4 Proof of Lemma A.1 43
A.4.5 Proof of Theorem A.3 43
A.4.6 Proof of Corollary A.2 44
A.4.7 Proof of Lemma A.2 45
A.4.8 Proof of Corollary A.3 45
A.4.9 Proof of Theorem A.5 46
A.5 Returns to Education in the UK: Details 47

Table A1: Review of Articles

Authors	Year	Journal	Topic	Data	Year	Type of treatment	Misclass. rate	False positive (%)	False negative (%)
Almada, McCarthy, Tchernis	2016	American Journal of Agricultural Economics	SNAP	National Longitudinal Survey of Youth 1979 Cohort (NLSY79)	1996-2004	Binary	33	4	29
Acerenza, Ban, Kedagni	2021	Working paper	Education	Indonesia Family Life Survey (IFLS)	2000	Binary	16	-	-
Baker, Stabile, Deri	2004	Journal of Human Resources	Health	Canadian National Population Health Survey and Ontario Health Insurance Plan	1996-1997	Binary Cancer Diabetes Migraines Stroke Asthma	 74.5 36.7 55 49.9 48.7	 0.5 0.7 7 0.9 4.7	 74 36 48 49 44
Battistin, De Nadai, Sianesi	2014	Journal of Econometrics	Education	National Child Development Survey (NCDS)	1981 & 1991	Binary	37	26	11
Black, Berger, Scott	2000	Journal of the American Statistical Association	Health Insurance	Upjohn Institute Survey	1993	Binary	20.9	15.4	5.5
Black, Sanders, Taylor	2003	Journal of the American Statistical Association	Education	post-1991 Current Population Survey (CPS)		Binary	9.55	-	-
Brachet	2008	Working paper	Maternal smoking	US Natality	1989-1996	Binary	16.6-35.0	0	16.6-35.0
Card	1996	Econometrica	Union	Current Population Surveys (CPS)	1987 & 1988	Binary	5.2	2.5	2.7
Card, Hildreth, Shore-Sheppard	2004	Journal of Business & Economic Statistics	Medicaid	Survey of Income and Program Participation (SIPP) and California's Medi-Cal Eligibility File	1990-1993	Binary	16 to 17.5	1.3 to 2.8	14.7
Courtemanche, Denteh, Tchernis	2019	Southern Economic Journal	Food Security	FoodAPS-ADMIN FoodAPS-ALERT	2012-2013	Binary	20.1 19.3	8.4 7.8	11.7 11.5
Dustmann and van Soest	2001	Review of Economics and Statistics	Language	German SocioEconomic Panel (GSOEP)	1984–1987, 1989, 1991, and 1993	Discrete	82.3	70	12.3
Hu, Xiao, Zhong	2012	Working paper	Education	National Longitudinal Study of High School of 1972 (NLS-72) Self-reported Education	1972-1986	Discrete High School Some College Bachelor's Degree	 7.5 9.8 0.4	 7.5 2.8 0	 0 7 0.4
Johnston, Propper, Shields	2009	Journal of Health Economics	Health	Health Survey for England (HSE)	1998 & 2003	Binary	90.9	3.7	87.2
Kane, Rouse, Staiger	1999	Working paper	Education	NLS-72 Self-reported Education	1972-1979	Discrete High School Some College Bachelor's Degree	 6 7 5	 6 1 0	 0 6 5
Klerman, Ringel, Roth	2005	Working paper	Medicaid	CPS and Medi-Cal Eligibility Data System	1990-2000	Binary	30	2	28
Kreider and Pepper	2007	Journal of the American Statistical Association	Disability	Health and Retirement Survey (HRS) and SIPP	1992-1993;1996	Binary Disability beneficiaries Claimed no disability Gainfully employed No work limitation	 10 27 66 78	 - - - -	 - - - -
Kreider, Pepper, Gundersen, Jolliffe	2012	Journal of the American Statistical Association	SNAP	National Health and Nutrition Examination Survey (NHANES)	2001-2006	Binary	4	0	4
Kreider, Manski, Moeller, Pepper	2015	Health Economics	Insurance	HRS	2004 & 2006	Binary	6.5	-	-
Krueger and Rouse	1998	Journal of Labor Economics	Training	Survey and administrative data	1991-1994	Binary Manufacturing Company Service Company	 27 16	 20 13	 7 3
Meyer and Mittag	2019	Southern Economic Journal	Food stamp	FoodAPS	2012	Binary	19.5	1.2	18.3
Meyer, Mittag, George	2020	Journal of Human Resources	Food stamp	American Community Survey (ACS) CPS SIPP	2001 2002-2005 2001-2005	Binary Binary Binary	33.81 49.82 24.46	0.73 0.84 1.64	33.08 48.98 22.82
Meyer and Mittag	2019	American Economic Journal: Applied Economics	Poverty	CPS-ASEC	2008-2011	Binary SNAP Public assistant Housing assistant	 45 63.7 39	 2 0.7 3	 43 63 36
Mitchell	2010	Journal of Marriage and Family	Divorce	Life Events and Satisfaction Study	1995	Binary	3.6	-	-
Possebom	2022	Working paper	Crime	Justice Court System in the State of São Paulo, Brazil	2010-2017	Binary	4.4	3.5	0.9

Notes: We reviewed articles published between 1996 and 2022 and found 28 articles and 54 reportings of treatment misclassification in most fields of applied economics. The misclassification rate is the sum of false negative and false positive probabilities.

Table A2: Review of Articles (continue)

Authors	Year	Journal	Topic	Data	Year	Type of treatment		Misclass. rate	False positive (%)	False negative (%)
Savoca	2000	Health Services & Outcomes Research Methodology	Psychiatric Diseases	U.S. National Institute of Mental Health Epidemiological Catchment Area Survey	1980s	Binary	Drug abuse	57.1	2.2	54.9
							Alcohol abuse	56.4	39.7	16.7
							Anti-social personality	46.6	3.8	42.8
							Somatization	98	0.1	97.9
							Panic disorders	98.2	0.1	98.1
							Major depression	94.5	0.6	93.9
							Agoraphobia	84.4	1.0	83.4
							Social phobia	83	0.6	82.4
							Simple phobia	99.7	2.9	96.8
							Obsessive-compulsive disorder	97.4	0.6	96.8
Wineman et al.	2020	Journal of Agricultural Economics	Agriculture	Varietal Monitoring for Realized Productivity and Value in Tanzania Survey	2016-2017	Binary		30	16	14
Wossen et al.	2019	American Journal of Agricultural Economics	Productivity	Cassava Monitoring Survey	2015-2016	Binary		35	10	25
Wossen et al.	2019	Food Policy	Poverty	Agricultural Technology Adoption Household Survey	2015-2016	Binary		34	15	19

Notes: This table continues from the previous Table A1.

A.1 MR-LATE and the Local Average Response Function

We illustrate the link between MR-LATE and the local average response function (LARF) (Abadie, 2003) in the case with binary treatment and binary instrument. Recall that $C = \{D_0 = 0, D_1 = 1\}$ stands for the collection of compliers. We have

$$\mathbb{E}[Y|D,C] = D\mathbb{E}[Y_1|C] + (1-D)\mathbb{E}[Y_0|C], \tag{A1}$$

where the equality is because $D = Z$ given compliers and thus $\mathbb{E}[Y|D = d, C] = \mathbb{E}[Y_d|C]$ for $d \in \{0,1\}$. If D is observable, Abadie (2003) shows that the LARFs, $\mathbb{E}[Y_1|C]$ and $\mathbb{E}[Y_0|C]$, can be identified by the observable (Y, D, Z):

$$\begin{aligned} \mathbb{E}[Y|D=1,C] = \mathbb{E}[Y_1|C] &= \frac{\mathbb{E}[YD|Z=1]-\mathbb{E}[YD|Z=0]}{\mathbb{E}[D|Z=1]-\mathbb{E}[D|Z=0]}, \\ \mathbb{E}[Y|D=0,C] = \mathbb{E}[Y_0|C] &= \frac{\mathbb{E}[Y(1-D)|Z=1]-\mathbb{E}[Y(1-D)|Z=0]}{\mathbb{E}[(1-D)|Z=1]-\mathbb{E}[(1-D)|Z=0]}. \end{aligned} \tag{A2}$$

Then, the LATE of Imbens and Angrist (1994) is exactly the difference between the two identifiable LARFs in (A2).

