Bounding Program Benefits When Participation is Misreported

Denni Tommasi[1] and Lina Zhang*[2]

[1]University of Bologna, IZA and CDES

[2]University of Amsterdam

October 29, 2021

Abstract

Instrumental variables (IV) are commonly used to estimate treatment effects in case of noncompliance. However, program participation is often misreported in survey data and standard techniques are not sufficient to point identify and consistently estimate the effects of interest. In this paper, we first derive a new link between the true and mismeasured treatment effect. Second, we provide three IV strategies to partially identify the heterogeneous treatment effects when both noncompliance and misreporting are present. Third, we use our new Stata command, `ivbounds`, and obtain novel results of the benefits of participating in the 401(k) pension plan on savings.

JEL Codes: C14, C21, C26, C35, C51.

Keywords: heterogeneous treatment effects, causality, binary treatment, endogenous measurement error, discrete or multiple instruments, weighted average of LATEs, endogeneity, program evaluation.

*Tommasi: Department of Economics, University of Bologna. Piazza Scaravilli 2, Bologna, 40126, Italy. E-mail: `denni.tommasi@unibo.it`. Zhang: Amsterdam School of Economics, University of Amsterdam. Roetersstraat 11, 1018 WB Amsterdam, The Netherlands. E-mail: `l.zhang5@uva.nl`. This is a substantial revision of a paper that circulated with the same title. We would like to thank Isaiah Andrews, Peng Ding, David Frazier, Martin Huber, Kosuke Imai, Zhichao Jiang, Desire Kedagni, Umair Khalil, Arthur Lewbel, Charles Manski, Francesca Molinari, Akanksha Negi, Whitney Newey, Didier Nibbering, Tatsushi Oka, John Pepper, Mervyn Silvapulle, Alessandro Tarozzi, Takuya Ura and Kaspar Wuthrich, participants at the Italian Congress of Econometrics and Empirical Economics, and seminar participants at the University of Bologna for valuable comments. Hai Duong provided excellent research assistance. Denni acknowledges financial support from SEED fund of Monash University. The `ivbounds` Stata command is available from the Statistical Software Components (SSC) repository. See Lin, Tommasi, and Zhang (2021) for explanations on how to use the command. All remaining errors are ours.

1 Introduction

The instrumental variables (IV) method is commonly used to estimate treatment effects in case of noncompliance (Athey and Imbens, 2017). Standard approaches for the identification and inference of causal parameters require that the treatment variable is correctly measured. In program evaluation, misreporting (misclassification) of key variables due to 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" is an increasing problem for social scientists (Meyer et al., 2015, p. 219).[1] Since participation is usually binary, attempting to evaluate the benefits of a program using standard techniques would lead to biased estimates (Kreider, 2010; Millimet, 2011).[2] In this paper, we focus on the average causal effect for compliers (Imbens and Angrist, 1994) and develop an IV method that can be used to measure the benefits of a program when both noncompliance and misreporting of treatment status are present. When the IV is binary, our target parameter is the local average treatment effect (LATE); with discrete or multiple-discrete IVs, it is the weighted average of LATEs (WLATE).[3]

Recent works have developed methods to deal with an endogenous and misclassified binary treatment variable. Identification of our target parameter(s) in case of exogenous (nondifferential) misclassification of the treatment is achieved assuming resurvey data (Battistin et al., 2014), two instrumental variables (Yanagi, 2019), two measurements of the same treatment (Calvi, Lewbel, and Tommasi, 2018) or homogeneous treatment effects (DiTraglia and García-Jimeno, 2019). Sharp bounds on the causal effects are provided by Jiang and Ding (2020). A few papers provide solutions for the general case of endogenous (differential) misclassification of the treatment. Kreider et al. (2012) formalized an approach to partially identify the average treatment effect, which requires auxiliary data. Nguimkeu et al. (2018) achieved point identification of our target parameter(s) under the assumption of homogeneous treatment effects and a strong parametric structure. Ura (2018) considered the heterogeneous treatment effects and provided a partial identification strategy of LATE by using a binary instrument.

This paper considers the problem of endogenous (differential) misclassification in a heterogeneous-treatment-effects framework and extends the above works in several directions. First, we characterize the bias of the mismeasured IV estimand and establish a link between the true and mismeasured treatment effects. This link is mediated by a new parameter, defined in terms of the misclassification probabilities, which can be used to approximate the possible level of bias of the estimated benefits of a program in a binary, discrete or multiple-discrete-instruments setting. Second, we generalize the partial identification result of Ura (2018) to allow the support of the instrument to grow from

[1] Also Meyer and Mittag (2019a,b) and Meyer et al. (2020) documented that item nonresponse rates in all major United States (US) surveys for various transfer programs were large and increasing.

[2] In a measurement error scenario, instrumental variables correct for endogeneity and measurement error of the treatment variable simultaneously. However, measurement error is always nonclassical for a binary treatment because of the negative correlation between the true treatment status and the error term.

[3] We provide results valid for both parameters; hence, to avoid any confusion, throughout the text, one should carefully note when we refer to LATE or WLATE.

binary to discrete to multiple-discrete, which is likely to be useful for practitioners.[4] Different to a binary IV setting, we gain identification power by using multiple total variation distances, which capture the distributional effect of the instrument(s) on observable variables. Two strategies yield the bounds for the WLATE: first, via the bounds of LATEs; second, via the bounds of a new estimand that we introduce, the local average of treatment misclassifications (LATMs), which captures the average of the misclassification probabilities for compliers. We also provide sufficient conditions under which the bounds of LATEs and LATMs are sharp. Finally, we formalize a third partial identification strategy to combine external information about the extent of misclassification to obtain tighter bounds or a point estimate of LATE or WLATE. Our strategy improves upon others that rely on auxiliary data because (i) external information are only used to narrow our bounds (e.g., Kreider and Pepper, 2007; Kreider et al., 2012), (ii) we do not need to assume an exogenous treatment (e.g., Imai and Yamamoto, 2010; Battistin and Sianesi, 2011), nor to observe who has missing treatment (e.g., Molinari, 2010), and (iii) treatment misclassification in our case can also be endogenous (differential).

Our approach, using external information, is based on the observation that an increasing number of researchers can link administrative records of program receipts to standard survey data and report the extent of misreporting. Hence, for a wide range of programs, some information about treatment misclassification probabilities is available even to researchers who do not have access to administrative data.[5] Alternatively, one could also retrieve these data using small validation studies or repeated measurements of the same individual. In cases where the practitioner has a good approximation of the misclassification probabilities in the survey, our proposed bounds, using external information, degenerate to a point estimate. Even if these misclassification probabilities are approximate, our approach provides the least-biased point estimate, which is much closer to the truth than the naive IV estimate, which ignores treatment misclassification. Hence, our estimator has a bias reduction property.

Regarding inference, for each partial identification strategy, we construct confidence intervals for the bounds with uniform and asymptotic size control. They are built based on a two-step bootstrap procedure following the work by Chernozhukov, Chetverikov, and Kato (2019). We demonstrate the finite sample properties of the proposed inference methods through a series of Monte Carlo simulations.

We extend these results in two main directions. First, to further improve the bounds, we show the benefits of having multiple treatment indicators or repeated measures of the same treatment.

[4]Instrumental variables with a support larger than two values are commonly applied in empirical research. Moreover, allowing also the case of multiple-discrete-instruments is likely to be useful to practitioners because, according to Mogstad et al. (2020b), more than half of the empirical papers using instrumental variables and published in top journals in the last 20 years "make use of multiple instrumental variables for a single treatment."

[5]For example, Meyer et al. (2020) used administrative data from the food stamp program, Supplemental Nutrition Assistance Program (SNAP), participation and link them to the ACS, the CPS, and the Survey of Income and Program Participation (SIPP). They found that the extent of true food-stamp-recipient households that did not report receipt (false negative) in these surveys was 35%, 23%, and 50%, respectively. Misclassification probabilities of other US government transfer programs were reported in Meyer and Mittag (2019a,b). In a similar vein, Dushi and Iams (2010) merged tax records information with the SIPP data. They found that over 17% of participants in defined contribution (DC) pension plans self-reported as nonparticipants (false negative), and almost 10% of nonparticipants self-reported as participants (false positive). As more researchers gain access to linked administrative data, similar information about misclassification probabilities can be obtained for other countries and programs.

Importantly, we do not restrict the dependence among the treatment measures; thus, these variables can be considered endogenous. Second, since the instrument(s) may be confounded without conditioning on some covariates, or treatment effects may be heterogeneous across the population characterized by different attributes, we provide a strategy to use the propensity score index to include covariates in the analysis.

Overall, this paper shows that researchers measuring the benefits of a program can obtain bounds of LATE or WLATE if the binary treatment is misclassified. To do so, they can use the `ivbounds` Stata command, which is available from the SSC repository (see Lin, Tommasi, and Zhang (2021) for how to use this command). We use this new command to reassess the benefits of participating in the 401(k) pension plan on savings in the US. The empirical results are twofold. First, we find that the benefits of the plan estimated with a naive approach are biased (overestimated) approximately by 37%. Second, using our preferred strategy in the presence of external information, we obtain bounds of the true benefits that can be up to 36% narrower in width than comparable results in the literature (Ura, 2018).

Our method has at least three applications. First, it can be used as the leading identification strategy in any setting where the practitioner knows that the endogenous binary treatment is not well measured. Second, it can be used as the leading robustness check if misreporting is only suspected. Third, it can assess the sensitivity of program benefits under different assumptions of the misclassification probabilities. Although our method is primarily motivated by (and directed to practitioners in) the program evaluation literature, it is not limited to applications within this context. It can be applied to any setting where the endogenous binary treatment is contaminated by endogenous measurement error, and the researcher considers LATE or WLATE the relevant parameter(s) for evaluating the policy change.

This paper relates to a long-standing tradition in the program evaluation literature concerned with using IVs to infer causal parameters when the treatment is misclassified. Papers empirically documenting substantial misclassification error in the treatment include Bollinger (1996), Angrist and Krueger (1999), Kane et al. (1999), Bound et al. (2001), Card (2001), Black et al. (2003) and Hernandez et al. (2007). A few earlier papers considered techniques for dealing with treatment misclassification. In the context of homogeneous treatment effects of a mismeasured binary regressor, Aigner (1973), Bollinger (1996), Kane et al. (1999) and Black et al. (2000) used IV techniques to estimate the effects of an exogenous treatment. Under more general conditions, Klepper (1988) provided bounds on average treatment effects with multiple misclassified treatments. In the context of heterogeneous treatment effects, Mahajan (2006), Lewbel (2007) and Hu (2008) also used instruments to point-identify average treatment effects in the case of an exogenous and mismeasured binary (or discrete) treatment indicator.

Our estimation problem has the standard LATE structure; that is, a binary treatment is correlated with a binary, discrete or multiple-discrete instrument(s). The LATE or WLATE, "may be the only relevant information that is credibly identifiable under weak conditions" (Imbens, 2014) and it "is of intrinsic interest when the instrument itself represents an intervention, like a policy change or a

randomized control trial" (Mogstad et al., 2018).[6] In situations where the causal effect for compliers might not represent the effect of interest, the LATE or WLATE could be used to extrapolate to causal effects for individuals other than those affected by the instrument available. Recently, Heckman et al. (2006), Brinch et al. (2017), Vuong and Xu (2017), Chen et al. (2017) and Mogstad et al. (2018, 2020a) among others, have provided methods along this line. In particular, Mogstad et al. (2018, 2020a) consider multiple IVs and propose a partial identification method for policy-relevant treatment effects (PRTE), via the identifiable LATE or WLATE and the marginal treatment effect (MTE) framework. Note that the papers above all rely on the treatment variable to be correctly measured. Thus, our approach can be used as the first step when the targets are the PRTEs and the treatment suffers from misclassification. In addition, Acerenza et al. (2021) and Possebom (2021) conduct bounding analysis for the MTEs and PRTEs with misclassified treatment. This paper can be viewed as an alternative and complement to them, because our method provides the bounds for LATEs and WLATE, but also serves as the starting point of studying PRTEs.

The remainder of the paper is organized as follows. Section 2 presents our framework and the main results. Section 3 develops an inference procedure for the parameters of interest. Section 4 discusses extensions, how to use our partial identification strategies in practice, simulations and an application using our new `ivbounds` command. Concluding remarks are in Section 5. Proofs and additional material are in the Appendix.

2 Theoretical Framework

This section proceeds in four acts. First, we describe our theoretical framework and show the limitations of the standard IV approach when the treatment variable is contaminated by measurement error. This leads to a simple relationship between the true and mismeasured effect, which can be captured by a summary statistic of the misclassification probabilities. Second, we present tractable outer sets of the LATEs and the LATMs, and provide sufficient conditions for the sharp identified sets. Third, we outline different strategies to partially identify the parameter(s) of interest based on different sources of information. Fourth, we show how to use external information regarding the extent of the misclassification probabilities to obtain tighter bounds or, under certain conditions, to obtain a point estimate of the effect.

2.1 Setup and Limitations of the Standard IV Approach

We introduce some notation which will be used throughout the text. For the moment, we derive our results without conditioning on covariates. Later, we extend the partial identification and inference procedure to accommodate a generic vector X of observable characteristics.