Now, consider the case where D is *not* observed. Based on the procedure of MR-LATE, suppose two binary treatment indicators T^a and T^b are available satisfying Assumptions 3.1-3.3.[5] Given the observable (Y, T^a, T^b, Z), one can show that:

$$\begin{aligned} \lambda^a &= \frac{\mathbb{E}[YT^a|Z=1]-\mathbb{E}[YT^a|Z=0]}{\mathbb{E}[T^a|Z=1]-\mathbb{E}[T^a|Z=0]} = q^a\mathbb{E}[Y_1|C] + (1-q^a)\mathbb{E}[Y_0|C], \\ \lambda^b &= \frac{\mathbb{E}[YT^b|Z=1]-\mathbb{E}[YT^b|Z=0]}{\mathbb{E}[T^b|Z=1]-\mathbb{E}[T^b|Z=0]} = q^b\mathbb{E}[Y_1|C] + (1-q^b)\mathbb{E}[Y_0|C], \end{aligned} \tag{A3}$$

where we denote $q^j = p_1^j/(p_1^j - p_0^j)$ with $j = a, b$ and $p_1^j = \mathbb{E}[T_1^j|C]$, $p_0^j = \mathbb{E}[T_0^j|C]$. A comparison between the LARFs in (A2) and λ^a, λ^b in (A3) reveals three important connections between MR-LATE and LARF. Firstly, λ^a and λ^b have the same expression with LARF in terms of the conditional means of the observable variables, via replacing D and $1-D$ in (A2) with the observable T^a and T^b. Thus, λ^a and λ^b aim to mimic the two LARFs with observable treatment proxies, capturing the information in D by T^a and in $1-D$ by T^b. Secondly, due to the potential misclassification error in T^a and T^b, in general, λ^a and λ^b fail to recover the exact LARFs, but provide a linear combination of them with weights q^j and $1-q^j$ (see the right hand side of (A3)). Lastly and most importantly, MR-LATE is able to point identify the LARFs with well-chosen T^a and T^b:

$$\lambda^a = \mathbb{E}[Y_1|C] \text{ if } q^a = 1 \text{ and } \lambda^b = \mathbb{E}[Y_0|C] \text{ if } q^b = 0.$$

[5] For binary instrument, Assumption 3.3-(ii) is not needed.

Without loss of generality, we rule out the cases where $p_1^j - p_0^j = 0$. By definition of q^j:[6]

$$
\begin{aligned}
&q^a = 1, \text{ if and only if } \Pr(T^a = D|D = 0, C) = 1, \\
&q^b = 0, \text{ if and only if } \Pr(T^b = 1 - D|D = 1, C) = 1.
\end{aligned}
\tag{A4}
$$

Under the conditions in (A4), MR-LATE point identifies the LARFs and the LATE.

The connection between λ^a and λ^b to the LARFs discussed above thus sheds lights on the point identification mechanism of the MR-LATE approach. Based on (A4), MR-LATE aims to point identifies two LARFs by utilizing two treatment proxies satisfying that T^a never mistakes the true untreated compliers as treated and $1 - T^b$ never mistakes the true treated compliers as untreated. In a special case where D is observed, simply setting $T^a = D$ and $T^b = 1 - D$ ensures desirable point identification.

[6]Due to $D = Z$ given compliers $p_1 = \mathbb{E}[T_1|C] = \Pr(T = 1|D = 1, C)$, $p_0 = \mathbb{E}[T_0|C] = \Pr(T = 1|D = 0, C)$. Then, $\Pr(T^a = D|D = 0, C) = 1 \iff p_0^a = 0 \iff q^a = 1$. Similarly, $\Pr(T^b = 1 - D|D = 1, C) = 1 \iff p_1^b = 0 \iff q^b = 0$.

A.2 Proofs of Section 3

A.2.1 Proof of Lemma 3.1

Consider $p_{1,k}$ as an example. The proof for $p_{0,k}$ is the same. Given C_k, we have that $D = 1[Z = z_k]$. Due to Assumption 3.2-(i) (extended unconfoundedness), D is independent to T_1 given C_k. Then,

$$\begin{aligned} p_{1,k} &= \mathbb{E}[T_1|C_k] = \mathbb{E}[T_1|D = 1, C_k] \\ &= \mathbb{E}[T|D = 1, C_k] \\ &= \mathbb{E}[T|D = 1], \end{aligned} \tag{A5}$$

where the last equality is due to Assumption 3.2-(ii). Denoting $p_1 = \mathbb{E}[T|D = 1]$ fulfills the proof.

A.2.2 Proof of Theorem 3.1

First, we show that $\lambda_k^a - \lambda_k^b = (q_k^a - q_k^b)\mathbb{E}[Y_1 - Y_0|C_k]$. By definition, $YT = [Y_0 + D(Y_1 - Y_0)][T_0 + D(T_1 - T_0)]$. By the virtue of Assumption 3.2-(i) (extended unconfoundedness),

$$\begin{aligned} &\mathbb{E}(YT|Z = z_k) - \mathbb{E}(YT|Z = z_{k-1}) \\ =&\mathbb{E}\{[Y_0 + D_k(Y_1 - Y_0)][T_0 + D_k(T_1 - T_0)]|Z = z_k\} \\ &\qquad - \mathbb{E}\{[Y_0 + D_{k-1}(Y_1 - Y_0)][T_0 + D_{k-1}(T_1 - T_0)]|Z = z_{k-1}\} \\ =&\mathbb{E}[Y_0T_0 + D_k(Y_1T_1 - Y_0T_0)] - \mathbb{E}[Y_0T_0 + D_{k-1}(Y_1T_1 - Y_0T_0)] \\ =&\mathbb{E}[(D_k - D_{k-1})(Y_1T_1 - Y_0T_0)] \\ =&\mathbb{E}[Y_1T_1 - Y_0T_0|C_k]\Pr(C_k), \end{aligned} \tag{A6}$$

where the last equality is due to $\Pr(D_k - D_{k-1} = -1) = 0$. Replacing Y in the above derivations with one gives us

$$\mathbb{E}(T|Z = z_k) - \mathbb{E}(T|Z = z_{k-1}) = \mathbb{E}[T_1 - T_0|C_k]\Pr(C_k).$$

By the fact that $D = 1[Z = z_k]$ given C_k, we have $\mathbb{E}[Y_dT_d|C_k] = \mathbb{E}[Y_d|C_k, T_d = 1]\mathbb{E}[T_d|C_k] = \mathbb{E}[Y_d|D = d, C_k, T_d = 1]\mathbb{E}[T_d|C_k] = \mathbb{E}[Y_d|C_k]\mathbb{E}[T_d|C_k]$, where the last equality is because of Assumption 3.3-(i) and the independence of the instrument. Thus,

$$\lambda_k := \frac{\mathbb{E}(YT|Z = z_k) - \mathbb{E}(YT|Z = z_{k-1})}{\mathbb{E}(T|Z = z_k) - \mathbb{E}(T|Z = z_{k-1})} = \mathbb{E}[q_kY_1 + (1 - q_k)Y_0|C_k],$$

and simple calculation leads to $\lambda_k^a - \lambda_k^b = (q_k^a - q_k^b)\mathbb{E}[Y_1 - Y_0|C_k]$.

Next, we move on to λ. The numerator of λ can be rewritten as

$$\begin{aligned}
&\mathbb{E}[YT(g(Z)-\mathbb{E}[g(Z)])] \\
&=\sum_{l=0}^{K}\mathbb{E}(YT|Z=z_l)(g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
&=\sum_{l=0}^{K}\Big[\mathbb{E}(YT|Z=z_0)+\mathbb{E}(YT|Z=z_1)-\mathbb{E}(YT|Z=z_0)+\ldots \\
&\qquad +\mathbb{E}(YT|Z=z_l)-\mathbb{E}(YT|Z=z_{l-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
&=\sum_{l=0}^{K}\mathbb{E}(YT|Z=z_0)(g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
&\qquad +\sum_{l=0}^{K}\sum_{k=1}^{l}\Big[\mathbb{E}(YT|Z=z_k)-\mathbb{E}(YT|Z=z_{k-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
&=\sum_{k=1}^{K}\Big[\mathbb{E}(YT|Z=z_k)-\mathbb{E}(YT|Z=z_{k-1})\Big]\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l. \qquad \text{(A7)}
\end{aligned}$$

Substitute (A6) into (A7) and apply Assumption 3.3,

$$\begin{aligned}
\mathbb{E}[YT(g(Z)-\mathbb{E}[g(Z)])] &= \sum_{k=1}^{K}\mathbb{E}[Y_1T_1-Y_0T_0|C_k]\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
&= \sum_{k=1}^{K}\mathbb{E}[p_{1,k}Y_1-p_{0,k}Y_0|C_k]\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l.
\end{aligned}$$

The denominator of λ can be obtained by replacing Y with one. By Assumption 3.3 (non-differential misclassification) we have

$$\begin{aligned}
\lambda &= \sum_{k=1}^{K}\frac{(p_{1,k}-p_{0,k})\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}{\sum_{m=1}^{K}(p_{1,m}-p_{0,m})\Pr(C_m)\sum_{l=m}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}\times\frac{\mathbb{E}\left[p_{1,k}Y_1-p_{0,k}Y_0|C_k\right]}{(p_{1,k}-p_{0,k})} \\
&= \sum_{k=1}^{K}w_k\mathbb{E}\left[q_kY_1+(1-q_k)Y_0|C_k\right], \qquad \text{(A8)}
\end{aligned}$$

where denote $w_k=\frac{(p_{1,k}-p_{0,k})\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}{\sum_{m=1}^{K}(p_{1,m}-p_{0,m})\Pr(C_m)\sum_{l=m}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l}$, and denote w_k^a and w_k^b as the w_k associated with T^a and T^b. Then, by (A8):

$$\begin{aligned}
\rho=\lambda^a-\lambda^b &= \sum_{k=1}^{K}w_k^a\mathbb{E}\left[Y_0+q_k^a(Y_1-Y_0)|C_k\right]-\sum_{k=1}^{K}w_k^b\mathbb{E}\left[Y_0+q_k^b(Y_1-Y_0)|C_k\right] \\
&= \sum_{k=1}^{K}(w_k^a-w_k^b)\mathbb{E}\left[Y_0|C_k\right]+\sum_{k=1}^{K}(w_k^aq_k^a-w_k^bq_k^b)\mathbb{E}\left[Y_1-Y_0|C_k\right]. \qquad \text{(A9)}
\end{aligned}$$

It yields from Lemma 3.1 that if Assumption 3.2-(ii) holds, then $p_{1,k}^a=p_1^a$, $p_{0,k}^a=p_0^a$, $p_{1,k}^b=p_1^b$ and

$p_{0,k}^b = p_0^b$, which implies $q_k^a = q^a$, $q_k^b = q^b$. In addition, $w_k^a = w_k^b = \gamma_k^{IV}$. Thus, by (A9) we get

$$\begin{aligned}\rho &= \sum_{k=1}^{K} (\gamma_k^{IV} q^a - \gamma_k^{IV} q^b)\mathbb{E}[Y_1 - Y_0|C_k] = (q^a - q^b)\sum_{k=1}^{K}\gamma_k^{IV}\mathbb{E}[Y_1 - Y_0|C_k] \\ &= (q^a - q^b)\alpha^{IV}. \end{aligned} \tag{A10}$$

A.2.3 Proof of Theorem 3.2

Recall $W_i = (Y_i, T_i^a, T_i^b, Z_i)$ and $g(Z_i) = g(Z_i;\theta)$. Suppose that $\hat{\theta}_n$ solves a $d_\theta \times 1$ vector $\sum_{i=1}^n \psi(W_i;\theta) = 0$ where d_θ is the dimension of θ. Assume that there is a unique solution to $\mathbb{E}[\psi(W;\theta)] = 0$ and $\partial\mathbb{E}[\psi(W;\theta)]/\partial\theta'$ is of full rank. Let $\epsilon_i^j = Y_i T_i^j - \gamma^j - \lambda^j T_i^j$. Denote a $d_\eta \times 1$ moment function

$$h(W_i;\eta) = \begin{bmatrix} \psi(W_i;\theta) \\ Y_i T_i^a - \gamma^a - \lambda^a T_i^a \\ Y_i T_i^b - \gamma^b - \lambda^b T_i^b \\ g(Z_i;\theta)(Y_i T_i^a - \gamma^a - \lambda^a T_i^a) \\ g(Z_i;\theta)(Y_i T_i^b - \gamma^b - \lambda^b T_i^b) \end{bmatrix} = \begin{bmatrix} \psi(W_i;\theta) \\ \epsilon_i^a \\ \epsilon_i^b \\ g(Z_i;\theta)\epsilon_i^a \\ g(Z_i;\theta)\epsilon_i^b \end{bmatrix}.$$

We have that $\mathbb{E}[h(W_i;\eta^0)] = 0$, where the last two moment conditions come from the definition of λ^j. Then $\hat{\eta}_n = (\hat{\theta}_n', \hat{\gamma}_n^a, \hat{\gamma}_n^b, \hat{\lambda}_n^a, \hat{\lambda}_n^b)$ solves $\frac{1}{n}\sum_{i=1}^n h(W_i;\eta) = 0$. By the mean-value theorem, we have

$$0 = \frac{1}{n}\sum_{i=1}^n h(W_i;\hat{\eta}_n) = \frac{1}{n}\sum_{i=1}^n h(W_i;\eta^0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial h(W_i;\tilde{\eta}_n)}{\partial\eta'}(\hat{\eta}_n - \eta^0),$$

where $\tilde{\eta}_n$ is element-wise between η^0 and $\hat{\eta}_n$. Denote

$$H = \mathbb{E}\left[\frac{\partial h(W_i;\eta^0)}{\partial\eta'}\right] = \mathbb{E}\begin{pmatrix} \frac{\partial\psi(W_i;\theta^0)}{\partial\theta'} & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -T_i^a & 0 \\ 0 & 0 & -1 & 0 & -T_i^b \\ \epsilon_i^a\frac{\partial g(Z_i;\theta^0)}{\partial\theta'} & -g(Z_i;\theta^0) & 0 & -g(Z_i;\theta^0)T_i^a & 0 \\ \epsilon_i^b\frac{\partial g(Z_i;\theta^0)}{\partial\theta'} & 0 & -g(Z_i;\theta^0) & 0 & -g(Z_i;\theta^0)T_i^b \end{pmatrix}.$$

Since $\mathrm{Cov}(g(Z_i,\theta^0), T_i^a) \neq 0$ and $\mathrm{Cov}(g(Z_i,\theta^0), T_i^b) \neq 0$, we have that H is invertible. For large enough sample size n, $\frac{1}{n}\sum_{i=1}^n \frac{\partial h(W_i;\tilde{\eta}_n)}{\partial\eta'}$ is also invertible and we can obtain that

$$\sqrt{n}(\hat{\eta}_n - \eta^0) = -\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial h(W_i;\tilde{\eta}_n)}{\partial\eta'}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n h(W_i;\eta^0),$$

Let $\Sigma = \mathrm{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n h(W_i;\eta^0)\right]$. Under standard regularity conditions, by central limit theorem we can show $\sqrt{n}(\hat{\eta}_n - \eta^0) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1'})$. By the delta method, the asymptotic variance of $\sqrt{n}(\hat{\rho}_n - \rho^0)$ is $\delta H^{-1}\Sigma H^{-1'}\delta'$, with $\delta = (\mathbf{0}_{1\times d_\theta}, 0, 0, 1, -1)$.

A.3 MR-LATE with a Discrete Treatment Variable: Details

We begin by introducing some notation using a version of the Rubin's causal model that allows for variable treatment intensity and discrete or multiple discrete instruments (Angrist and Imbens, 1995). We derive our results without conditioning on covariates X, as everything immediately extends to conditioning on them.

Model Setup. Let $D \in \Omega_D = \{0, 1, 2, ..., J\}$ be the true discrete treatment variable that affects the outcome of interest. The instrument Z is defined the same as in Section 3.1. Let the random variable D_k, for $k = 0, 1, ..., K$, be the potential treatments associated to $Z = z_k$. Denote by Y_j the potential outcome if an individual was assigned to $D = j$. Then, we can write

$$D = \sum_{k=0}^{K} 1[Z = z_k] D_k, \qquad Y = \sum_{j=0}^{J} 1[D = j] Y_j.$$

Assumption A.1. *Y, D, and Z satisfy the standard Angrist and Imbens (1995) assumptions:*

(i) $(\{Y_j\}_{j=0}^{J}, \{D_k\}_{k=0}^{K}, Z)$ *are i.i.d. across all individuals and have finite first and second moments;*

(ii) *(Unconfoundedness)* $Z \perp (\{Y_j\}_{j=0}^{J}, \{D_k\}_{k=0}^{K})$ *and* $\mathbb{E}(D|Z = z)$ *is a nontrivial function of* z*;*

(iii) *(First stage)* $Cov(D, g(Z)) \neq 0$*;*

(iv) *(Monotonicity) With probability one, either* $D_l \geq D_w$ *for all individuals, or* $D_l \leq D_w$ *for all individuals. Furthermore, for all* $z_l, z_w \in \Omega_Z$*, either* $Pr(z_l) \leq Pr(z_w)$ *implies* $g(z_l) \leq g(z_w)$*, or* $Pr(z_l) \leq Pr(z_w)$ *implies* $g(z_l) \geq g(z_w)$*.*

True Effect. Under Assumption A.1, the average causal response (ACR) of Angrist and Imbens (1995) is defined as

$$\beta_{k,k-1} = \frac{\mathbb{E}[Y|Z = z_k] - \mathbb{E}[Y|Z = z_{k-1}]}{\mathbb{E}[D|Z = z_k] - \mathbb{E}[D|Z = z_{k-1}]} = \sum_{j=1}^{J} \omega_j \mathbb{E}[Y_j - Y_{j-1}|D_k \geq j > D_{k-1}],$$

with $\omega_j = \frac{\Pr(D_k \geq j > D_{k-1})}{\sum_{j=1}^{J} \Pr(D_k \geq j > D_{k-1})}$ for $j = 1, ..., J$, where ω_j is nonnegative and $\sum_{j=1}^{J} \omega_j = 1$. When D is observed, the IV estimand identifies the weighted average of ACRs (WACR):

$$\beta^{IV} = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(D, g(Z))} = \sum_{k=1}^{K} \mu_k^{IV} \beta_{k,k-1}, \tag{A11}$$

with $\mu_k^{IV} = \frac{(\mathbb{E}[D|Z=z_k] - \mathbb{E}[D|Z=z_{k-1}]) \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}{\sum_{m=1}^{K} (\mathbb{E}[D|Z=z_m] - \mathbb{E}[D|Z=z_{m-1}]) \sum_{l=m}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}$ where μ_k^{IV} is nonnegative and $\sum_{k=1}^{K} \mu_k^{IV} = 1$.