[6] This is an estimand subject to a heated debate and many distinguished researcher, both in economics and statistics, have contributed to this debate (e.g. Heckman et al., 2006; Manski, 2007; Deaton, 2010; Heckman and Urzúa, 2010; Imbens, 2010). More recently, considerable effort have been put into developing connections between instrumental variables and structural estimators (e.g. Kline and Walters, 2019).

True effect. Let D be the true binary treatment variable that affects the outcome of interest. D is *not* observed and its effects cannot be consistently estimated. Let Z be a $h \times 1$ vector of discrete instruments. Let $\Omega_Z = \{z_0, z_1, ..., z_K\}$ be the support of Z with $z_k \in \mathbb{R}^h$. Denote $D_k \in \{0, 1\}$, for $k = 0, 1, ..., K$, as the potential treatment corresponding to possible realization z_k of Z. By definition,

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

where $1[\cdot]$ denotes the indicator function. Denote $\Pr(z_k) = \mathbb{E}(D|Z = z_k)$ the propensity score. Let Y be an observed outcome of interest and let Y_d be the potential outcome with $d \in \{0, 1\}$ for possible realization of D. Denote by $\Omega_Y \subset \mathbb{R}$ the support of Y, Y_1 and Y_0. Then,

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

A common way to exploit multiple instruments is to introduce a scalar function $g : \Omega_Z \mapsto \mathbb{R}$, for example, $g(z)$ can be an estimate of $\Pr(z)$ or other known functions.[7]

Assumption 2.1. *Y, D and Z satisfy the standard Imbens and Angrist (1994) assumptions:*

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

(ii) (Unconfoundedness) $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K)$ and $Pr(z) = \mathbb{E}(D|Z = z)$ for $z \in \Omega_Z$ is a nontrivial function of z; $0 < \pi_k = Pr(Z = z_k) < 1$, $k = 0, 1, ..., K$;

(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 $g(z_l) \geq g(z_w)$.

The monotonicity assumption is satisfied if no subjects respond in the opposite way to their instrument assignment status (no defiers). When there is more than one instrument, Mogstad et al. (2020a,b) point out that the monotonicity assumption can only be satisfied if the treatment choice behavior is homogeneous, which means that all individuals respond to the same shift in the instrument value in the same direction.[8] Throughout the paper, we denote compliers ($D_{k-1} = 0, D_k = 1$) as C_k. If D was observed, under Assumption 2.1, the Imbens and Angrist (1994)'s weighted average of local average treatment effect (WLATE) would be identified by the instrumental

[7] If Z is a scalar binary or discrete instrument satisfying monotonicity assumption, we can simply set $g(z) = z$. If Z includes multiple instruments, $g(z)$ can be set as, for example, an estimate of $E[Y|Z = z]$ or of $\Pr(T = 1|Z = z)$ for $z \in \Omega_Z$, where T represents a proxy of the true treatment and will be introduced later.

[8] Mogstad et al. (2020b) refer to Assumption 2.1 (iv) as "IA Monotonicity" and distinguish it from the "Actual Monotonicity", meaning that if $z_l \geq z_w$ component-wise then $D_l \geq D_w$ for all individuals. It is unclear whether the analysis in this paper can be extended using only the "Actual Monotonicity" assumption, unless all elements in Ω_Z can be ranked by the component-wise "$\leq$" or "$\geq$". Słoczyński (2020) also studies the relaxation of IA Monotonicity in the LATE setting, but focusing on correctly measured treatment.

variables estimand:

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

where $\gamma_k^{IV} := \frac{\Pr(C_k)\sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}{\sum_{m=1}^{K} \Pr(C_m)\sum_{l=m}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}$ are the weights, $\Pr(C_k) = \Pr(z_k) - \Pr(z_{k-1})$ and $\alpha_{k,k-1} := \mathbb{E}[Y_1 - Y_0 | C_k]$ is the local average treatment effect (LATE) for each subgroup of compliers C_k. The weights $\{\gamma_k^{IV}\}_{k=1}^{K}$ are nonnegative and $\sum_{k=1}^{K} \gamma_k^{IV} = 1$. However, since in practice we do not observe D, we cannot implement this standard approach.

Mismeasured effect. Instead of D, suppose we can observe a binary treatment indicator T, which could be a proxy for D, or could correspond to reported values of D that are misclassified for some observations. This means that T does not equal D for some individuals because of misclassification errors. Define $T_d \in \{0, 1\}$ as the potential observed treatment with $d \in \{0, 1\}$ for possible realization of D. Then by definition:

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

The variables T_0 and T_1 can be interpreted as indicators of whether treatment is correctly measured or not. That is, if $T_0 = 0$ and $T_1 = 1$, then the true treatment D is not misclassified. This shows that, in a binary treatment setting, there are two possible measurement or misclassification errors: if $T_0 = 1$, then a true $D = 0$ is misclassified as treated (false positive), and if $T_1 = 0$, then a true $D = 1$ is misclassified as untreated (false negative).

Assumption 2.2. *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 2.2-(i) combines the LATE unconfoundedness assumption that $Z \perp \left(Y_1, Y_0, \{D_k\}_{k=0}^{K}\right)$ with the assumption that the instruments are also independent of the potential measurement errors, and hence of (T_1, T_0). Random assignment of Z would be sufficient to make 2.2-(i) hold. Assumption 2.2-(ii) is used to ensure that the identifiable estimand from observable data is well-defined and it is a minimal relevance condition. It says that, although T suffers from potential misclassification error, it still provides some information regarding D. We do not restrict the extent of misclassification here and will revisit this matter in the next section.

Using the proxy T in place of D leads to the identification of a new parameter, which is useful to characterize. Let $p_{d,k} = \mathbb{E}(T_d \mid C_k)$ for $d \in \{0, 1\}$ and $k = 1, 2, ..., K$. By definition, $p_{1,k}$ is the probability that compliers C_k would have their treatment correctly observed if they were treated. That is, $p_{1,k}$ is the probability that the compliers would have $T = 1$ if they were assigned $D = 1$. In contrast, $p_{0,k}$ is the probability that compliers C_k would have their treatment incorrectly observed

if they were untreated. That is, $p_{0,k}$ is the probability that the compliers would have $T = 1$ if they were assigned $D = 0$.

Theorem 2.1. *Let Assumption 2.1 and 2.2 hold. Then:*

$$\alpha^{Mis} := \frac{Cov(Y, g(Z))}{Cov(T, g(Z))} = \frac{\mathbb{E}[(Y - \mathbb{E}(Y))(g(Z) - \mathbb{E}[g(Z)])]}{\mathbb{E}[(T - \mathbb{E}(T))(g(Z) - \mathbb{E}[g(Z)])]} = \sum_{k=1}^{K} \gamma_k^{Mis} \alpha_{k,k-1}, \tag{2}$$

where $\gamma_k^{Mis} := \frac{Pr(C_k) \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}{\sum_{m=1}^{K} (p_{1,m} - p_{0,m}) Pr(C_m) \sum_{l=m}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z)])}$ *are the weights for each subgroup of compliers* C_k.

Proof of Theorem 2.1. See Appendix A.1. □

Intuitively, α^{Mis} denotes the new estimand that can be identified if we ignore the misclassification error and use a mismeasured treatment indicator T in place of the true treatment D. If the treatment indicator contains sufficient information about the true treatment, that is, if the summation of the false positive and false negative rates is less than one for all complier groups, then we have $0 \leq p_{1,k} - p_{0,k} \leq 1$ and γ_k^{Mis} is positive for all k. However, the summation $\sum_{l=k}^{K} \gamma_k^{Mis}$ is likely to be greater than one because each γ_k^{Mis} is inflated by the misclassification error in the denominator. In contrast, if the misclassification is severe, it is possible that $p_{1,k} - p_{0,k} < 0$ and weight γ_k^{Mis} becomes negative for some k. In this case, α^{Mis} may be negative (positive) even if treatment effects are positive (negative) for everyone in the population. Clearly, $\alpha^{Mis} \neq \alpha^{IV}$ because $\gamma_k^{Mis} \neq \gamma_k^{IV}$. A sufficient condition for $\alpha^{Mis} = \alpha^{IV}$ is that $p_{1,k} = 1$ and $p_{0,k} = 0$ for all k (no misclassification error).

Relationship between the true and mismeasured effect. There is a simple relationship between α^{IV} and α^{Mis} which can be captured by a summary statistic of the misclassification probabilities.

Corollary 2.1. *Let Assumption 2.1 and 2.2 hold and, without loss of generality, assume* $\gamma_k^{IV} \neq 0$ *and* $\gamma_k^{Mis} \neq 0$ *for* $\forall k$. *Then, there exists a summary statistic* ξ *such that:*

$$\alpha^{Mis} = \sum_{k=1}^{K} \gamma_k^{IV} \alpha_{k,k-1} \times \frac{\gamma_k^{Mis}}{\gamma_k^{IV}} \implies \alpha^{IV} = \xi \alpha^{Mis} \tag{3}$$

where the ratio $\xi = \gamma_k^{IV} / \gamma_k^{Mis} = \sum_{k=1}^{K} \gamma_k^{IV} (p_{1,k} - p_{0,k})$.

Proof of Corollary 2.1. See Appendix A.2. □

The parameter ξ is a weighted average of the difference between misclassification probabilities, it is constant across k, with absolute value less than or equal to one, and unobserved in practice. A similar link between the causal and the identifiable parameter has been established in an IV setting by Hausman et al. (1998), Frazis and Loewenstein (2003), Lewbel (2007), Battistin and Sianesi (2011), Stephens Jr and Unayama (2019) and Calvi, Lewbel, and Tommasi (2018), under a

variety of different conditions.[9] Our main contribution with respect to this literature is to show that such relationship can be expressed in terms of the probabilities of false positive and false negative, which are statistics that are increasingly available in the data, in a heterogeneous-treatment-effects framework.

Indeed, denote the weighted average probability of false negative as $w^n = 1 - \sum_{k=1}^{K} \gamma_k^{IV} p_{1,k}$ (this is the probability of treated individuals misclassified as untreated) and false positive as $w^p = \sum_{k=1}^{K} \gamma_k^{IV} p_{0,k}$ (this is the probability of untreated individuals misclassified as treated). Then, by definition,

$$\xi = 1 - w^n - w^p,^{10} \tag{4}$$

which makes clear that ξ can be interpreted as a measure of how severe the treatment misclassification is. When there is no misclassification ($w^n = w^p = 0$), the parameter $\xi = 1$ and hence $\alpha^{Mis} = \alpha^{IV}$ because the bias is 0. As misclassification worsen ($w^n > 0, w^p > 0$), ξ falls and the bias becomes increasingly severe. When $0 < \xi < 1$, the bias in α^{Mis} is $1/\xi - 1 \times \alpha^{IV}$. When $\xi < 0$, α^{Mis} and α^{IV} are of opposite signs.

Table 1: Bias of α^{Mis} relative to α^{IV} for different misclassification probabilities

Bias $= (1/\xi - 1) \times \alpha^{IV}$						
$w^n \downarrow$ | $w^p \rightarrow$	0	0.05	0.10	0.20	0.30	0.40
0.00	0.000	0.053	0.111	0.250	0.429	0.667
0.05	0.053	0.111	0.176	0.333	0.538	0.818
0.10	0.111	0.176	0.250	0.429	0.667	1.000
0.20	0.250	0.333	0.429	0.667	1.000	1.500
0.30	0.429	0.538	0.667	1.000	1.500	2.333
0.40	0.667	0.818	1.000	1.500	2.333	4.000

Notes: Each cell reports $(1/\xi - 1) \times \alpha^{IV}$ for different values of w^n (false negative) and w^p (false positive). α^{IV} is normalized to 1.

A practitioner can use relationship (4) to approximate the possible level of bias of the estimated benefits of a program for different misclassification probabilities. In Table 1 we report the difference in the values between α^{Mis} and α^{IV} for different values of w^n and w^p. The true effect α^{IV} is normalized to 1. Hence, if the sample contains, e.g., 10% false negative and 5% false positive, this

[9] First, similarly to Frazis and Loewenstein (2003) and Stephens Jr and Unayama (2019), but differently from Lewbel (2007) and Battistin and Sianesi (2011), we establish a link between the causal and identifiable parameter by assuming an endogenous treatment. However, the assumed model in these papers is parametric, therefore the treatment effects are homogeneous. Second, similarly to Lewbel (2007) and Battistin and Sianesi (2011), we assume a nonparametric model, therefore the treatment effects are heterogeneous. However, they assume an exogenous treatment, hence unconfoundedness, which is not required in our context. Third, differently from Frazis and Loewenstein (2003), we assume monotonicity of the instrumental variable(s), which allows us to derive the misclassification probabilities in terms of the compilers. Finally, α^{Mis} generalizes the B-LATE (for Biased LATE) estimator of Calvi, Lewbel, and Tommasi (2018) to a discrete and multiple-discrete-instruments setting. Our factor $1/\xi$ becomes their factor $1/p$ (the fraction of individuals correctly reporting their treatment status) when a scalar binary instrument is used. All these papers, including ours, are related to one another and benefited from the result by Hausman et al. (1998).

[10] Notice that a practitioner does not need to know each value of $p_{1,k}$ and $p_{0,k}$, for $k = 1, 2, ..., K$, to be able to approximate the value of ξ. This is because, in practice, the type of information that is increasingly reported in the data is the overall misclassification probabilities, w^n and/or w^p, like in our application to the 401(k) Pension Plan. These are the only information actually required to approximate ξ. We come back to this point in Section 2.4 when we discuss the partial identification strategy that allows to incorporate external information to shrink the bounds of α^{IV}.

table tells the practitioner that the estimated effect α^{Mis} is approximately 17.6% larger than the (unknown) true effect.