Mismeasured Effect. Suppose D is unobservable and we can only observe a treatment proxy T, which may suffer from the misclassification error. Denote by $T_j \in \Omega_D$ the potential observed treatment for possible realization $D = j$. By definition:

$$T = T_D = \sum_{j=0}^{J} 1[D = j] T_j.$$

Assumption A.2. *The treatment proxy T is such that the following conditions are satisfied:*

(i) (Extended unconfoundedness) $Z \perp (\{Y_j\}_{j=0}^{J}, \{D_k\}_{k=0}^{K}, \{T_j\}_{j=0}^{J})$;

(ii) (Extended first stage) $Cov(T, g(Z)) \neq 0$.

If we replace D by the proxy T, we obtain a mismeasured IV estimand which is useful to characterize the bias caused by treatment misclassification.

Theorem A.1. *Let Assumptions A.1 and A.2 hold. Then, we have*

$$\beta^{Mis} = \frac{Cov(Y, g(Z))}{Cov(T, g(Z))} = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))(g(Z) - \mathbb{E}[g(Z)])]}{\mathbb{E}[(T - \mathbb{E}(T))(g(Z) - \mathbb{E}[g(Z)])]} = \sum_{k=1}^{K} \mu_k^{Mis} \beta_{k,k-1},$$

with $\mu_k^{Mis} = \frac{(\mathbb{E}[D|Z=z_k] - \mathbb{E}[D|Z=z_{k-1}]) \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}{\sum_{m=1}^{K} (\mathbb{E}[T|Z=z_m] - \mathbb{E}[T|Z=z_{m-1}]) \sum_{l=m}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}$.

Proof of Theorem A.1. See Appendix A.4.1. □

Thus, the IV estimand β^{Mis} denotes the mismeasured treatment effect that is identifiable if the misclassification error is ignored. The results above generalizes the earlier result in Tommasi and Zhang (2020) to settings with discrete treatment intensity. In general, due to misclassification, we know that $\beta^{Mis} \neq \beta^{IV}$ because $\mu_k^{Mis} \neq \mu_k^{IV}$.

Relationship between the true and mismeasured effect. We find that a simple relationship between β^{Mis} and β^{IV} which can be captured by a summary statistics of the weighted average of misclassification probabilities.

Corollary A.1. *Under Assumptions A.1 and A.2, there exists a summary statistic ξ such that:*

$$\beta^{Mis} = \sum_{i=1}^{K} \mu_k^{IV} \beta_{k,k-1} \times \frac{\mu_k^{Mis}}{\mu_k^{IV}} \implies \beta^{IV} = \xi \beta^{Mis} \tag{A12}$$

where the ratio $\xi = \mu_k^{IV} / \mu_k^{Mis} = \sum_{k=1}^{K} \mu_k^{IV} \xi_{k,k-1}$ is a constant for all k and

$$\xi_{k,k-1} := \frac{\mathbb{E}[T|Z = z_k] - \mathbb{E}[T|Z = z_{k-1}]}{\mathbb{E}[D|Z = z_k] - \mathbb{E}[D|Z = z_{k-1}]} = \sum_{j=1}^{J} \omega_j \mathbb{E}[T_j - T_{j-1}|D_k \geq j > D_{k-1}].$$

Proof of Corollary A.1. See Appendix A.4.2. □

The parameter ξ is a weighted average of the difference between mismeasured potential observed treatments. Corollary A.1 demonstrates that, when the treatment variable is misclassified, the magnitude of the bias in β^{Mis} is determined by the factor $1/\xi$.

A.3.1 Point Identification with a Binary IV

Consider a simple case where the instrument is binary. From the results in the main text, we know that when the true treatment is binary, we only need two binary proxies to point identify α^{IV} (one for $D = 1$ and the other for $D = 0$). As a straightforward extension, when the true treatment is discrete, with $J + 1$ possible realizations, we require one binary proxy for each realization of D. Given the observable discrete treatment proxy $T \in \Omega_D$, we can construct an indicator $T^{(d)}$ for each treatment level $T = d$ with $d \in \Omega_D$ as

$$T^{(d)} := 1[T = d] = \sum_{j=0}^{J} 1[D = j] T_j^{(d)}, \tag{A13}$$

where the unobservable binary variable $T_j^{(d)} := 1[T_j = d]$ can be understood as an indicator of the treatment misclassification in the treatment level $D = j$. Specifically, $T_j^{(d)} = 1$ implies that the true status $D = j$ can be correctly indicated by $T^{(d)}$, while $T_j^{(d)} = 0$ implies that the true status $D = j$ cannot be correctly indicated by $T^{(d)}$. Because there are $J + 1$ possible treatment values, we can construct $\{T^{(d)}\}_{d=0}^{J}$ indicators based on T. In this case, $\{T^{(d)}\}_{d=0}^{J}$ are extensions of the T^a and T^b in Section 3.2. If the treatment is binary, we have $J = 1$ and $\{T^{(d)}\}_{d=0}^{J} = \{T^a, T^b\}$.

When D is observed with no error, the $J+1$ treatment indicators can be defined as $T^{(d)} = 1[D = d]$ so that it is a correct indicator of the event $(D = d)$. In the presence of misclassification, the treatment proxy T can seldom provide correct information about the realizations of D. Because the instrument is binary, denote $p_{j,k}^{(d)} = \mathbb{E}[T_j^{(d)} | D_1 \geq k > D_0]$ for $j, k \in \Omega_D$, which is the probability of $T^{(d)}$ correctly indicating the treatment realization of individuals in subgroup $D_1 \geq k > D_0$, if they were assigned to $D = j$. Similarly to the binary treatment setting, define $\lambda^{(d)}$ as

$$\lambda^{(d)} = \frac{\mathrm{Cov}(Y T^{(d)}, Z)}{\mathrm{Cov}(T^{(d)}, Z)}. \tag{A14}$$

Note that the subgroup $(D_1 \geq j > D_0)$ for $j = 1, 2, ..., J$ is potentially overlapping with each other because the change in the instrument can trigger larger than one-unit change in the treatment level. However, as argued by Angrist and Imbens (1995), the instrument typically would not cause more than one-unit incremental change in the treatment. Thus, it is reasonable to make this assumption to ease the illustration.

Assumption A.3. *For $\forall j \in \{1, 2, ..., J\}$ and for $\forall d \in \Omega_D$, we have*

(i) (One-unit incremental change) $(D_1 \geq j > D_0)$ is equivalent to $(D_1 = j, D_0 = j - 1)$;

(ii) (Non-differential misclassification) $\mathbb{E}[Y | T^{(d)}, D, D_1 \geq j > D_0] = \mathbb{E}[Y | D, D_1 \geq j > D_0]$ for all d;

(iii) (Homogeneous misclassification) $\mathbb{E}\left[T^{(d)} | D, D_1, D_0\right] = \mathbb{E}[T^{(d)} | D]$.

The Assumption A.3-(ii) is in line with the *non-differential misclassification*. It implies that the treatment indicator $T^{(d)}$ has no direct effects on the local average response function $\mathbb{E}[Y | D, D_1 \geq$

$j > D_0$], once the actual treatment D is controlled for. Assumption A.3-(iii) is apparently an analog of the *homogeneous misclassification* assumption employed in the binary treatment case. It requires that conditional on the actual treatment intensity D, the treatment indicator $T^{(d)}$ does not depend on (D_0, D_1). Therefore, two types of measurement errors are accommodated by Assumptions A.3: those random errors that are independently generated, $(T_0^{(d)}, T_1^{(d)}) \perp (Y, D, Z)$, and those errors that correlated with potential outcomes but only via the actual treatment D.

Theorem A.2. *Suppose Assumptions A.1 to A.3 hold. Then we have*

$$\lambda^{(d)} = \sum_{j=1}^{J} \frac{Pr(D_1 \geq j > D_0)}{\sum_{k=1}^{J}(p_{k,k}^{(d)} - p_{k-1,k}^{(d)})Pr(D_1 \geq k > D_0)} \mathbb{E}[p_{j,j}^{(d)} Y_j - p_{j-1,j}^{(d)} Y_{j-1} | D_1 \geq j > D_0].$$

Proof of Theorem A.2. See Appendix A.4.3. □

To point identify the ACR, we need to introduce more assumptions on the treatment indicator $T^{(d)}$, as well as on the potential outcomes.

Assumption A.4. *For $j = 1, 2, ..., J-1$ and $k > j$,*

$$\mathbb{E}[Y_j | D_1 \geq k > D_0] = \mathbb{E}[Y_j | D_1 \geq j > D_0].$$

In addition, we have $\mathbb{E}[Y_0 | D_1 \geq k > D_0] = \mathbb{E}[Y_0 | D_1 \geq 1 > D_0]$, for all $k > 1$.

Assumption A.4 means that the conditional mean of the potential outcome Y_j is indifferent for the subgroups $D_1 \geq k > D_0$ and $D_1 \geq j > D_0$, as long as $k > j$. The mean independence of this sort does not impose the mean independence across all subgroups. This assumption is needed to show the point identification of ACR. It is quite strong and may not be testable. Later, we also consider to relax it and to obtain set identification results. In the following theorem, we demonstrate that under some conditions that are similar to the one type misclassification assumption in the binary treatment case, we can point identify an ACR. Denote $p_{0,0}^{(d)} = \mathbb{E}[T_0^{(d)} | D_1 = 0, D_0 = 0]$.

Theorem A.3. *Let Assumptions A.1 to A.3 hold.*

(i) If there exists a $T^{(J)}$ such that $p_{J,J}^{(J)} \neq 0$ and $p_{j,j}^{(J)} = 0$ for $\forall j \neq J$, then $\lambda^{(J)} = \mathbb{E}[Y_J | D_1 \geq J > D_0]$.

(ii) If there exists a $T^{(0)}$ such that $p_{0,0}^{(0)} \neq 0$ and $p_{j,j}^{(0)} = 0$ for $\forall j \neq 0$, then $\lambda^{(0)} = \mathbb{E}[Y_0 | D_1 \geq 1 > D_0]$.

(iii) Further assume Assumption A.4 holds. If there exists a $T^{(d)}$ with $d \in \{1, 2, ..., J-1\}$ such that $p_{d,d}^{(d)} \neq 0$ and $p_{j,j}^{(d)} = 0$ for any $j \neq d$, then $\lambda^{(d)} = \mathbb{E}[Y_d | D_1 \geq d > D_0]$.

Proof of Theorem A.3. See Appendix A.4.5. □

Corollary A.2. *For $\forall j, j' \in \{0, 1, .., J\}$ such that $j' < j$, if there exist $T^{(j)}$ and $T^{(j')}$ as described in Theorem A.3 (i)-(iii), then $\lambda^{(j)} - \lambda^{(j')} = \mathbb{E}[Y_j - Y_{j'} | D_1 \geq j > D_0]$.*

Proof of Corollary A.2. See Appendix A.4.6. □

A.3.2 Set Identification with a Binary IV

We consider now relaxing Assumption A.4 to achieve set identification using binary IV.

Assumption A.5. *$\mathbb{E}[Y_j|D_1 \geq i > D_0] \geq \mathbb{E}[Y_j|D_1 \geq i' > D_0]$, for all $i, i', j \in \{0, 1, ..., J\}$ and $i > i'$.*

Assumption A.5 above imposes mean dominance restriction among complier subgroups, which is empirically suitable and may be implied by economic theory. Chen et al. (2018) employ similar mean stochastic dominance across strata, in binary treatment and binary instrument setting. Here, we extend it to a discrete treatment framework. The direction of the inequality in Assumption A.5 should be decided based on the case by case study.

In what follows, we use a simple example to explain the set identification.

Corollary A.3. *Consider $J = 2$. Suppose Assumptions A.1 to A.3 and A.5 hold, and assume $\mathbb{E}[T^{(1)}|Z = 1] - \mathbb{E}[T^{(1)}|Z = 0] > 0$. If $T^{(0)}$, $T^{(1)}$, and $T^{(2)}$ satisfy the conditions in Theorem A.3, then*

$$\begin{aligned}
&\lambda^{(1)} - \lambda^{(0)} \leq \mathbb{E}[Y_1 - Y_0|D_1 \geq 1 > D_0], \\
&\lambda^{(2)} - \lambda^{(1)} \geq \max\Big\{\mathbb{E}[Y_2 - Y_1|D_1 \geq 2 > D_0],\ \mathbb{E}[Y_2 - Y_1|D_1 \geq 1 > D_0]\Big\}, \\
&\lambda^{(2)} - \lambda^{(0)} \geq \max\Big\{\mathbb{E}[Y_2 - Y_0|D_1 \geq 2 > D_0],\ \mathbb{E}[Y_2 - Y_0|D_1 \geq 1 > D_0]\Big\}.
\end{aligned}$$

Proof of Corollary A.3. See Appendix A.4.8. □

A.3.3 Set Identification in the General Case with a Discrete IV

In this section, we provide a set identification result for β^{IV} under more plausible conditions.

Theorem A.4. *Let Assumptions A.1 and A.2 hold. If $Cov(D, g(Z))$ and $Cov(T, g(Z))$ have the same sign, then β^{Mis} signs β^{IV}. In addition, assume there exist known constants $\underline{\xi}$ and $\overline{\xi}$ such that $\underline{\xi} \leq \xi \leq \overline{\xi}$.*

(i) If $\beta^{Mis} \geq 0$, then $0 \leq \underline{\xi}\beta^{Mis} \leq \beta^{IV} \leq \overline{\xi}\beta^{Mis}$.

(ii) If $\beta^{Mis} < 0$, then $\overline{\xi}\beta^{Mis} \leq \beta^{IV} \leq \underline{\xi}\beta^{Mis} < 0$.

The proof of Theorem A.4 follows directly from the expression in (A12), therefore omitted. Compared to the assumptions required by point identification in the previous section, the advantages here are twofold. First, we are able to deal with endogenous and heterogeneous treatment misclassifications, since no restrictions on the dependence of the misclassification errors to the potential outcomes and the potential treatments, are imposed. Second, we do not require the number of treatment indicators to be the same with the number of treatment categories, since one treatment proxy under relatively weak condition suffices the set identification.

A.3.4 Inference

Assume the function $g(Z) = g(Z; \theta)$ with $\theta \in \mathbb{R}^{d_\theta}$ which can be estimated as $\hat{g}(Z_i) := g(Z; \hat{\theta}_n)$. Denote $\boldsymbol{\lambda} = (\lambda^{(0)}, \lambda^{(1)}, ..., \lambda^{(J)})'$, $\gamma^{(d)} = \mathbb{E}[Y_i T_i^{(d)}] - \lambda^{(d)}\mathbb{E}[T_i^{(d)}]$ and $\eta = (\theta', \gamma^{(0)}, ..., \gamma^{(J)}, \lambda^{(0)}, ..., \lambda^{(J)})' \in$

Γ with true value η^0. Given $W_i = \{Y_i, T_i^{(0)}, ..., T_i^{(J)}, Z_i\}_{i=1}^n$, let

$$\hat{\lambda}_n^{(j)} = \frac{\sum_{i=1}^n \hat{g}(Z_i)(Y_i T_i^{(j)} - \overline{YT}_n^{(j)})}{\sum_{i=1}^n \hat{g}(Z_i)(T_i^{(j)} - \overline{T}_n^{(j)})}, \quad \text{where} \quad \overline{YT}_n^{(j)} = \frac{1}{n}\sum_{i=1}^n Y_i T_i^{(j)} \text{ and } \overline{T}_n^{(j)} = \frac{1}{n}\sum_{i=1}^n T_i^{(j)}.$$

Let $\hat{\boldsymbol{\lambda}}_n = (\hat{\lambda}_n^{(0)}, \hat{\lambda}_n^{(1)}, ..., \hat{\lambda}_n^{(J)})'$ and $\boldsymbol{\lambda}^0$ to be its true value. Denote by $\mathbf{I}_k$ a $k \times k$ identity matrix and $\mathbf{0}_{k\times k'}$ a $k \times k'$ matrix of zeros.

Theorem A.5. *Under Assumptions A.1 and A.2, we have*

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^0) \xrightarrow{d} \mathcal{N}\left(0, \delta \tilde{H}^{-1} \tilde{\Sigma} \tilde{H}^{-1'} \delta'\right),$$

where denote $\delta = (\mathbf{0}_{(J+1)\times(d_\theta+J+1)}, \mathbf{I}_{J+1})$, $\tilde{H} = \mathbb{E}\left[\frac{\partial \tilde{h}(W_i;\eta^0)}{\partial \eta'}\right]$ *a* $d_\eta \times d_\eta$ *matrix,* $\tilde{\Sigma} = Var\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{h}(W_i;\eta^0)\right]$, *and* $\tilde{h}(W_i;\eta)$ *is a* $d_\eta \times 1$ *vector of moment functions defined in the proof of this theorem.*

Proof of Theorem A.5. See Appendix A.4.9. □

Similar to the discussion in Section 3.4, let $\tilde{\mathbf{h}}_n(\eta) = (\tilde{h}(W_1;\eta), ..., \tilde{h}(W_n;\eta))'$ and G be a $n \times n$ matrix that describes the dependence structure of $\{W_i\}_{i=1}^n$. Then, a consistent estimator of $\tilde{H}$ is $\frac{1}{n}\sum_{i=1}^n \frac{\partial \tilde{h}(W_i;\hat{\eta}_n)}{\partial \eta'}$ and a consistent estimator of the covariance matrix $\tilde{\Sigma}$ can be expressed as

$$\hat{\tilde{\Sigma}} = \frac{1}{n}\tilde{\mathbf{h}}_n(\hat{\eta}_n)' G \tilde{\mathbf{h}}_n(\hat{\eta}_n). \tag{A15}$$

When samples are i.i.d., G is an identity matrix and $\hat{\tilde{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \tilde{h}(W_i;\hat{\eta}_n)\tilde{h}(W_i;\hat{\eta}_n)'$. By setting G to be a block-diagonal matrix with all the blocks as matrices of ones and zeros in the off-diagonals, $\hat{\tilde{\Sigma}}$ becomes an estimator accounting for clustered standard errors. With more general data correlation or heteroskedasticity, Newey and West (1987) can be applied via adjusting G accordingly.