2.2 Bounds of the LATEs and the LATMs

We introduce an additional assumption needed for partial identification results.

Assumption 2.3 (Informative Treatment Proxy)**.** *For all* $k = 1, 2, ..., K$, $Pr(T = d|C_k, D = d) > Pr(T = d|C_k, D = 1-d)$, $d = \{0, 1\}$.

Assumption 2.3 states that T is an informative proxy of the actual treatment status and it also rules out the negative weights in α^{Mis}. One sufficient condition for it is that $\max\{\Pr(T = 1|C_k, D = 0), \Pr(T = 0|C_k, D = 1)\} < 1/2$ for all $k = 1, 2, ..., K$, meaning that the observations of T are more accurate than pure guesses about the true treatment. A similar restriction is widely invoked in the measurement error literature.[11] Moreover, in a similar vein to the monotonicity condition of Hausman et al. (1998), in the following Lemma we show that Assumption 2.3 ensures that, when the instrumental variable Z varies, the mismeasured propensity score $\Pr(T = 1|Z)$ moves in the same direction of the true propensity score $\Pr(D = 1|Z)$.

Lemma 2.1. *Under Assumption 2.1, 2.2 and 2.3, we have that* $\mathbb{E}(T|Z = z_l) \leq \mathbb{E}(T|Z = z_w)$ *implies* $Pr(z_l) \leq Pr(z_w)$ *for* $\forall z_l, z_w \in \Omega_Z$.

Proof of Lemma 2.1. See Appendix A.3. □

Lemma 2.1 says that, even though the magnitude of propensity score $\Pr(z)$ cannot be recovered from the observed data, the proxy T reveals relevant information about how the actual treatment responds to the changes of Z. Given this Lemma, without loss of generality, hereafter we assume the elements $\{z_0, z_1, ..., z_K\}$ of the support Ω_Z follow an *ascending order*, in the sense that $\forall l, w \in \{0, 1, ..., K\}$, $l < w$ implies $\Pr(z_l) \leq \Pr(z_w)$. This order is identifiable given Lemma 2.1.

Bounding the probability of compliers. To bound the probability of compliers, we use the concept of total variation (TV) distance. For any generic random variable (or vector) A and $z_k, z_{k-1} \in \Omega_Z$, TV is a L^1 distance between the two conditional distribution functions $f_{A|Z=z_k}$ and $f_{A|Z=z_{k-1}}$, defined as follows:

$$TV_{A,k} = \frac{1}{2} \int |f_{A|Z=z_k}(a) - f_{A|Z=z_{k-1}}(a)| d\mu_A(a),$$

where μ_A denotes a dominating measure for the distribution of A.[12] If A is discrete, the integral is replaced by summation across all possible values of A. The $TV_{A,k}$ is identifiable and it captures the extent of the distributional effect of Z on A, when Z changes from z_{k-1} to z_k. If $A = Y$, then $TV_{Y,k}$

[11] See e.g. Bollinger (1996), Lewbel (2007), Hu (2008), Chen et al. (2011), Battistin and Sianesi (2011), and Battistin et al. (2014).

[12] For two σ-finite measures μ and μ', the measure μ' is dominated by μ, if, for any measurable set $\mathscr{A}$, $\mu(\mathscr{A}) = 0$ implies $\mu'(\mathscr{A}) = 0$. For more detailed definition, see the Radon-Nikodym Theorem in Billingsley (2008).

is the distribution version of the "intent-to-treat" effect. The TV will play a crucial role in bounding the probability of compliers when the actual treatment D is unobservable.

Lemma 2.2. *Let Assumption 2.1-(ii) to (iv), 2.2-(i) and 2.3 hold. We have that, for* $\forall k = 1, 2, ..., K$*:*

$$TV_{(Y,T),k} \leq Pr(C_k) \leq 1 - \sum_{k' \neq k} TV_{(Y,T),k'}.$$

Proof of Lemma 2.2. See Appendix A.4. □

By the definition of total variation distance, we know that $\Pr(C_k) = TV_{D,k}$, therefore $TV_{(Y,T),k}$ can be understood as the smallest effect of Z changing from z_{k-1} to z_k on the true treatment D. The bound for $\Pr(C_k)$ in Lemma 2.2 depends on the strength of the instrument(s). For example, if the change of Z from z_{k-1} to z_k causes no distributional variation of the outcome and the treatment proxy, the lower bound of $\Pr(C_k)$ reduces to 0. Similarly, if no distributional variation is triggered by the change of Z from $z_{k'-1}$ to $z_{k'}$ for all $k' \neq k$, the upper bound of $\Pr(C_k)$ increases to 1.[13]

Given Lemma 2.2, we can now proceed to consider the bounds for (i) the LATE ($\alpha_{k,k-1}$) and (ii) the difference between misclassification probabilities ($\Delta p_k = p_{1,k} - p_{0,k}$). For convenience, hereafter we refer to Δp_k as the local average of treatment misclassification (LATM), because it is analogous to the LATE if we replace $Y_1 - Y_0$ by $T_1 - T_0$:

$$\text{LATM} = \Delta p_k = \mathbb{E}[T_1 - T_0 | C_k]$$

Let $\mathbf{P}$ be an arbitrary data generating process (DGP) of (Y, T, Z). Denote the class of DGPs of $\mathbf{P}$ as $\mathscr{P}_0$, then we have $\mathbf{P} \in \mathscr{P}_0$. Denote Θ to be the parameter space of α^{IV}, α^{Mis} and of all $\alpha_{k,k-1}$. For example, $\Theta = \{-1, 1\}$ if outcome Y is binary, and $\Theta = \Omega_Y$ if outcome Y is continuous.[14] For $A = \{Y, T\}$, denote $\Delta_k \mathbb{E}(A|Z) = \mathbb{E}(A|Z = z_k) - \mathbb{E}(A|Z = z_{k-1})$.

Bounding the LATEs. Theorem 1 in Imbens and Angrist (1994) says that under Assumption 2.1 in this paper, we have:

$$\Delta_k \mathbb{E}(Y|Z) = \alpha_{k,k-1} \Pr(C_k). \tag{5}$$

Multiplying both sides of (5) by $\alpha_{k,k-1}$, we obtain that:

$$\alpha_{k,k-1} \Delta_k \mathbb{E}(Y|Z) = \alpha_{k,k-1}^2 \Pr(C_k) \geq 0. \tag{6}$$

[13] Lemma 2.2 generalizes Lemma 3 of Ura (2018) to accommodate multiple or multi-valued IV(s). Notice that, when instrument is discrete, this author proposes to use only the subpopulation where the instrument takes two values, and bound $\Pr(C_k)$ by $[TV_{(Y,T),k}, 1]$. However, we demonstrate the identification power gain of discrete IV(s), as we can actually bound $\Pr(C_k)$ from above by $1 - \sum_{k' \neq k} TV_{(Y,T),k'}$ instead of 1.

[14] The parameter space for each $\alpha_{k,k-1}$ may be different for each k. However, we ignore this possibility for notational simplicity.

Moreover, by applying Lemma 2.2 to the absolute value of (5), we have:

$$|\Delta_k \mathbb{E}(Y|Z)| \leq |\alpha_{k,k-1}| \left[1 - \sum_{k' \neq k} TV_{(Y,T),k'}\right], \tag{7}$$

$$|\Delta_k \mathbb{E}(Y|Z)| \geq |\alpha_{k,k-1}| TV_{(Y,T),k}. \tag{8}$$

Thus, under Assumptions 2.1, 2.2 and 2.3, each LATE $\alpha_{k,k-1}$ satisfies the inequalities (6)-(8). Inequality (6) indicates that the sign of $\alpha_{k,k-1}$ is identified by $\Delta_k \mathbb{E}(Y|Z)$ whenever $\Pr(C_k)$ is nonzero. In addition, when $\Delta_k \mathbb{E}(Y|Z) \neq 0$, inequalities (7) and (8) give the lower and upper bounds of $|\alpha_{k,k-1}|$, respectively. Denote the set of $\alpha_{k,k-1}$, characterized by (6)-(8), as $\Theta_k^{\alpha}(\mathbf{P}) \subset \Theta$. In the next Lemma, we derive explicit expressions for $\Theta_k^{\alpha}(\mathbf{P})$, and provide sufficient conditions under which $\Theta_k^{\alpha}(\mathbf{P})$ is a sharp identified set of LATE.

Lemma 2.3. *Let Assumption 2.1-(ii)-(iv), 2.2-(i) and 2.3 hold. Then, for* $\forall k = 1, 2, ..., K$*:*

(i) If $TV_{(Y,T),k} = 0$*, then* $\Theta_k^{\alpha}(\mathbf{P}) = \Theta$*. Whereas if* $TV_{(Y,T),k} > 0$*, then:*

$$\Theta_k^{\alpha}(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) < 0; \end{cases} \tag{9}$$

(ii) If $\max_{0 \leq m \leq K} TV_{(Y,T),m} = 0$*, then* $\Theta_k^{\alpha}(\mathbf{P}) = \Theta$ *is the sharp identified set of* $\alpha_{k,k-1}$*. Whereas, if* $TV_{(Y,T),k} > 0$ *and* $TV_{(Y,T),k'} = 0$ *for all* $k' \neq k$*, then* $\Theta_k^{\alpha}(\mathbf{P})$ *in* (9) *is the sharp identified set of* $\alpha_{k,k-1}$*.*

Proof of Lemma 2.3. See Appendix A.5. □

Lemma 2.3 (i) shows that, if $TV_{(Y,T),k} = 0$, then no useful information about how the instrument's value changing from z_{k-1} to z_k affects the treatment can be extracted from the observable data, so that we fail to exclude any values from the parameter space of the LATE, Θ. Once $TV_{(Y,T),k} > 0$, the instrument has nontrivial identification power, and analytic bounds can be derived for the LATE. To be more specific, if $\Delta_k \mathbb{E}(Y|Z) = 0$, then $\alpha_{k,k-1}$ is point identified as zero. If $\Delta_k \mathbb{E}(Y|Z) \neq 0$, the sign of $\alpha_{k,k-1}$ is identified by the sign of $\Delta_k \mathbb{E}(Y|Z)$. Importantly, $\Theta_k^{\alpha}(\mathbf{P})$ can be seen as an informative bound for LATE, because it excludes the intention to treat (ITT) effect $\Delta_k \mathbb{E}(Y|Z)$ and the naive Wald estimand $\Delta_k \mathbb{E}(Y|Z)/\Delta_k \mathbb{E}(T|Z)$,[15] which are the two trivial bounds of the LATE. Moreover, if the total variation is nonzero only when Z changes from z_{k-1} to z_k, then $\Theta_k^{\alpha}(\mathbf{P})$ is the sharp identified set and it reduces to the identified set of Ura (2018). This is intuitive because $TV_{(Y,T),k'} = 0$ for all $k' \neq k$ implies that $Z = z_{k-1}$ and $Z = z_k$ are the only two values inducing nonzero changes in the outcome or the treatment. Thus, the multiple total variation distances generated from the discrete IV(s) are essentially equivalent to that generated from a binary IV.

[15] This is because, by Lemma A.1 in Appendix, we have $TV_{(Y,T),k} \geq |\Delta_k \mathbb{E}(T|Z)|$.

For more general cases, where more than two total variation distances are nonzero, although sharpness result is not established for $\Theta_k^{\alpha}(\mathbf{P})$, this outer set still possesses desirable properties. First, it is tighter than the bound provided by Ura (2018) using two values of Z. Second, since an outer set is always a superset of the sharp identified set, inference based on the outer set is conservative yet valid. Moreover, the outer set is often considered useful in practice, since it may be sufficient to answer important empirical questions, such as whether the treatment effect is negative or what is the possible range of program benefits (see Molinari, 2020, for more details). Third, point identification of LATE can be established, if either (i) there is no misclassification or (ii) the IV(s) has perfect explanatory power of the treatment, e.g., when $\Pr(D=1|Z=z_k)=1$ and $\Pr(D=0|Z=z_{k-1})=1$. Intuitively, the bounds will be tighter in cases that are "closer" to either of these two extreme cases.

Bounding the LATMs. Similar arguments can be applied to obtain the inequalities (10)-(12) below, satisfied by each Δp_k:

$$\Delta p_k \Delta_k \mathbb{E}(T|Z) \geq 0, \tag{10}$$

$$|\Delta_k \mathbb{E}(T|Z)| \leq |\Delta p_k| \left[1 - \sum_{k' \neq k} TV_{(Y,T),k'} \right], \tag{11}$$

$$|\Delta_k \mathbb{E}(T|Z)| \geq |\Delta p_k| TV_{(Y,T),k}. \tag{12}$$

Denote the set of Δp_k, characterized by (10)-(12), as $\Theta_k^p(\mathbf{P})$. The Lemma below gives analytic bounds of Δp_k, as well as sufficient conditions for the sharp identified set.