A.4 Proofs of Section A.3

A.4.1 Proof of Theorem A.1

Assumption A.2-(ii) guarantees that the denominator of β^{Mis} is nonzero and thus β^{Mis} is well-defined. From the proof of Theorem 2 in Angrist and Imbens (1995), we know that

$$\mathbb{E}[Y(g(Z)-\mathbb{E}[g(Z)])]=\sum_{k=1}^{K}\Big[\mathbb{E}(D|Z=z_k)-\mathbb{E}(D|Z=z_{k-1})\Big]\sum_{l=k}^{K}\pi_l\,(g(z_l)-\mathbb{E}[g(Z)])\,\beta_{k,k-1}, \tag{A16}$$

so we only need to consider the denominator of β^{Mis}. We can obtain

$$\begin{aligned}
&\mathbb{E}[T(g(Z)-\mathbb{E}[g(Z)])]\\
=&\sum_{l=0}^{K}\mathbb{E}\big[T\big|Z=z_l\big](g(z_l)-\mathbb{E}[g(Z)])\,\pi_l\\
=&\sum_{l=0}^{K}\Big[\mathbb{E}(T|Z=z_0)+\mathbb{E}(T|Z=z_1)-\mathbb{E}(T|Z=z_0)+\ldots\\
&\qquad+\mathbb{E}(T|Z=z_l)-\mathbb{E}(T|Z=z_{l-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\,\pi_l\\
=&\sum_{l=0}^{K}\mathbb{E}(T|Z=z_0)(g(z_l)-\mathbb{E}[g(Z)])\pi_l+\sum_{l=0}^{K}\sum_{k=1}^{l}\Big[\mathbb{E}(T|Z=z_k)-\mathbb{E}(T|Z=z_{k-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\pi_l\\
=&\sum_{k=1}^{K}\Big[\mathbb{E}(T|Z=z_k)-\mathbb{E}(T|Z=z_{k-1})\Big]\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l.
\end{aligned} \tag{A17}$$

A.4.2 Proof of Corollary A.1

For $\forall k$, by definitions of μ_k^{IV} and μ_k^{Mis} we have:

$$\begin{aligned}
\frac{\mu_k^{IV}}{\mu_k^{Mis}}=&\sum_{k=1}^{K}\Bigg\{\frac{(\mathbb{E}[D|Z=z_k]-\mathbb{E}[D|Z=z_{k-1}])\sum_{l=k}^{K}\pi_l\,(g(z_l)-\mathbb{E}[g(Z)])}{\sum_{m=1}^{K}[\mathbb{E}(D|Z=z_k)-\mathbb{E}(D|Z=z_{k-1})]\sum_{l=m}^{K}\pi_l\,(g(z_l)-\mathbb{E}[g(Z)])}\\
&\qquad\times\frac{\mathbb{E}[T|Z=z_k]-\mathbb{E}[T|Z=z_{k-1}]}{\mathbb{E}[D|Z=z_k]-\mathbb{E}[D|Z=z_{k-1}]}\Bigg\}\\
=&\sum_{k=1}^{K}\mu_k^{IV}\times\frac{\mathbb{E}[T|Z=z_k]-\mathbb{E}[T|Z=z_{k-1}]}{\mathbb{E}[D|Z=z_k]-\mathbb{E}[D|Z=z_{k-1}]}.
\end{aligned} \tag{A18}$$

It is clear that the right hand side is the same for all k. By denoting $\xi=\frac{\mu_k^{IV}}{\mu_k^{Mis}}$, we can obtain $\beta^{Mis}\xi=\beta^{IV}$. In addition, by similar proof of Theorem 1 in Angrist and Imbens (1995), replacing Y with T leads to

$$\frac{\mathbb{E}[T|Z=z_k]-\mathbb{E}[T|Z=z_{k-1}]}{\mathbb{E}[D|Z=z_k]-\mathbb{E}[D|Z=z_{k-1}]}=\sum_{j=1}^{J}\omega_j\mathbb{E}[T_j-T_{j-1}|D_k\geq j>D_{k-1}]. \tag{A19}$$

A.4.3 Proof of Theorem A.2

Let us first consider the numerator of $\lambda^{(d)}$:

$$YT^{(d)} = Z\sum_{j=0}^{J} 1[D_1 = j]Y_j T_j^{(d)} + (1-Z)\sum_{j=0}^{J} 1[D_0 = j]Y_j T_j^{(d)}. \tag{A20}$$

Applying the arguments in the proof of Theorem 1 in Angrist and Imbens (1995), we can get

$$\begin{aligned}
&\mathbb{E}[YT^{(d)}|Z=1] - \mathbb{E}[YT^{(d)}|Z=0] \\
=&\sum_{j=1}^{J} \mathbb{E}[Y_j T_j^{(d)} - Y_{j-1}T_{j-1}^{(d)}|D_1 \geq j > D_0]\Pr(D_1 \geq j > D_0).
\end{aligned} \tag{A21}$$

Because by Assumption A.3, $(D_1 \geq j > D_0)$ is equivalent to $(D_1 = j, D_0 = j-1)$ and $\mathbb{E}[Y|T^{(d)}, D, D_1 \geq j > D_0] = \mathbb{E}[Y|D, D_1 \geq j > D_0]$ for all d, we have

$$\begin{aligned}
\mathbb{E}[Y_j T_j^{(d)}|D_1 \geq j > D_0] &= \mathbb{E}[Y_j T_j^{(d)}|D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[Y_j T_j^{(d)}|Z=1, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[Y_j T_j^{(d)}|D=j, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[YT^{(d)}|D=j, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}\left\{T^{(d)}\mathbb{E}[Y|T^{(d)}, D=j, D_1 = j, D_0 = j-1]|D=j, D_1 = j, D_0 = j-1\right\} \\
&= \mathbb{E}[T_j^{(d)}|D=j, D_1 = j, D_0 = j-1]\mathbb{E}[Y_j|D=j, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[T_j^{(d)}|D_1 = j, D_0 = j-1]\mathbb{E}[Y_j|D_1 = j, D_0 = j-1] \\
&= p_{j,j}^{(d)}\mathbb{E}[Y_j|D_1 \geq j > D_0].
\end{aligned}$$

Similarly, $\mathbb{E}[Y_{j-1}T_{j-1}^{(d)}|D_1 \geq j > D_0] = \mathbb{E}[Y_{j-1}T_{j-1}^{(d)}|Z=0, D_1 = j, D_0 = j-1] = p_{j-1,j}^{(d)}\mathbb{E}[Y_{j-1}|D_1 \geq j > D_0]$. The expression of the denominator can be easily obtained via replacing Y with one. Thus,

$$\begin{aligned}
\lambda^{(d)} &= \sum_{j=1}^{J} \frac{\mathbb{E}[T_j^{(d)}Y_j - T_{j-1}^{(d)}Y_{j-1}|D_1 \geq j > D_0]\Pr(D_1 \geq j > D_0)}{\sum_{k=1}^{J}\mathbb{E}[T_k^{(d)} - T_{k-1}^{(d)}|D_1 \geq k > D_0]\Pr(D_1 \geq k > D_0)} \\
&= \sum_{j=1}^{J} \frac{\Pr(D_1 \geq j > D_0)}{\sum_{k=1}^{J}(p_{k,k}^{(d)} - p_{k-1,k}^{(d)})\Pr(D_1 \geq k > D_0)}\mathbb{E}[p_{j,j}^{(d)}Y_j - p_{j-1,j}^{(d)}Y_{j-1}|D_1 \geq j > D_0].
\end{aligned}$$

Lemma A.1. *Under Assumption A.3, we have*

$$\mathbb{E}[T_j^{(d)}|D_1 \geq j > D_0] = \mathbb{E}[T_j^{(d)}|D_1 \geq j+1 > D_0] = \mathbb{E}[T^{(d)}|D=j], \ \textit{for}\ j = 1, 2, ..., J-1.$$

In addition, $\mathbb{E}[T_0^{(d)}|D_1 \geq 1 > D_0] = \mathbb{E}[T_0^{(d)}|D_1 = D_0 = 0]$.

A.4.4 Proof of Lemma A.1

Given Assumption A.3, $(D_1 \geq j > D_0) = (D_1 = j, D_0 = j-1)$. Then, for $j = 1, 2, ..., J-1$,

$$\begin{aligned}
\mathbb{E}[T_j^{(d)}|D_1 \geq j > D_0] &= \mathbb{E}[T_j^{(d)}|D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[T_j^{(d)}|Z = 1, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[T_j^{(d)}|D = j, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[T^{(d)}|D = j, D_1 = j, D_0 = j-1] \\
&= \mathbb{E}[T^{(d)}|D = j], \qquad \text{(A22)}
\end{aligned}$$

where the second equality in (A22) is due to the independence of Z in Assumption A.2, and the third equality is because that D is solely determined by Z given D_1, D_0. In addition, the last equality follows from Assumption A.3 (iii). Similarly, we can show that for $j = 1, 2, ..., J-1$,

$$\mathbb{E}[T_j^{(d)}|D_1 \geq j+1 > D_0] = \mathbb{E}[T_j^{(d)}|Z = 0, D_1 = j+1, D_0 = j] = \mathbb{E}[T^{(d)}|D = j]. \qquad \text{(A23)}$$

The same arguments can show $\mathbb{E}[T_0^{(d)}|D_1 \geq 1 > D_0] = \mathbb{E}[T_0^{(d)}|D_1 = D_0 = 0]$.

A.4.5 Proof of Theorem A.3

Without loss of generality, assume $\Pr(D_1 \geq j > D_0) \neq 0$ for all $j = 1, ..., J$. For $d = 0, 1, ..., J$,

$$\lambda^{(d)} = \sum_{j=1}^{J} \frac{\Pr(D_1 \geq j > D_0)}{\sum_{i=1}^{J}(p_{i,i}^{(d)} - p_{i-1,i}^{(d)})\Pr(D_1 \geq i > D_0)} \mathbb{E}[p_{j,j}^{(d)} Y_j - p_{j-1,j}^{(d)} Y_{j-1}|D_1 \geq j > D_0].$$

For its denominator, it can be rewritten as

$$\begin{aligned}
\sum_{i=1}^{J}(p_{i,i}^{(d)} - p_{i-1,i}^{(d)})\Pr(D_1 \geq i > D_0) = &-p_{0,1}^{(d)}\Pr(D_1 \geq 1 > D_0) + p_{J,J}^{(d)}\Pr(D_1 \geq J > D_0) \\
&+ \sum_{i=1}^{J-1}\Big[p_{i,i}^{(d)}\Pr(D_1 \geq i > D_0) - p_{i,i+1}^{(d)}\Pr(D_1 \geq i+1 > D_0)\Big] \\
= &-p_{0,0}^{(d)}\Pr(D_1 \geq 1 > D_0) + p_{J,J}^{(d)}\Pr(D_1 \geq J > D_0) \\
&+ \sum_{i=1}^{J-1} p_{i,i}^{(d)}\left[\Pr(D_1 \geq i > D_0) - \Pr(D_1 \geq i+1 > D_0)\right], \qquad \text{(A24)}
\end{aligned}$$

where the second equality is by Lemma A.1. Similarly, the numerator of $\lambda^{(d)}$ can be represented as

$$
\begin{aligned}
&\sum_{j=1}^{J}\mathbb{E}[p_{j,j}^{(d)}Y_j - p_{j-1,j}^{(d)}Y_{j-1}|D_1 \geq j > D_0]\Pr(D_1 \geq j > D_0)\\
=&-p_{0,1}^{(d)}\mathbb{E}[Y_0|D_1 \geq 1 > D_0]\Pr(D_1 \geq 1 > D_0) + p_{J,J}^{(d)}\mathbb{E}[Y_J|D_1 \geq J > D_0]\Pr(D_1 \geq J > D_0)\\
&+\sum_{j=1}^{J-1}\Big[p_{j,j}^{(d)}\mathbb{E}[Y_j|D_1 \geq j > D_0]\Pr(D_1 \geq j > D_0) - p_{j,j+1}^{(d)}\mathbb{E}[Y_j|D_1 \geq j+1 > D_0]\Pr(D_1 \geq j+1 > D_0)\Big]\\
=&-p_{0,0}^{(d)}\mathbb{E}[Y_0|D_1 \geq 1 > D_0]\Pr(D_1 \geq 1 > D_0) + p_{J,J}^{(d)}\mathbb{E}[Y_J|D_1 \geq J > D_0]\Pr(D_1 \geq J > D_0)\\
&+\sum_{j=1}^{J-1}p_{j,j}^{(d)}\big[\mathbb{E}[Y_j|D_1 \geq j > D_0]\Pr(D_1 \geq j > D_0) - \mathbb{E}[Y_j|D_1 \geq j+1 > D_0]\Pr(D_1 \geq j+1 > D_0)\big],
\end{aligned}
\tag{A25}
$$

where the second equality is by Lemma A.1.

(i) Under condition (a), it yields from (A24) and (A25) that,

$$
\lambda^{(J)} = \frac{\Pr(D_1 \geq J > D_0)}{p_{J,J}^{(J)}\Pr(D_1 \geq J > D_0)}\mathbb{E}[p_{J,J}^{(J)}Y_J|D_1 \geq J > D_0] = \mathbb{E}[Y_J|D_1 \geq J > D_0]. \tag{A26}
$$

(ii) By condition (b), (A24) and (A25), we have

$$
\lambda^{(0)} = \frac{\Pr(D_1 \geq 1 > D_0)}{p_{0,0}^{(0)}\Pr(D_1 \geq 1 > D_0)}\mathbb{E}[p_{0,0}^{(0)}Y_0|D_1 \geq 1 > D_0] = \mathbb{E}[Y_0|D_1 \geq 1 > D_0]. \tag{A27}
$$

(iii) Given conditions in (c), it is easy to see

$$
\begin{aligned}
\lambda^{(d)} &= \frac{\Pr(D_1 \geq d > D_0)\mathbb{E}[Y_d|D_1 \geq d > D_0] - \Pr(D_1 \geq d+1 > D_0)\mathbb{E}[Y_d|D_1 \geq d+1 > D_0]}{\Pr(D_1 \geq d > D_0) - \Pr(D_1 \geq d+1 > D_0)}\\
&= \frac{\Pr(D_1 \geq d > D_0) - \Pr(D_1 \geq d+1 > D_0)}{\Pr(D_1 \geq d > D_0) - \Pr(D_1 \geq d+1 > D_0)}\mathbb{E}[Y_d|D_1 \geq d > D_0]\\
&= \mathbb{E}[Y_d|D_1 \geq d > D_0],
\end{aligned}
\tag{A28}
$$

where the second equality is due to $\mathbb{E}[Y_d|D_1 \geq d+1 > D_0] = \mathbb{E}[Y_d|D_1 \geq d > D_0]$ by Assumption A.4.

A.4.6 Proof of Corollary A.2

Given (A26), (A27) and (A28), by Assumption A.4, for $j, j' = 0, 1, .., J$ and $j' < j$, if $j' = 0$

$$
\begin{aligned}
\lambda^{(j)} - \lambda^{(j')} &= \mathbb{E}[Y_j|D_1 \geq j > D_0] - \mathbb{E}[Y_0|D_1 \geq 1 > D_0]\\
&= \mathbb{E}[Y_j|D_1 \geq j > D_0] - \mathbb{E}[Y_0|D_1 \geq j > D_0]\\
&= \mathbb{E}[Y_j - Y_{j'}|D_1 \geq j > D_0].
\end{aligned}
$$

Otherwise,

$$
\begin{aligned}
\lambda^{(j)} - \lambda^{(j')} &= \mathbb{E}[Y_j | D_1 \geq j > D_0] - \mathbb{E}[Y_{j'} | D_1 \geq j' > D_0] \\
&= \mathbb{E}[Y_j | D_1 \geq j > D_0] - \mathbb{E}[Y_{j'} | D_1 \geq j > D_0] \\
&= \mathbb{E}[Y_j - Y_{j'} | D_1 \geq j > D_0].
\end{aligned}
$$

Lemma A.2. *Consider $J = 2$. Under Assumptions A.1 to A.3, suppose there exists a $T^{(1)}$ as in Theorem A.3 (iii). Then, $\mathbb{E}[T^{(1)}|Z = 1] - \mathbb{E}[T^{(1)}|Z = 0]$ signs $Pr(D_1 \geq 1 > D_0) - Pr(D_1 \geq 2 > D_0)$.*

A.4.7 Proof of Lemma A.2

By replacing Y with one in (A21) in the proof of Theorem A.2 and $J = 2$, we know that

$$
\begin{aligned}
&\mathbb{E}[T^{(1)}|Z = 1] - \mathbb{E}[T^{(1)}|Z = 0] \\
=&\mathbb{E}[T_1^{(1)} - T_0^{(1)} | D_1 \geq 1 > D_0]\Pr(D_1 \geq 1 > D_0) + \mathbb{E}[T_2^{(1)} - T_1^{(1)} | D_1 \geq 2 > D_0]\Pr(D_1 \geq 2 > D_0).
\end{aligned}
$$

In addition, $T^{(1)}$ satisfying conditions in Theorem A.3 (iii) implies that

$$
\begin{aligned}
\mathbb{E}[T_1^{(1)} | D_1 \geq 1 > D_0] &= \mathbb{E}[T_1^{(1)} | D_1 \geq 2 > D_0] = p_{1,1}^{(1)}, \\
\mathbb{E}[T_0^{(1)} | D_1 \geq 1 > D_0] &= \mathbb{E}[T_2^{(1)} | D_1 \geq 2 > D_0] = 0.
\end{aligned}
$$

Thus, we get $\mathbb{E}[T^{(1)}|Z = 1] - \mathbb{E}[T^{(1)}|Z = 0] = p_{1,1}^{(1)}\Big[\Pr(D_1 \geq 1 > D_0) - \Pr(D_1 \geq 2 > D_0)\Big]$, and $0 < p_{1,1}^{(1)} \leq 1$ fulfils the proof.