Lemma 2.4. *Let Assumption 2.1-(ii)-(iv), 2.2-(i) and 2.3 hold. For* $\forall k = 1, 2, ..., K$,

(i) If $TV_{(Y,T),k} = 0$, *then* $\Theta_k^p(\mathbf{P}) = [-1, 1]$. *Whereas, if* $TV_{(Y,T),k} > 0$, *then:*

$$\Theta_k^p(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(T|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}}, \frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}} \right], & \text{if } \Delta_k \mathbb{E}(T|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(T|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}}, \frac{\Delta_k \mathbb{E}(T|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \right], & \text{if } \Delta_k \mathbb{E}(T|Z) < 0; \end{cases} \tag{13}$$

(ii) If $\max_{0 \leq m \leq K} TV_{(Y,T),m} = 0$, *then* $\Theta_k^p(\mathbf{P}) = [-1, 1]$ *is the sharp identified set of* Δp_k. *Whereas, if* $TV_{(Y,T),k} > 0$ *and* $TV_{(Y,T),k'} = 0$ *for all* $k' \neq k$, *then* $\Theta_k^p(\mathbf{P})$ *in* (13) *is the sharp identified set of* Δp_k.

Proof of Lemma 2.4. See Appendix A.6. □

We are interested in the identified set of Δp_k because it plays a crucial role in characterizing the bias of α^{Mis} relative to the object of interest, α^{IV}. As shown in Lemma 2.4, the sign and an informative bound for Δp_k can be obtained as long as $TV_{(Y,T),k} > 0$.[16] It is also clear that, in order

[16] Since, without loss of generality, we assume $\{z_0, z_1, ..., z_K\}$ follow the ascending order, then it is clear that $\Delta_k \mathbb{E}(T|Z) \geq 0$ for all k. For the sake of completeness, however, in Lemma 2.4 we still present the result for the case $\Delta_k \mathbb{E}(T|Z) < 0$.

to partial identify Δp_k, we do not need any prior or external information about how severely the treatment proxy T is contaminated by measurement error.

2.3 Partial Identification of α^{IV}

The bounds of the LATEs and the LATMs provide the fundamental basis for the identification of the estimand α^{IV}. In this section, we begin by proposing two strategies to partially identify α^{IV}. Both strategies do not rely on additional or external sources of information.

First strategy. Recall that the estimand α^{IV} is a weighted average of LATEs $\{\alpha_{k,k-1}\}_{k=1}^{K}$ with nonnegative weights $\{\gamma_k^{IV}\}_{k=1}^{K}$ summing up to one. Hence the first partial identification strategy is based on $\{\alpha_{k,k-1}\}_{k=1}^{K}$:

$$\min_{k=1,2,...,K}\{\alpha_{k,k-1}\} \leq \alpha^{IV} = \sum_{i=1}^{K} \gamma_k^{IV} \alpha_{k,k-1} \leq \max_{k=1,2,...,K}\{\alpha_{k,k-1}\} \tag{14}$$

Given (14), our first partial identification result of α^{IV} can be obtained from the bounds of LATEs given in Lemma 2.3.

Theorem 2.2. *Let Assumption 2.1, 2.2, and 2.3 hold. Denote* $\Theta^{\alpha}(\mathbf{P}) = \bigcup_{k\in\{1,2,...,K\}} \Theta_k^{\alpha}(\mathbf{P})$. *Then we have* $\alpha^{IV} \in \Theta^{\alpha}(\mathbf{P})$.

Proof of Theorem 2.2. See Appendix A.7. □

The superscript α of $\Theta^{\alpha}(\mathbf{P})$ means that it is constructed from $\{\Theta_k^{\alpha}(\mathbf{P})\}_{k=1}^{K}$. Theorem 2.2 shows that α^{IV} lies in the union of the partially identified sets of LATEs $\{\alpha_{k,k-1}\}_{k=1}^{K}$. In principle, the set $\Theta^{\alpha}(\mathbf{P})$ might be uninformative about the direction of the WLATE in situations where at least two LATEs, $\alpha_{k,k-1}$ and $\alpha_{k',k'-1}$, have opposite signs. Fortunately, because we can recover the sign of all the LATEs from the observed data (Lemma 2.3), we are able to recover the sign of α^{IV} as long as all the LATEs stand on the same side of zero. We refer to this feature of the data as "direction consistency" of LATEs. This knowledge reveals partly how the treatment affects the outcome which, in many empirical applications, is supported by economic theory. For example, in a study of the returns to schooling, a higher education level indicates, on average, higher wages. Hence, in this case, the "direction consistency" of LATEs is positive.[17]

Corollary 2.2. *Let Assumption 2.1, 2.2, and 2.3 hold.*

(i) If $\Delta_k \mathbb{E}(Y|Z) > 0$ *for all* $k = 1, 2, ..., K$, *then* $\alpha^{IV} > 0$ *and*

$$\Theta^{\alpha}(\mathbf{P}) = \left[\min_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z)}{1 - \sum_{k'\neq k} TV_{(Y,T),k'}} \right\}, \max_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}} \right\} \right].$$

[17] Such sign restriction, or direction consistency, is commonly assumed in the treatment effects partial identification literature. For example, the monotone treatment response $Y_1 \geq Y_0$ (or $Y_1 \leq Y_0$) for all individuals in Manski (1997), Manski and Pepper (2000, 2009) and Bhattacharya et al. (2008) among others. Another weaker condition is the monotonicity of average outcomes in treatment at strata level, $\mathbb{E}(Y_1|C_k) \geq \mathbb{E}(Y_0|C_k)$, proposed by Chen et al. (2018). The strata level monotonicity is more plausible in practice, without restricting the sign for all individuals.

(ii) If $\Delta_k \mathbb{E}(Y|Z) < 0$ *for all* $k = 1, 2, ..., K$*, then* $\alpha^{IV} < 0$ *and*

$$\Theta^{\alpha}(\mathbf{P}) = \left[\min_{k \in \{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}} \right\}, \max_{k \in \{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \right\} \right].$$

Proof of Corollary 2.2. The proof follows from Lemma A.1 in Appendix and Theorem 2.2. □

The Corollary above provides the identification of the sign of α^{IV}, as well as the explicit expression of $\Theta^{\alpha}(\mathbf{P})$, when the direction consistency of LATEs is satisfied. If for some k, we have $\Delta_k \mathbb{E}(Y|Z) = 0$, the results above still hold with one side of the bounds being zero and the possibility $\alpha^{IV} = 0$.

In finite samples, estimators taking maximums and minimums are systematically biased (Kreider and Pepper, 2007; Chernozhukov et al., 2013). For $\Theta^{\alpha}(\mathbf{P})$, because we take the union of the bounds of LATEs, then its lower bound, i.e., the minimum of the lower bound estimators of the LATEs, is biased downward, and its upper bounds, i.e., the maximum of the upper bound estimators of the LATEs, is biased upward. Hence, in our setting, this bias is not of particular concern because it results in wider bounds rather than narrower bounds in finite samples. Nevertheless, we note that the confidence interval of our proposed bounds will be conservative in the sense its size is less than the nominal size. The same logic applies to the bounds in second strategy in the next section.

Second strategy. Our second strategy is built upon the relation between α^{IV} and α^{Mis}, and the bounds of LATMs. Recall from Corollary 2.1 that $\alpha^{IV} = \xi \alpha^{Mis}$, where $\xi = \sum_{k=1}^{K} \gamma_k^{IV} \Delta p_k$. Based on the definition of ξ, we have:

$$\min_{k=1,2,...,K} \{\Delta p_k\} \leq \xi = \sum_{k=1}^{K} \gamma_k^{IV} \Delta p_k \leq \max_{k=1,2,...,K} \{\Delta p_k\}. \tag{15}$$

Based on (15), our second partial identification result of α^{IV} can be characterized using the bounds of LATMs.

Theorem 2.3. *Let Assumption 2.1, 2.2, and 2.3 hold. Denote*

$$\Theta^{p}(\mathbf{P}) = \left\{ \alpha^{Mis} \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,...,K} \Theta_k^{p}(\mathbf{P}) \right\}, \tag{16}$$

where Δp *represents any generic value in the union* $\bigcup_{k=1,2,...,K} \Theta_k^{p}(\mathbf{P})$*. Then we have* $\alpha^{IV} \in \Theta^{p}(\mathbf{P})$*.*

Proof of Theorem 2.3. See Appendix A.8. □

The superscript p of $\Theta^{p}(\mathbf{P})$ represents its key components $\{\Theta_k^{p}(\mathbf{P})\}_{k=1}^{K}$. Theorem 2.3 gives the general form of the set $\Theta^{p}(\mathbf{P})$, based on both the identifiable estimand α^{Mis} and the bounds of $\{\Delta p_k\}_{k=1}^{K}$. If $\alpha^{Mis} = 0$, α^{IV} is point identified as zero.

Corollary 2.3. *Let Assumption 2.1, 2.2, and 2.3 hold. Suppose* $\Delta_k \mathbb{E}(T|Z) > 0$ *for all* $k = 1, 2, ..., K$*.*

(i) If $\alpha^{Mis} \geq 0$*, then* $\alpha^{IV} \geq 0$ *and*

$$\Theta^p(\mathbf{P}) = \left[\alpha^{Mis} \times \min_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \right\}, \alpha^{Mis} \times \max_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}} \right\} \right],$$

(ii) If $\alpha^{Mis} < 0$*, then* $\alpha^{IV} < 0$ *and*

$$\Theta^p(\mathbf{P}) = \left[\alpha^{Mis} \times \min_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}} \right\}, \alpha^{Mis} \times \max_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \right\} \right].$$

Proof of Corollary 2.3. It follows directly from Lemma A.1 in Appendix and Theorem 2.3. □

The Corollary above gives the sign of α^{IV}, and the explicit expression of $\Theta^p(\mathbf{P})$. Note that $\Delta_k \mathbb{E}(T|Z) \geq 0$ for all k is satisfied under the ascending order of Ω_Z. If $\Delta_k \mathbb{E}(T|Z) = 0$ for some k, then one side of the bounds for α^{IV} reduces to zero.

First vs Second strategy. The partial identification results of the two strategies introduced thus far are both compatible with the observable data under Assumption 2.1, 2.2 and 2.3. Moreover, they make distinct contributions to the identification of α^{IV} because they are based on different sources of information. Since the two sets $\Theta^{\alpha}(\mathbf{P})$ and $\Theta^p(\mathbf{P})$ are likely to be different, it is important to determine their relative performance. In order to facilitate this comparison, we re-write $\Theta^{\alpha}(\mathbf{P})$ and $\Theta^p(\mathbf{P})$ as the unions of the re-scaled $\Theta^p_k(\mathbf{P})$.

Corollary 2.4. *Under Assumption 2.1, 2.2 and 2.3,* $\Theta^{\alpha}(\mathbf{P})$ *and* $\Theta^p(\mathbf{P})$ *can be rewritten as follows:*

$$\Theta^{\alpha}(\mathbf{P}) = \bigcup_{k=1,2,...,K} \left\{ \frac{\alpha_{k,k-1}}{\Delta p_k} \times \Delta p : \ \Delta p \in \Theta^p_k(\mathbf{P}) \right\},$$

$$\Theta^p(\mathbf{P}) = \bigcup_{k=1,2,...,K} \left\{ \frac{\alpha^{IV}}{\xi} \times \Delta p : \ \Delta p \in \Theta^p_k(\mathbf{P}) \right\},$$

where Δp *is any generic value in* $\Theta^p_k(\mathbf{P})$.

Proof of Corollary 2.4. See Appendix A.9. □

Corollary 2.4 delivers four crucial messages. First, in general, unless more information are available, it is not a-priori obvious which set outperforms the other, since $\alpha_{k,k-1}/\Delta p_k$ may not be uniformly larger or smaller than α^{IV}/ξ across all k. Second, when the ratios $\{\alpha_{k,k-1}/\Delta p_k\}_{k=1}^K$ are the same across all k, we have that $\Theta^{\alpha}(\mathbf{P}) = \Theta^p(\mathbf{P})$. This special case, however, relies on both unconfounded treatment and homogenous misclassification, which may be quite restrictive in practice. Third, for all $\Delta p \in \Theta^p_k(\mathbf{P})$ and all k, the closer to 1 is the ratio $\Delta p/\xi$, the narrower is the set delivered by Strategy 2 (that is, the narrower is the set $\Theta^p(\mathbf{P})$). Fourth, at the limit, if $\Delta p_k = \xi$ for all k, that is, the data satisfy homogeneous misclassification, then $\Theta^p_k(\mathbf{P}) = \xi$. In this last case,

point identification is achieved by $\Theta^p(\mathbf{P})$ as follows:

$$\Theta^p(\mathbf{P}) = \bigcup_{k=1,2,...,K} \{\alpha^{IV}\} = \alpha^{IV}.$$

However, for $\Delta p_k = \xi$, the improvement of Strategy 1 is not as good as that of Strategy 2. This is because, although $\alpha_{k,k-1}$ can be point identified by $\xi\Delta_k\mathbb{E}(Y|Z)/\Delta_k\mathbb{E}(T|Z)$,[18] from Corollary 2.4 we have:

$$\Theta^{\alpha}(\mathbf{P}) = \left[\min_{k=1,2,...,K}\{\alpha_{k,k-1}\}, \max_{k=1,2,...,K}\{\alpha_{k,k-1}\}\right]$$

which only partially identifies α^{IV}. Thus, whenever the misclassification error is close to be homogenous (that is, the correlation between the misclassification error and the potential treatments is small), Strategy 2 should, in general, outperform Strategy 1.

Two final remarks. First, following the method of intersecting the bounds, which is commonly applied in the treatment effect partial identification literature, there is no issue preventing us from intersecting $\Theta^{\alpha}(\mathbf{P})$ and $\Theta^p(\mathbf{P})$ to achieve a tighter bound. Intersection bounds can be biased in finite samples, with the estimated bounds being too narrow and its size being routinely underestimated (Kreider and Pepper, 2007; Chernozhukov et al., 2013). Note that it is different from the bias resulting from taking the union of bounds. One solution to avoid such bias is to apply the bias correction method proposed by Chernozhukov et al. (2013). For practical purpose, however, adopting only one strategy may be beneficial for computational simplicity. Second, it is interesting to note that, if the instrument is binary, α^{IV} is just the LATE and α^{Mis} with $g(x) = x$ reduces to:

$$\alpha^{Mis} = \frac{\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]}{\mathbb{E}[T|Z=1]-\mathbb{E}[T|Z=0]} = \frac{\mathbb{E}[Y_1-Y_0|D_1=1,D_0=0]}{p_1-p_0}.$$

Then, Theorem 2.2 and 2.3 will be identical. In addition, the results of our first two partial identification strategies will also coincide with that in Ura (2018), because $K = 1$ and $\sum_{k'\neq k} TV_{(Y,T),k'}$ degenerates to zero.

2.4 Partial Identification of α^{IV} Using External Information

Suppose the practitioner has some prior or external information about the possible range of ξ, and this range is narrower than $\bigcup_{k=1,2,...,K}\Theta_k^p(\mathbf{P})$ which is obtained using the observable data. Then we can utilize this narrower range to further tighten the bounds. This strategy is based on the observation that administrative records of program receipts are not easily accessible to all researchers, and hence we cannot know exactly who is misclassified in a survey. However, an increasing number of studies report the average extent of misreporting for a wide range of programs. This information often comes in the form of average number of false negative (w^n) and/or false positive (w^p) in the sample. In light of Equation (4), this information is sufficient to restrict the possible range of ξ. Validation studies or repeated measurements of the same individual can also provide valuable

[18] This is because $\frac{\Delta_k\mathbb{E}(Y|Z)}{\Delta_k\mathbb{E}(T|Z)} = \frac{\alpha_{k,k-1}}{\Delta p_k}$. If $\Delta p_k = \xi$ for a known ξ, then $\alpha_{k,k-1} = \xi\frac{\Delta_k\mathbb{E}(Y|Z)}{\Delta_k\mathbb{E}(T|Z)}$ is point identified.

information.

At the risk of repetition, recall from Corollary 2.1 that $\alpha^{IV} = \xi\alpha^{Mis}$, where $\xi = 1 - w^n - w^p$. Our third identification strategy is described below.

Theorem 2.4. *Let Assumption 2.1 and 2.2 hold. Suppose there exist two known constants* $\underline{\xi} \leq \overline{\xi}$ *and* $\underline{\xi}, \overline{\xi} \in [0,1]$*, such that* $\underline{\xi} \leq \xi \leq \overline{\xi}$*.*

(i) If $\alpha^{Mis} \geq 0$*, denote* $\Theta^{\xi}(\mathbf{P}) = \left[\underline{\xi}\alpha^{Mis}, \overline{\xi}\alpha^{Mis}\right]$*. Then,* $\alpha^{IV} \geq 0$ *and* $\alpha^{IV} \in \Theta^{\xi}(\mathbf{P})$*.*

(ii) If $\alpha^{Mis} < 0$*, denote* $\Theta^{\xi}(\mathbf{P}) = \left[\overline{\xi}\alpha^{Mis}, \underline{\xi}\alpha^{Mis}\right]$*. Then,* $\alpha^{IV} < 0$ *and* $\alpha^{IV} \in \Theta^{\xi}(\mathbf{P})$*.*

Proof of Theorem 2.4. See Appendix A.10. □

Intuitively, the constants $\underline{\xi}$ and $\overline{\xi}$ are two bounds of the weighted average of LATMs. By using these extra information, the set $\Theta^{\xi}(\mathbf{P})$ will be at least as good as that in Corollary 2.3 (second strategy). If no extra information about the measurement accuracy is available, one could simply set $\underline{\xi}$ and $\overline{\xi}$ as the ending points of $\bigcup_{k=1,2,\ldots,K} \Theta_k^p(\mathbf{P})$. Therefore, compared to the first two identification strategies, which are based purely on the observable data, by following our third strategy one can further tighten the bounds of α^{IV} and obtain (potentially) significant improvements.

Point estimate. From Theorem 2.4, two sets of conditions suffice to obtain tighter bounds. Firstly, having $\underline{\xi}$ close to 1 means less overall misclassification. At the extreme, when $\underline{\xi} = 1$, we have no misclassification error at all ($w^n = w^p = 0$), hence we can achieve point identification of $\alpha^{IV} = \alpha^{Mis}$. Secondly, having $(\underline{\xi}, \overline{\xi})$ close to each other indicates more accurate knowledge of the overall misclassification probabilities, which produces a narrower bound as well. At the extreme, when $\underline{\xi} = \overline{\xi} = \xi$, we can also achieve point identification of $\alpha^{IV} = \xi\alpha^{Mis}$. Notice that, in application, the constants $\underline{\xi}$ and $\overline{\xi}$ are going to be two approximations of the bounds of the misclassification probabilities. Hence, if the practitioner can set $\underline{\xi} = \overline{\xi} = \xi$, the point estimate delivered by the estimator $\xi\alpha^{Mis}$ is going to be biased with respect to α^{IV}, unless ξ is the exact value of misclassification. If $\underline{\xi}$ and $\overline{\xi}$ are approximations, then our approach can be used as a bias reduction method with respect to a naïve IV estimator.

3 Inference

A feasible approach to compute the confidence interval for the partially identified α^{IV} is via the bootstrap-based testing of moment inequalities proposed by Chernozhukov et al. (2019). In the present paper, the inferential procedure is based on a modification of their approach that accommodates our setting. Specifically, since the partial identification of α^{IV} is based on the union of either $\Theta_k^{\alpha}(\mathbf{P})$ or $\Theta_k^{p}(\mathbf{P})$, we proceed with the estimation in three steps. First, we construct the moment inequalities representations of the sets $\Theta_k^{\alpha}(\mathbf{P})$ and $\Theta_k^{p}(\mathbf{P})$. Second, we construct the confidence intervals for $\alpha_{k,k-1}$ and Δp_k. Third, depending on the chosen identification strategy, we construct the appropriate confidence intervals of α^{IV} by taking the unions of the confidence intervals of either

$\alpha_{k,k-1}$ or Δp_k. To save on space, we here report only the proposed confidence intervals of our partial identification strategies and their asymptotic properties. All the technical details and derivations are left in Appendix A.1.

To obtain the confidence interval of α^{IV}, the nuisance parameters π_k and α^{Mis} are estimated in advance. Suppose a $(1-\eta_\pi)$-confidence interval of π_k and a $(1-\eta_{\alpha^{Mis}})$-confidence interval of α^{Mis} are available to the practitioner. For any $\beta > 0$, denote the $(1-\beta)$-confidence interval of $\alpha_{k,k-1}$, Δp_k and α^{Mis} as $\mathscr{C}_{\alpha_{k,k-1}}(\beta)$, $\mathscr{C}_{\Delta p_k}(\beta)$ and $\mathscr{C}_{\alpha^{Mis}}(\beta)$, respectively.

First, based on the first partial identification strategy in Theorem 2.2, we propose a $(1-\beta^\alpha)$-confidence interval for α^{IV}:

$$\mathscr{C}^\alpha(\beta^\alpha) := \bigcup_{k=1,2,..,K} \mathscr{C}_{\alpha_{k,k-1}}(\eta + 2\eta_\pi), \tag{17}$$

where the size $\beta^\alpha = \eta + 2\eta_\pi$.

Second, based on the second partial identification strategy in Theorem 2.3, we propose another $(1-\beta^p)$-confidence interval for α^{IV}:

$$\mathscr{C}^p(\beta^p) := \bigcup_{\alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{ \alpha \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,..,K} \mathscr{C}_{\Delta p_k}(\eta + 2\eta_\pi) \right\}, \tag{18}$$

where the size $\beta^p = \eta_{\alpha^{Mis}} + \eta + 2\eta_\pi$.

The last confidence interval comes from our partial identification strategy with external sources of information in Theorem 2.4:

$$\mathscr{C}^\xi(\beta^\xi) := \bigcup_{\alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{ \alpha \times \Delta p : \ \Delta p \in \left[\underline{\xi}, \overline{\xi}\right] \right\}, \tag{19}$$

where $\beta^\xi = \eta_{\alpha^{Mis}}$ and $\underline{\xi}, \overline{\xi}$ are known values such that the true value of $\xi \in \left[\underline{\xi}, \overline{\xi}\right]$.

The next Corollary gives the asymptotic properties of $\mathscr{C}^\alpha(\beta^\alpha)$, $\mathscr{C}^p(\beta^p)$ and $\mathscr{C}^\xi(\beta^\xi)$.

Corollary 3.1. *Let the assumptions in Theorem A.1 hold. Furthermore, let θ be any point in $\Theta^j(\mathscr{P})$, $j \in \{\alpha, p, \xi\}$. Then, $\mathscr{C}^\alpha(\beta^\alpha)$, $\mathscr{C}^p(\beta^p)$ and $\mathscr{C}^\xi(\beta^\xi)$ defined in (17)-(19) all control their sizes asymptotically and uniformly over $\mathscr{P}_0$, i.e.*

$$\liminf_{n\to\infty} \inf_{\mathbf{P}\in\mathscr{P}_0,\ \theta\in\Theta^j(\mathbf{P})} Pr\left[\theta \in \mathscr{C}^j\left(\beta^j\right)\right] \geq 1-\beta^j, \text{ for all } j \in \{\alpha, p, \xi\}.$$

Proof of Corollary 3.1. See Appendix A.1.6. □

Corollary 3.1 proves that, for all the three confidence intervals of α^{IV}, their asymptotic coverage rate, at any point inside the associated identified set, achieves the desired level. Moreover, for given η and η_π, $\mathscr{C}^\alpha(\beta^\alpha)$ has a higher coverage rate than $\mathscr{C}^p(\beta^p)$, because $\mathscr{C}^p(\beta^\alpha)$ is constructed also based on the $(1-\eta_{\alpha^{Mis}})$-confidence interval of α^{Mis}. The coverage rate of $\mathscr{C}^\xi(\beta^\xi)$ is in general the highest, since we can set $\eta_{\alpha^{Mis}} \leq \beta^\alpha$.

4 Extensions and Applications

This section is organized in four parts. First, we sketch two extensions of our partial identification strategies, which are fully developed in the appendix. Second, we show how to implement our method in practice. Third, we describe the main ideas and results of our Monte Carlo simulations, which are fully presented in the older version of our paper (Tommasi and Zhang, 2020). Finally, we use our Stata command, `ivbounds`, to measure the benefits of participating in the 401(k) pension plan on savings.

4.1 Multiple treatments or repeated measurements

The results in section 2.3 and 2.4 require only one binary treatment indicator, T. Nevertheless, we show that if there are multiple treatment proxies (or repeated measurements), we can further tighten the bounds of α^{IV}, since each proxy may carry different relevant information about the actual treatment, D.[19] Based on the full set of results presented in Appendix A.2, when multiple treatment proxies are available, all three confidence intervals can be obtained in the same manner as in (17)-(19). Moreover, there are two main differences between our approach and those commonly used in the literature when multiple treatments are available.[20] First, we do not restrict the dependence among our treatment proxies; therefore, the extra measures might be endogenous and do not have to be instruments. In addition, our proxies may be built upon the same, not repeated, measurement by creating multiple treatment dummies from the same discrete treatment variable and capturing various pieces of useful information in the same measurement.

Note that, in practice, one could argue that an easier way to use multiple treatment proxies is to combine them into a new, single treatment measure that might suffer from less misclassification. We verify in a simple case that if two treatment proxies, denoted as T and S, are both binary, the bounds of α^{IV} constructed as in Appendix A.2 are tighter than the bounds using a single proxy $1[T = 1, S = 1]$. Therefore, when multiple proxies are available, our proposed method is preferable.

4.2 Including covariates

In many applications, the instrumental variable(s) may be confounded without conditioning on some covariates. In addition, treatment effects may be heterogeneous across the population and characterized by different attributes. Hence, in the identification of causal effects, particular attention has been paid to accounting for covariates (e.g., see Abadie (2003), Frölich (2007) and Angrist and Fernandez-Val (2013) among many others). Following this literature, we extend all our partial identification results to accommodate for covariates. To save on space, we here report only the main insights and all details are relegated to Appendix A.3.

[19] Here we refer to multiple treatment measures as "multiple treatment proxies", in the sense that the extra treatment measures (other than the binary T used in the previous sections), can be binary, discrete or continuous.

[20] Indeed, in the presence of misclassification error, multiple treatment proxies in the form of repeated measurements are widely used in both point and partial identification of treatment effects literature (see e.g. Hausman et al. (1991) among others).