A.4.8 Proof of Corollary A.3

Under $J = 2$, it follows from the proof of (A28) that

$$
\lambda^{(1)} = \frac{\Pr(D_1 \geq 1 > D_0)\mathbb{E}[Y_1 | D_1 \geq 1 > D_0] - \Pr(D_1 \geq 2 > D_0)\mathbb{E}[Y_1 | D_1 \geq 2 > D_0]}{\Pr(D_1 \geq 1 > D_0) - \Pr(D_1 \geq 2 > D_0)}.
$$

Given $\mathbb{E}[T^{(1)}|Z = 1] - \mathbb{E}[T^{(1)}|Z = 0] > 0$, we have that $\Pr(D_1 \geq 1 > D_0) - \Pr(D_1 \geq 2 > D_0) > 0$ from Lemma A.2. Then, Assumption A.5 implies

$$
\begin{aligned}
\lambda^{(1)} &\leq \frac{\Pr(D_1 \geq 1 > D_0)\mathbb{E}[Y_1 | D_1 \geq 1 > D_0] - \Pr(D_1 \geq 2 > D_0)\mathbb{E}[Y_1 | D_1 \geq 1 > D_0]}{\Pr(D_1 \geq 1 > D_0) - \Pr(D_1 \geq 2 > D_0)} \\
&= \mathbb{E}[Y_1 | D_1 \geq 1 > D_0]. \qquad \text{(A29)}
\end{aligned}
$$

(ii) For $T^{(1)}$ to $T^{(2)}$ described in Theorem A.3, we have $\lambda^{(0)} = \mathbb{E}[Y_0 | D_1 \geq 1 > D_0]$ and $\lambda^{(2)} =$

$\mathbb{E}[Y_2|D_1 \geq 2 > D_0]$. Then, from Assumption A.5 and (A29) we have

$$\begin{aligned}
\lambda^{(1)} - \lambda^{(0)} &\leq \mathbb{E}[Y_1|D_1 \geq 1 > D_0] - \mathbb{E}[Y_0|D_1 \geq 1 > D_0], \\
\lambda^{(2)} - \lambda^{(1)} &\geq \mathbb{E}[Y_2|D_1 \geq 2 > D_0] - \mathbb{E}[Y_1|D_1 \geq 1 > D_0] \geq \mathbb{E}[Y_2 - Y_1|D_1 \geq 2 > D_0], \\
\lambda^{(2)} - \lambda^{(1)} &\geq \mathbb{E}[Y_2|D_1 \geq 2 > D_0] - \mathbb{E}[Y_1|D_1 \geq 1 > D_0] \geq \mathbb{E}[Y_2 - Y_1|D_1 \geq 1 > D_0], \\
\lambda^{(2)} - \lambda^{(0)} &= \mathbb{E}[Y_2|D_1 \geq 2 > D_0] - \mathbb{E}[Y_0|D_1 \geq 1 > D_0] \geq \mathbb{E}[Y_2 - Y_0|D_1 \geq 2 > D_0], \\
\lambda^{(2)} - \lambda^{(0)} &= \mathbb{E}[Y_2|D_1 \geq 2 > D_0] - \mathbb{E}[Y_0|D_1 \geq 1 > D_0] \geq \mathbb{E}[Y_2 - Y_0|D_1 \geq 1 > D_0].
\end{aligned}$$

A.4.9 Proof of Theorem A.5

Suppose that $\hat{\theta}_n$ solves a $d_\theta \times 1$ vector $\sum_{i=1}^n \psi(W_i;\theta) = 0$ where d_θ is the dimension of θ. Assume that there is a unique solution to $\mathbb{E}[\psi(W;\theta)] = 0$ and $\partial \mathbb{E}[\psi(W;\theta)]/\partial \theta'$ is of full rank. Denote $\epsilon_i^{(d)}(\eta) = Y_i T_i^{(d)} - \gamma^{(d)} - \lambda^{(d)} T_i^{(d)}$. Let a $d_\eta \times 1$ vector $\tilde{h}(W_i;\eta)$ be

$$\tilde{h}(W_i;\eta) = \begin{bmatrix} \psi(W_i;\theta) \\ Y_i T_i^{(0)} - \gamma^{(0)} - \lambda^{(0)} T_i^{(0)} \\ \vdots \\ Y_i T_i^{(J)} - \gamma^{(J)} - \lambda^{(J)} T_i^{(J)} \\ Z_i(Y_i T_i^{(0)} - \gamma^{(0)} - \lambda^{(0)} T_i^{(0)}) \\ \vdots \\ Z_i(Y_i T_i^{(J)} - \gamma^{(J)} - \lambda^{(J)} T_i^{(J)}) \end{bmatrix} = \begin{bmatrix} \psi(W_i;\theta) \\ \epsilon_i^{(0)}(\eta) \\ \vdots \\ \epsilon_i^{(J)}(\eta) \\ Z_i \epsilon_i^{(0)}(\eta) \\ \vdots \\ Z_i \epsilon_i^{(J)}(\eta) \end{bmatrix}.$$

We have that $\mathbb{E}\left[\tilde{h}(W_i;\eta)\right] = 0$ holds, where the last $J+1$ moment conditions come from the definition of $\lambda^{(d)}$. Then $\hat{\eta}_n = (\hat{\theta}_n', \hat{\gamma}_n^{(0)}, ..., \hat{\gamma}_n^{(J)}, \hat{\lambda}_n^{(0)}, ..., \hat{\lambda}_n^{(J)})$ solves $\frac{1}{n}\sum_{i=1}^n \tilde{h}(W_i;\eta) = 0$. Denote a $d_\eta \times d_\eta$ matrix $\tilde{H} = \mathbb{E}\left[\frac{\partial \tilde{h}(W_i;\eta^0)}{\partial \eta'}\right]$. Because $\mathrm{Cov}(g(Z_i), T_i^j) \neq 0$, $\tilde{H}$ is invertible. By the mean value theorem, we get

$$0 = \frac{1}{n}\sum_{i=1}^n \tilde{h}(W_i;\hat{\eta}_n) = \frac{1}{n}\sum_{i=1}^n \tilde{h}(W_i;\eta^0) + \frac{1}{n}\sum_{i=1}^n \frac{\partial \tilde{h}(W_i;\tilde{\eta}_n)}{\partial \eta'}(\hat{\eta}_n - \eta^0),$$

where $\tilde{\eta}_n$ is element-by-element between $\hat{\eta}_n$ and η^0. For large enough sample size, we know that $\frac{1}{n}\sum_{i=1}^n \frac{\partial \tilde{h}(W_i;\tilde{\eta}_n)}{\partial \eta'}$ is invertible. Therefore,

$$\sqrt{n}(\hat{\eta}_n - \eta^0) = -\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial \tilde{h}(W_i;\tilde{\eta}_n)}{\partial \eta'}\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{h}(W_i;\eta^0),$$

Denote $\tilde{\Sigma}$ =Var$[\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{h}(W_i;\eta^0)]$. Under standard regularity conditions, we can show that

$$\sqrt{n}(\hat{\eta}_n - \eta^0) \xrightarrow{d} \mathcal{N}(0, \tilde{H}^{-1}\tilde{\Sigma}\tilde{H}^{-1'}).$$

Thus, by the delta method, the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}^0)$ is the lower-right $(J+1)\times(J+1)$ block of $\tilde{H}^{-1}\tilde{\Sigma}\tilde{H}^{-1'}$.

A.5 Returns to Education in the UK: Details

Table A3: Descriptive Statistics

	(1) Mean	(2) S.D.	(3) Min	(4) Max
Key Variables:				
Log(wage)	2.06	0.43	0	4
Qualitication, school transcript [0,1]	0.59	0.49	0	1
Qualitication, self-reported in 1981 [0,1]	0.65	0.48	0	1
Qualitication, self-reported in 1991 [0,1]	0.64	0.48	0	1
Parents' interest in child's education at age 7 [0,1]	0.44	0.50	0	1
Mother's education > Father's education [0,1]	0.22	0.42	0	1
Covariates:				
White [0,1]	0.98	0.13	0	1
Comprehensive school [0,1]	0.49	0.50	0	1
Secondary modern school [0,1]	0.16	0.37	0	1
Grammar school [0,1]	0.11	0.31	0	1
Public school [0,1]	0.06	0.23	0	1
Father's education	7.67	4.62	0	18
Mother's education	7.76	4.39	0	18
Father's age	44.10	12.16	0	73
Mother's age	42.40	9.05	0	62
Professional [0,1]	0.05	0.21	0	1
Intermediate [0,1]	0.15	0.36	0	1
Skilled non-manual [0,1]	0.09	0.28	0	1
Skilled manual [0,1]	0.32	0.47	0	1
Semi'skilled non-manual [0,1]	0.01	0.10	0	1
Semi-skilled manual [0,1]	0.10	0.29	0	1
Unskilled manual [0,1]	0.03	0.16	0	1
Mother is employed [0,1]	0.54	0.50	0	1
Number of siblings	1.72	1.75	0	11
London [0,1]	0.14	0.34	0	1
Wales [0,1]	0.06	0.24	0	1
Scotland [0,1]	0.10	0.31	0	1
Observations	2454			

Notes: The table reports the summary statistics of the sample used in our empirical illustration. We use a version of the dataset constructed by Battistin and Sianesi (2011) and Battistin et al. (2014).