Let X be a vector of observables with support Ω_X. Define the conditional IV estimand $\alpha^{IV}(x)$, which can be expressed as a weighted average of the conditional LATEs $\{\alpha_{k,k-1}(x)\}_{K=1}^{k}$:

$$\alpha^{IV}(x) := \frac{\text{Cov}(Y, g(Z)|X=x)}{\text{Cov}(D, g(Z)|X=x)} = \sum_{k=1}^{K} \gamma_k^{IV}(x)\alpha_{k,k-1}(x), \tag{20}$$

where $\gamma_k^{IV}(x)$ is the weight and $\sum_{k=1}^{K} \gamma_k^{IV}(x) = 1$ for $\forall x \in \Omega_X$. Given (20), the extension of our three main partial identification strategies to their conditional version is straightforward.

We can distinguish two main cases based on the support of instrument(s). First, if the instrument is binary and covariates are included, one can target the unconditional IV estimand, $\mathbb{E}[\alpha^{IV}(X)]$, by equipping our bounding strategies with the result in Frölich (2007):[21]

$$\mathbb{E}[\alpha^{IV}(X)] = \mathbb{E}[\alpha_{1,0}(X)] = \frac{\mathbb{E}[\mathbb{E}(Y|X,Z=1) - \mathbb{E}(Y|X,Z=0)]}{\mathbb{E}[\mathbb{E}(D|X,Z=1) - \mathbb{E}(D|X,Z=0)]} = \frac{\mathbb{E}[\Delta_1 \mathbb{E}(Y|X,Z)]}{\mathbb{E}[\Pr(C_1|X)]},$$

where the numerator is identifiable and the denominator can be bounded via the conditional version of our method.

Second, if the instrument(s) are discrete or multiple-discrete and covariates are included, one can use our strategies to target the conditional IV estimand $\alpha^{IV}(x)$. We focus on $\alpha^{IV}(x)$ for two reasons. First, since $\mathbb{E}[\alpha^{IV}(X)] \neq \sum_{k=1}^{K} \mathbb{E}[\gamma_k^{IV}(X)]\mathbb{E}[\alpha_{k,k-1}(X)]$, unless the weight $\gamma_k^{IV}(X)$ is degenerate in X, the result in Frölich (2007) cannot be applied. One sufficient condition for $\gamma_k^{IV}(X)$ invariant to X is that covariates are independent of (D, Z), which may not be feasible in many studies and goes against our attempts to include the covariates. Thus, targeting the unconditional IV estimand $\mathbb{E}[\alpha^{IV}(X)]$ in this general case is not straightforward (more details are provided in the end of Appendix A.3) and a rigorous solution is beyond the scope of this paper.[22] Moreover, $\alpha^{IV}(x)$ has a clear relationship with the conditional LATEs, which are the foundation of our partial identification strategies in the presence of covariates.

We need to further consider two scenarios in the general case when instrument(s) are discrete or multiple-discrete and covariates are included. First, the simplest scenario is when all covariates in X take on a finite number of values. Angrist and Imbens (1995) and Angrist and Fernandez-Val (2013) study the conditional treatment effects when covariates are discrete. Assuming covariates are discrete is not required for partial identification of $\alpha^{IV}(x)$, while, in this case, a practitioner can simply implement the same inference process outlined in Section 3 for each subpopulation with $X = x$ and $x \in \Omega_X$. The only requirement is that there must be a large enough sample size for each covariate-cell.

Second, when covariates are continuous and/or high-dimensional, the inference procedure must be adjusted.[23] In this case, we suggest to follow a method adopted by Dehejia and Wahba (1999)

[21] See Ura (2018) for more details of this case.

[22] One way to construct a bound for the overall treatment effect $\mathbb{E}[\alpha^{IV}(X)]$ is to take expectations of the lower and upper bounds of the identified set of $\alpha^{IV}(x)$. In the final remarks of Appendix A.3, we point out another technical challenge that one would face if the identification target was the unconditional IV estimand.

[23] We do not attempt to solve issues in inference arising from infinite dimensional covariates. By "high-dimensional", we mean a relatively large but still finite number of covariates, which may cause the curse of dimensionality when using traditional semi or nonparametric estimation

and Battistin and Sianesi (2011) which is based on the idea of stratification matching. For the sake of dimension reduction, denote $e = e(x) = \Pr(T = 1|X = x)$ as the observable propensity score of the treatment proxy T, which is an index summarizing the information contained in the covariates. We can create strata by dividing the support of $e(X)$ and grouping individuals with e in the same range. By abuse of notation, denote the conditional IV estimand as $\alpha^{IV}(e \in A_s) = \frac{\text{Cov}(Y,g(Z)|e(X)\in A_s)}{\text{Cov}(D,g(Z)|e(X)\in A_s)}$ where A_s represents one of the S strata and $\bigcup_{s=1}^{S} A_s$ is the support $e(X)$. More specifically, each of our partial identification strategies can be implemented for $\alpha^{IV}(e \in A_s)$ following three simple steps:

Step 1. Estimate $e(x)$ from a linear, logit or probit regression,[24] where polynomials and interactions of X may be included as regressors to account for possible nonlinear effects of X on the probability of being observed as treated.

Step 2. Given the estimated propensity score $\hat{e}(x)$, stratify samples into a finite number of strata over the common support of the score. These strata can be either equally spaced, or user-specified, such that the number of observations within each stratum is large enough to conduct inference.[25] This step is equivalent to converting the continuous variable $e(x)$ into a discrete one.

Step 3. Within each stratum A_s with $s = 1, ..., S$, proceed with the chosen partial identification strategy for $\alpha^{IV}(e \in A_s)$ and conduct inference following the detailed procedure outlined in Section 3.[26]

4.3 Practical Guidance

We provide guidance about: (i) how to choose among the three partial identification strategies; (ii) how to incorporate external information about the misclassification error and calculate the bounds using strategy 3; and (iii) how to obtain a point estimate of the effect, which can be used as a bias reduction method in place of a conventional IV approach.

First, the choice of which strategy to adopt depends on the information available. If no prior or external information about measurement accuracy is available, strategies 1 and 2 can be applied with the available dataset. Moreover, based on the discussion of corollary 2.4, in situations where the practitioner suspects that the value of LATM, Δp_k, does not vary much across k (at the limit, the data exhibit homogeneous misclassification), we suggest to use strategy 2. Note that, as we

methods.

[24] A practitioner could also estimate $e(x)$ non-parametrically. That is, our choice of using a parametric method is only practical and it is not required by our theoretical framework. In particular, this means that we do not require additional identifying assumptions to perform inference in the case of covariates.

[25] As explained in Lin, Tommasi, and Zhang (2021), our specific routine divides the sample into equally spaced strata. For example, if the sample is divided into 10 strata, our routine divides the sample such that, strata = 1 contains the 10% of the observations with the lowest predicted values of $e(X)$; then strata = 2 contains the next 10% of the observations; and so on until strata = 10, which contains the 10% of the observations with the highest predicted values of $e(X)$.

[26] Specifically for Strategy 3, this means to obtain first an estimate of $\alpha^{Mis}(e \in A_s)$ and its confidence interval using samples for each stratum. Then, following Equation (19), construct the confidence interval of $\alpha^{IV}(e \in A_s)$ using information on the misclassification error (see Section 4.3 for guidance).

pointed out at the end of section 2.3, if the available instrument is binary, there is no choice to make between Strategy 1 and 2 because they are exactly the same. Lastly, when there are available information about the weighted average of LATMs, ξ, and we are quite confident about the accuracy of the range $[\underline{\xi}, \overline{\xi}]$, then strategy 3 is strongly recommended.

Second, prior information about the misclassification error is useful because it helps improve the bounds of α^{IV}. The objective of strategy 3 is to combine the estimated α^{Mis} with information about the misclassification error to obtain the tightest bounds of α^{IV}. Recall, the weighted average probability of false negative is $w^n = 1 - \sum_{k=1}^{K} \gamma_k^{IV} p_{1,k}$ (this is the probability of treated individuals misclassified as untreated), of false positive is $w^p = \sum_{k=1}^{K} \gamma_k^{IV} p_{0,k}$ (this is the probability of untreated individuals misclassified as treated), and hence, by definition, $\xi = 1 - w^n - w^p$. As explained before, researchers do not need to know each value of $p_{1,k}$ and $p_{0,k}$, for $k = 1, 2, ..., K$; it is sufficient to know the overall w^p and w^n in the sample. Without loss of generality, we assume the ascending order of Ω_Z so that $[\underline{\xi}, \overline{\xi}] \subset [0, 1]$. Assumption 2.3 implies that the probability of false positive is lower than the probability of false negative for each group of compliers, C_k,: that is, $0 \leq p_{0,k} \leq 1 - p_{1,k}$, implying $0 \leq w^p \leq w^n$.

To show how to implement the third strategy, we consider four cases and illustrate how one should set $[\underline{\xi}, \overline{\xi}]$ in each scenario. For illustrative purpose, suppose $\alpha^{IV} = 1$ but the practitioner can only obtain an estimate of α^{Mis}, using a conventional 2SLS, such that $\widetilde{\alpha}^{Mis} = 1.5$ and the 95% CI is $[0.52, 2.48]$. We calculate the corresponding confidence intervals (CI) of α^{IV} using (19).

Case 1: Approximation of false negative probability. The first case mimics a context where only w^n is known. This is a common situation under poor recalling of treatment status. Then, $\xi \leq 1 - w^n$ because w^p is nonnegative, and $\xi \geq 1 - 2w^n$ because $w^p \leq w^n$. In this case, a practitioner can set $\xi \in [\max\{0, 1 - 2w^n\}, 1 - w^n]$. Assume $w^n = 0.40$. Given the aforementioned CI of α^{Mis}, the 95% CI for α^{IV} in this case would be $[0.10, 1.49]$.[27]

Case 2: Approximation of false positive probability. In the second case, suppose only w^p is known. Then, $\xi \leq 1 - 2w^p$ because $w^p \leq w^n$. In this case, a practitioner can set $\xi \in [0, 1 - 2w^p]$. Suppose $w^p = 0.05$. Given the 95% CI of α^{Mis}, the 95% CI of α^{IV} would be $[0, 2.23]$.[28]

Three remarks follow directly from these first two cases. First, if the value of a false negative w^n is larger than 0.5 (that is, more than 50% of individuals who are truly treated report to be untreated), in case 1, the lower bound of ξ should be set to zero. This would occur if the data collected are heavily contaminated by misclassification error. Second, if the only available information is the probability of a false positive, such information is generally likely to be quite weak for providing

[27] The ending points of the CI for α^{IV} are the smallest and largest points of the interval $\mathscr{C}^{\xi}(\beta^{\xi})$. The rule to find these two extremes is straightforward. Multiplying the two ending points of CI of α^{Mis} by $\underline{\xi}$ and $\overline{\xi}$ respectively, gives us four values. Then, the smallest and largest value among these four values, will be the two ending points of the CI of α^{IV}. For example, given the CI of α^{Mis}, since both its extremes are positive, the CI is $[0.52 \times 0.2, 2.48 \times 0.6] = [0.104, 1.488]$. The calculation is slightly more complicated if the CI of the α^{Mis} contains both positive and negative values. For example, suppose the practitioner uses a smaller sample size, so that, $\widetilde{\alpha}^{Mis} = 1.5$, but the 95% CI of this 2SLS estimate is $[-0.08, 3.08]$. In this case, if we apply the same rule, the CI would be calculated as follows: $\alpha^{IV} \in [-0.08 \times 0.6, 3.08 \times 0.6] = [-0.05, 1.85]$.

[28] The CI of α^{IV} is $[2.48 \times 0, 2.48 \times 0.9] = [0, 2.23]$. Whereas, if $\widetilde{\alpha}^{Mis} = 1.5$ with 95% CI $[-0.08, 3.08]$, the CI of α^{IV} would be $[(-0.08) \times 0.9, 3.08 \times 0.9] = [-0.07, 2.77]$.

an informative bound of α^{IV}. This is because, in case 2, the lower bound of ξ is zero, which fails to recover the sign of the parameter of interest. In this situation, we recommend imposing further restrictions, such as setting the maximum probability of a false negative at 0.5, leading to a narrower bound $\xi \in [\max\{0, 0.5-w^p\}, 1-2w^p]$. The latter choice should be motivated by the specific context. Finally, even if $\underline{\xi}$ and $\overline{\xi}$ are both positive, it is possible that $\Theta^{\xi}(\mathbf{P})$ may fail to recover the sign of α^{IV} in the finite sample estimation, if the CI of α^{Mis} is on both side of zero.

Case 3: Bounds of false negative and false positive probabilities. The third case mimics a context where the practitioner know the bounds of w^n and w^p: $\underline{w}^n \leq w^n \leq \overline{w}^n$ and $\underline{w}^p \leq w^p \leq \overline{w}^p$. In this case, the range of ξ can be set as $[1-\overline{w}^n-\overline{w}^p, 1-\underline{w}^n-\underline{w}^p]$. For example, let us take $0.4 \leq w^n \leq 0.5$ and $0 \leq w^p \leq 0.05$ as prior information, then $\xi \in [0.45, 0.6]$. Given the CI of α^{Mis}, the 95% CI of α^{IV} would be $[0.23, 1.49]$.[29]

Case 4: Approximations of false negative and false positive probabilities. In the last case, we mimic a situation where the practitioner has a good approximation of both w^p and w^n, which is equivalent to having a good approximation of $\xi = 1-w^n-w^p$. In this case, $\Theta^{\xi}(\mathbf{P})$ degenerates to a point $\alpha^{Mis}(1-w^n-w^p)$. Suppose $w^n = 0.50$ and $w^p = 0.05$, then $\xi = 0.45$. Given the estimate of α^{Mis}, the point estimate of α^{IV} would be 0.675 with a 95% CI of $[0.23, 1.12]$.[30]

Case 4 is particularly interesting for a practitioner because we can obtain a point estimate of α^{IV}. However, it is worth noting that, since the value of w^n and w^p are likely to be only approximations, the point estimate obtained will be biased regarding the true α^{IV}. Nevertheless, our approach can be used in place of a conventional IV estimator as a bias reduction method. Moreover, our simulation results, fully presented in the older version of our paper (Tommasi and Zhang, 2020), demonstrate that the confidence interval of the point estimates $\alpha^{Mis}(1-w^n-w^p)$ yields a desirable coverage rate of the true value of α^{IV}.

One final remark. In practice, the possible values of misclassification rates might be estimated by matching survey data to administrative records and computing the fraction of false positives and negatives. If a confidence interval of $[\underline{\xi}, \overline{\xi}]$ or of ξ is available, we can additionally adjust the confidence interval of α^{IV} by taking into account the uncertainty captured by such confidence interval of $[\underline{\xi}, \overline{\xi}]$ or ξ.[31]

4.4 Monte Carlo Simulations

In Tommasi and Zhang (2020), we use Monte Carlo simulations to illustrate the finite sample properties of the confidence intervals $C^j(\beta^j)$, with $j = \alpha, p, \xi$, proposed in Section 3. We study the performance of the three strategies for practical applications, hence we compute the simplified version of the confidence intervals of $\alpha_{k,k-1}$ and Δp_k as in (A8). Based on this, the confidence intervals

[29] Whereas, if $\widetilde{\alpha}^{Mis} = 1.5$, but the 95% CI of this 2SLS estimate is $[-0.08, 3.08]$, the CI of α^{IV} would be again $[-0.048, 1.85]$.
[30] Whereas, if $\widetilde{\alpha}^{Mis} = 1.5$, but the 95% CI of this 2SLS estimate is $[-0.08, 3.08]$, the CI of α^{IV} would be $[-0.036, 1.39]$.
[31] More details can be found in the proof of Corollary 3.1 in Appendix A.1.6.

of α^{IV} are constructed in the same manners as in (17), (18) and (19). We extensively explore the sensitivity of the bounds along three dimensions: (i) strength of the instrumental variable, (ii) extent of the misclassification error, and (iii) external information. Overall, the conclusion is that our partial identification strategies represent a reliable alternative approach when practitioners can only use a mismeasured binary treatment T in place of D to estimate the benefits of a program. Moreover, our approach becomes very powerful and works best when external information about the accuracy of the measurement error can be considered.

4.5 Examples of Misreported Treatment

When doing applied work, economists are persistently challenged by mismeasured binary endogenous variables. The papers that we present in the following literature review highlight the value of applying our estimator when facing a mismeasured treatment.

Misreporting is an important and common feature of population surveys that collect socio-economic data. For example, it is no surprise that we observe misreporting when survey participants are asked to disclose illegal or shameful activities. Domestic violence and alcohol intake are two such areas that suffer from this phenomenon. More precisely, Alderman et al. (2013) found that there is a high mistreatment of women's outcomes and children's development because victims do not truthfully report domestic violence. Similarly, Agüero and Frisancho (2020) analysed a Peruvian data-set of the Demographic and Health Surveys and found a measurement error in responses to direct questions.[32] In particular, up to 30% of women underreported physical and sexual violence by their intimate partners. Palermo et al. (2014) extended this line of enquiry by exploring data from 24 developing countries and found that only seven percent of domestic violence victims actually made an official report. Whereas, Molinari (2010) points out that, in the National Longitudinal Survey on Youth, a survey which asks pregnant women to comment on their alcohol intake, there was an item nonresponse rate of 6 to 14% in 1984. Such a low response rate compounds the oversight of social problems, in this case, the effect of alcohol on birth outcomes. These low response rates are likely to affect causal analysis of a particular program aimed at reducing children's negative health outcomes.

Turning to the impact of misreporting on divorce and marriage statistics, self-reported divorce rates can be anywhere between 8% and 25% less than those held by government agencies, as highlighted by O'Connell (2006) and Mitchell (2010). In particular, O'Connell (2006) documents the impact of missing data by highlighting that up to 20% of people surveyed only partially reported on the required information in the Survey of Income and Program Participation. Such low reporting rates makes it difficult to predict social trends, such as single-parent families, and subsequently it makes it difficult for policy-makers to devise tailored interventions.

Given the effect of shame and distrust on self-reporting on domestic violence, alcohol-intake, illegal activity and divorce, it is not surprising that survey respondents inaccurately report on their

[32] This is a global data collection effort comprising 122 surveys in 61 developing countries.

physical body size (and in particular their weight), as well as access to food stamps. For instance, Zhang et al. (2016) make this observation as it relates to their analysis of the Consortium on the Safe Labor Survey. Meyer et al. (2015) examined this phenomenon by looking at the degree to which people are likely to disclose receipt of a government transfer. The Current Population Survey revealed nonresponse rates of 16–20% and the National Health Interview Survey had nonresponse rates of 24%. They also compared household survey data with official administrative data[33] and found that many survey respondents do not report receiving food stamps.[34]

These are a few examples where the estimator that we have designed can serve as the primary identification strategy as well as a robustness check for causal inference. However, there is a variety of other economic applications where the problem of misreporting of the treatment variable has been established, including the misreporting of: union status (Card, 1996), participation to trainings (Barron et al., 1997), coverage of health insurance (Black et al., 2000), language fluency (Dustmann and Soest, 2001), self-evaluation of health-related status (Crossley and Kennedy, 2002), educational attainment (Black et al., 2003), chemical emissions by firms (Marchi and Hamilton, 2006), disability status (Kreider and Pepper, 2007), types of corporate governance structure (Almeida et al., 2010), school meals (Gundersen et al., 2012), dental insurance (Kreider et al., 2015), firm's formality status (Gandelman and Rasteletti, 2017), and technology adoption (Wossen et al., 2018).

4.6 Application to the 401(k) Pension Plan

In this section, we use our method to measure the benefits of participating in the 401(k) pension plan on savings. The 401(k) pension plan is one of the most popular defined contribution retirement plans in the US. It aims at increasing financial savings through the tax deductibility of contributions to retirement accounts. Although the effects of this plan have been examined elsewhere (e.g., Abadie, 2003; Ura, 2018), the application contains all the ingredients to demonstrate the full extent of the usefulness of the approach proposed in our paper.

First, the participation to the program is binary and notoriously misreported in survey data. Second, the eligibility to the pension plan, which is provided only to workers in firms offering the plan, is arguably a valid instrument (e.g., Poterba et al., 1995). Third, the eligibility can be interacted with the year of introduction of the plan, which yields a discrete instrument that accounts for the duration of the exposure to the plan. Fourth, credible information on treatment misclassification probabilities are available from the literature and can be incorporated in estimation. Fifth, although our main theoretical results hold without covariates, including covariates is almost always crucial in application. In our specific context, for the instrument to be valid, it is really important to condition on family income and age. Hence, this specific application gives us also the opportunity to show the performance of our proposed suggestions to incorporate covariates in a realistic context. Finally,

[33] This is the Supplemental Nutrition Assistance Program (SNAP) or Food Stamp Program.

[34] 23% in the Survey of Income and Program Participation (SIPP), 35% in the American Community Survey, and 50% in the Current Population Survey (CPS).

the fact that the application is well known makes it easier for us to evaluate our results in light of the existing literature.

Given the three main contributions of the paper, we aim to answer the following questions: (i) What is the likely bias of the estimated program benefits if we do not account for treatment misclassification? (ii) In case of a binary instrument, how do the bounds of the program benefits shrink by incorporating external information on misclassification probabilities? How do they compare with the results of the existing literature? (iii) In case of a discrete instrument, how are the bounds (for each chosen strata) compared to a naive approach that does not account for treatment misclassification? How do these bounds shrink by incorporating external information on misclassification probabilities?

In measuring the benefits of the 401(k) pension plan, a researcher would face two main difficulties: endogenous participation in the plan and misreporting of participation. The first problem may arise due to unobserved differences in saving behaviors. That is, participants in the plan might save more in general than those who do not participate. Hence, a comparison of accumulated financial assets between participants and nonparticipants is likely to yield a positive bias of the true effect of the program. If this was the only problem in the data, a practitioner could just use the eligibility to the plan as a valid instrument and perform inference on the causal parameter as in Abadie (2003). However, the contemporaneous presence of the second problem makes the task difficult. Misreporting in this context may arise because individuals find it difficult to remember or understand their pension plan, leading to reporting error. Indeed, Gustman et al. (2007) documented that about one-fourth of respondents to the Health and Retirement Study (HRS) misreport their pension plan. Further, Dushi and Iams (2010) documented that in the SIPP, over 17% of participants in the 401(k) pension plan self-report as nonparticipants (false negative) and almost 10% of nonparticipants self-report as participants (false positive). Understanding plan benefits is relevant for the economic well-being of future retirees because these plans are important for retirement income security. This is the economic motivation underlying our efforts.

We use data from the SIPP round from 1991. The construction of the dataset and the choice of the covariates follows the work by Abadie (2003). Hence, our sample only includes households where at least one person is employed and has no income from self-employment. Moreover, the sample is restricted to individuals with an annual family income between $10,000 to $200,000, because eligibility for the plan is rare outside this range. Table 2 reports the summary statistics of the main variables used in the analysis. The average family net financial assets (outcome Y) is around $19,000. Roughly 27% of the observations report participating in the 401(k) pension plan (misreported treatment T), whereas 39% are eligible for the plan (instrument Z). The set of covariates, X, includes a constant, family income, age, age squared, marital status, and family size. The resulting sample size is 9,275.

First contribution. Given the available information regarding treatment misclassification probabilities, a researcher can use our new relationship between the true and mismeasured treat-

Table 2: Summary statistics

Variable	Mean	Standard deviation	Minimum	Maximum
Family net financial assets	19.0	63.9	-0.5	1,536
Participation to 401(k)	0.276	0.447	0	1
Eligibility to 401(k)	0.392	0.4356	0	1
Family income	39.2	24.1	10.0	199.0
Age	41.1	10.3	25	64
Family size	2.9	1.5	1	13

Notes: The Table reports the mean, standard deviation, minimum and maximum values of the main variables used in the paper. There is a total of 9,275 observations. The average family net financial assets (in 1,000$ units) is the outcome Y, the participation to the 401(k) pension plan is the misreported treatment T, whereas the eligibility to the plan is the instrument Z. The set of covariates X includes a constant, family income (in 1,000$ units), age, age squared, marital status and family size.

ment effect, Equation (4), to approximate the possible level of biases of the benefits of the 401(k) plan. In our case, $w^n = 17\%$ and $w^p = 10\%$, which means that the estimated (mismeasured) treatment effect reported in the literature is likely biased (upward) by approximately 37%. A similar approximation could be easily calculated for any program mentioned in Section 4.5, provided credible information regarding treatment misclassification probabilities.

Second contribution. We proceed by estimating the bounds of the unconditional IV estimand $E[\alpha^{IV}(X)]$ in case of a binary instrument. Panel A of Table 3 reports the results. Column (1) reports the conventional 2SLS estimate (assuming homogenous treatment effect) as shown in column (3) of Table 2 by Abadie (2003). This represents a biased point estimate because it ignores the potential treatment misclassification. The effect is statistically significant and says that participating in the 401(k) plan increases the total financial assets by roughly $9,400, with a 95% confidence interval of $5–13,000. Column (2) reports the estimate of mismeasured treatment effect following Frölich (2007) nonparametric approach to incorporate covariates.[35] This accounts for treatment effect heterogeneity, while ignoring the potential treatment misclassification. Next, Column (3) displays the 95% confidence interval for the unconditional IV estimand from Ura (2018) which accounts for the misclassification error of the treatment variable. This is our benchmark result from the literature, to which we compare the performance of our partial identification strategies.

Columns (4)–(8) report the 95% confidence interval of our partial identification strategies under different assumptions about the misclassification probabilities. The estimation error of the nuisance parameters $\pi(X) = \Pr(Z = 1|X)$ and $\mathbb{E}[\alpha^{Mis}(X)]$ are taken into account following the inference process in Section 3, where their confidence intervals are obtained by nonparametric bootstrapping. Column (4) assumes no information about the misclassification probabilities. Since the instrumental variable is binary, strategy 1 and 2 coincide and are equivalent to the method developed by Ura (2018). Column (5)-(8) use external information about the misclassification probabilities. In particular, Column (5) assumes that we know an approximation of the probability of false negative ($w^n = 17\%$) (Case 1 of Section 4.3); Column (6) assumes that we know an approximation of

[35] In this case, $\mathbb{E}[\alpha^{Mis}(X)]$ is calculated by $\mathbb{E}\left[\frac{Z-\pi(X)}{\pi(X)(1-\pi(X))}Y\right]/\mathbb{E}\left[\frac{Z-\pi(X)}{\pi(X)(1-\pi(X))}T\right]$, where $\pi(X) = \Pr(Z = 1|X)$ is estimated via a linear probability model. The confidence interval of $\mathbb{E}[\alpha^{Mis}(X)]$ is computed using a nonparametric bootstrap.

the probability of false positive ($w^p = 10\%$) (Case 2 of Section 4.3); Column (7) assumes that we know the bounds of these probabilities (assuming the probability of false negative is higher than the probability of false positive, we get $10\% \leq w^n \leq 17\%$ and $w^p = 10\%$) (Case 3 of Section 4.3). As one can see, using our partial identification strategies in the presence of external information, we obtain bounds of the true benefits that can be up to 36% narrower compared to Column (3). When both w^n and w^p are approximately known, as in this application, our approach can deliver a point estimate of the effect, which is reported in Column (8) (Case 4 of Section 4.3). Since these are both approximations of the misclassification probabilities, this point estimate is likely to be biased. However, it is closer to the true (unknown) effect than the value reported in Column (2), which ignores treatment misclassification.

Table 3: Empirical Illustration

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Binary instrument							
$\mathbb{E}[\alpha^{Mis}(X)]$		**Target parameter: unconditional LATE** $\mathbb{E}[\alpha^{IV}(X)]$					
2SLS Abadie (2003)	nonpara.	Ura (2018)	Strategy 1 ≡ 2	Strategy 3			
				appr. w^n	appr. w^p	bounds w^n and w^p	appr. w^n and w^p
9.4 (5.3, 13.5)	16.3 (6.0, 27.6)	(4.4, 28.3)	(4.3, 27.8)	(4.7, 21.2)	(0, 20.4)	(5.2, 20.4)	11.9 (5.2, 18.6)
Panel B: Discrete instrument							
	$\alpha^{Mis}(e \in A_s)$	**Target parameter: conditional WLATE** $\alpha^{IV}(e \in A_s)$					
		Strategy 1	Strategy 2	Strategy 3			
				appr. w^n	appr. w^p	bounds w^n and w^p	appr. w^n and w^p
Strata 1	21.8 (16.3, 27.3)	(2.5, 42.4)	(2.9, 29.4)	(11.2, 23.0)	(0, 22.1)	(12.2, 22.1)	15.9 (12.2, 20.2)
Strata 2	23.1 (19.2, 27.0)	(2.3, 70.1)	(4.6, 28.2)	(12.7, 22.4)	(0, 21.6)	(14.0, 21.6)	16.9 (14.0, 19.7)
Strata 3	54.5 (44.3, 64.8)	(19.2, 120.9)	(15.5, 68.2)	(29.6, 53.2)	(0, 51.2)	(32.8, 51.2)	39.8 (32.8, 46.7)

Notes: Results in this Table are in 1,000$ units. Confidence interval is in parentheses. Panel A reports the results using a binary IV. Panel B reports illustrative results using a discrete IV. In Panel A, Column (1) reports the conventional 2SLS estimates as shown in column (3) of Table 2 by Abadie (2003). Column (2) reports the mismeasured treatment effect taking unobserved heterogeneity into account following Frölich (2007). Column (3) reports the best 95% CI of the LATE as shown in Table 2 by Ura (2018). Column (4)-(7) report the 95% CI of our partial identification strategies under different assumptions regarding the misclassification probabilities. Finally, Column (8) delivers a point estimate of the effect. In Panel B, our target parameter is the conditional IV estimand. We stratify the samples into three strata based on their estimated $e = e(X) = \Pr(T = 1|X)$. Column (2) reports the results of $\alpha^{Mis}(e \in A_s)$ for each stratum A_s and $s = 1, 2, 3$. Column (3)-(7) report the 95% CI using our partial identification strategies under different assumptions regarding the misclassification probabilities. Finally, Column (8) delivers a point estimate of the effect.

Third contribution. Finally, we illustrate the performance of our method when the instrument is discrete by interacting the eligibility for the 401(k) plan and the duration of exposure to the plan. The duration of exposure is defined as how many years one has been exposed to the 401(k) program, which became active in 1981. Those with less than 10 years of exposure were 15 to 24 years old in 1981. Those with at least 10 years of exposure were 25 or older in 1981. The discrete instrument takes the value $Z = 0$ if an individual is not eligible and has been exposed for less than 10 years,

$Z = 1$ they are eligible and have less than 10 years of exposure or are ineligible and have at least 10 years of exposure, and $Z = 2$ if they are eligible and have at least 10 years of exposure. Naturally, with this instrument the ascending order requirement is satisfied.

When the instrument is discrete (or multiple-discrete), our target parameter is the conditional IV estimand as discussed in Section 4.2. Hence, for this case, we stratify the sample into three strata A_1, A_2, A_3, based on their estimated probability of self-reported participation $e = e(X)$ =Pr$(T = 1|X)$.[36] Each stratum consists of one-third of the samples, where samples in strata 1 have the smallest $e(X)$, and samples in strata 3 have the largest $e(X)$. We proceed with the estimation as in Panel A within each stratum. Panel B of Table 3 reports the results, which indicate that, without accounting for treatment misclassification, $\alpha^{Mis}(e \in A_s)$ overestimates the true effect of the program. Indeed, as one can notice, the lower bounds for $\alpha^{IV}(e \in A_s)$ of each estimate in Column (3)-(8) are much smaller than the left-end point of the 95% confidence interval of $\alpha^{Mis}(e \in A_s)$ in Column (2). In addition, as mentioned in Section 2.3 (and explored in simulation), strategy 2 produces a more informative bound than strategy 1; the lower bounds of strategy 1 and 2 are relatively comparable, while the bounds of strategy 2 are considerably tighter than those of strategy 1. Similar to the binary instrument case, in the discrete instrument case, and for each stratum, the more informative external information is incorporated in estimation, the better is the performance of strategy 3 in terms of tightest of the bounds.

5 Conclusion

In the evaluation of treatment effects, endogenous participation is often misreported in survey data. When treatment is binary, using a standard instrumental variable method would lead to biased estimates. Even with infrequent arbitrary errors in the binary treatment indicator, the bias can be severe. In this paper, we focus on the local average treatment effect (LATE) or the weighted average of LATEs (WLATE), which are parameters that can be estimated to measure the effects of a treatment in case of noncompliance. We start by showing the limitations of the standard LATE approach when the binary treatment is a mismeasured proxy of the true treatment and derive a simple relationship between the true and mismeasured treatment effects. This link is mediated by a new parameter, defined in terms of the misclassification probabilities, which can be used to approximate the possible level of bias of the estimated benefits of a program. Then, we provide three partial identification strategies to bound the LATE or WLATE and to further tighten the bounds using external information about misclassification probabilities.

Overall, this article shows that researchers who aim to measure treatment effects with a misclassified binary treatment can obtain bounds of the LATE or the WLATE. These bounds can potentially be tight, provided accurate information about the extent of misreporting in survey data can be obtained. This information can be accessed from treated individuals' administrative records, which are

[36] A practitioner can choose any number of strata. We choose three only for illustration. Results with higher number of strata do not change our analysis and are available upon request.

becoming increasingly available. In applications where this information is unavailable, one could also rely on small validation studies, or repeated measurements of the same individual, to retrieve useful information. Our main conclusion is that the proposed method is universally applicable as the leading identification strategy, or the leading robustness check, in any setting where the practitioner suspects that the endogenous binary treatment is not well measured and binary, discrete or multiple-discrete-instrument(s) are available.

References

ABADIE, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263. [20], [26], [27], [28], [29], [17]

ACERENZA, S., K. BAN, AND D. KÉDAGNI (2021): "Marginal Treatment Effects with Misclassified Treatment," Tech. rep. [5]

AGÜERO, J. M. AND V. FRISANCHO (2020): "Measuring Violence Against Women with Experimental Methods," Tech. rep. [25]

AIGNER, D. J. (1973): "Regression with a binary independent variable subject to errors of observation," *Journal of Econometrics*, 1, 49 – 59. [4]

ALDERMAN, H., J. DAS, AND V. RAO (2013): "Conducting ethical economic research: complications from the field," *World Bank Policy Research Working Paper*. [25]

ALMEIDA, H., M. CAMPELLO, AND J. GALVAO, ANTONIO F. (2010): "Measurement Errors in Investment Equations," *The Review of Financial Studies*, 23, 3279–3328. [26]

ANGRIST, J. AND I. FERNANDEZ-VAL (2013): "ExtrapoLATE-ing: External validity and overidentification in the late framework," in *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*, Cambridge University Press, vol. 51, 401. [20], [21]

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. [21]

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. [4]

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]

BARRON, J. M., M. C. BERGER, AND D. A. BLACK (1997): "How Well Do We Measure Training?" *Journal of Labor Economics*, 15, 507–528. [26]

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], [10]

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. [3], [8], [9], [10], [22]

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2008): "Treatment effect bounds under monotonicity assumptions: An application to swan-ganz catheterization," *American Economic Review*, 98, 351–56. [14]

BILLINGSLEY, P. (2008): *Probability and measure*, John Wiley & Sons. [10]

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. [4], [26]

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. [4], [26]

BOLLINGER, C. R. (1996): "Bounding mean regressions when a binary regressor is mismeasured," *Journal of Econometrics*, 73, 387 – 399. [4], [10]

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. [4]

BRINCH, C. N., M. MOGSTAD, AND M. WISWALL (2017): "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 125, 985–1039. [5]

CALVI, R., A. LEWBEL, AND D. TOMMASI (2018): "Women's empowerment and family health: Estimating LATE with mismeasured treatment," *Available at SSRN 2980250*. [2], [8], [9]

CARD, D. (1996): "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 64, 957–979. [26]

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. [4]

CHEN, X., C. A. FLORES, AND A. FLORES-LAGUNES (2017): "Going beyond LATE: Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*. [5]

——— (2018): "Going beyond LATE Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*, 53, 1050–1099. [14]

CHEN, X., H. HONG, AND D. NEKIPELOV (2011): "Nonlinear models of measurement errors," *Journal of Economic Literature*, 49, 901–37. [10]

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2019): "Inference on causal and structural parameters using many moment inequalities," *The Review of Economic Studies*, 86, 1867–1900. [3], [18], [4], [5], [19]

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81, 667–737. [15], [17]

CROSSLEY, T. F. AND S. KENNEDY (2002): "The reliability of self-assessed health status," *Journal of Health Economics*, 21, 643 – 658. [26]

DEATON, A. (2010): "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48, 424–55. [5]

DEHEJIA, R. H. AND S. WAHBA (1999): "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American statistical Association*, 94, 1053–1062. [21]

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]

DUSHI, I. AND H. M. IAMS (2010): "The impact of response error on participation rates and contributions to defined contribution pension plans," *Social Security Bulletin*, 70, 45—60. [3], [27]

DUSTMANN, C. AND A. V. SOEST (2001): "Language Fluency and Earnings: Estimation with Misclassified Language Indicators," *The Review of Economics and Statistics*, 83, 663–674. [26]

FRAZIS, H. AND M. A. LOEWENSTEIN (2003): "Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables," *Journal of Econometrics*, 117, 151 – 178. [8], [9]

FRÖLICH, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35–75. [20], [21], [28], [29], [16], [17]

GANDELMAN, N. AND A. RASTELETTI (2017): "Credit constraints, sector informality and firm investments: evidence from a panel of Uruguayan firms," *Journal of Applied Economics*, 20, 351 – 372. [26]

GUNDERSEN, C., B. KREIDER, AND J. PEPPER (2012): "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis," *Journal of Econometrics*, 166, 79 – 91, annals Issue on "Identification and Decisions", in Honor of Chuck Manski's 60th Birthday. [26]

GUSTMAN, A. L., T. STEINMEIER, AND N. TABATABAI (2007): "Imperfect knowledge of pension plan type," Working Paper 13379, National Bureau of Economic Research. [27]

HAUSMAN, J., J. ABREVAYA, AND F. SCOTT-MORTON (1998): "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 87, 239 – 269. [8], [9], [10]

HAUSMAN, J. A., W. K. NEWEY, H. ICHIMURA, AND J. L. POWELL (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50, 273–295. [20]

HECKMAN, J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. [5]

HECKMAN, J. J. AND S. URZÚA (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156, 27 – 37, structural Models of Optimization Behavior in Labor, Aging, and Health. [5]

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. [4]

HU, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144, 27 – 61. [4], [10]

IMAI, K. AND T. YAMAMOTO (2010): "Causal Inference with Differential Measurement Error: Nonparametric Identification and Sensitivity Analysis," *American Journal of Political Science*, 54, 543–560. [3]

IMBENS, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423. [5]

——— (2014): "Instrumental Variables: An Econometrician's Perspective," *Statist. Sci.*, 29, 323–358. [4]

IMBENS, G. W. AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [2], [6], [11], [36]

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. [4]

KLEPPER, S. (1988): "Bounding the effects of measurement error in regressions involving dichotomous variables," *Journal of Econometrics*, 37, 343 – 359. [4]

KLINE, P. AND C. R. WALTERS (2019): "On heckits, LATE, and numerical equivalence," *Econometrica*, 87, 677–696. [5]

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., R. J. MANSKI, J. MOELLER, AND J. PEPPER (2015): "The Effect of Dental Insurance on the Use of Dental Care for Older Adults: A Partial Identification Analysis," *Health Economics*, 24, 840–858. [26]

KREIDER, B. AND J. V. PEPPER (2007): "Disability and employment: Reevaluating the evidence in light of reporting errors," *Journal of the American Statistical Association*, 102, 432–441. [3], [15], [17], [26]

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], [3]

LEWBEL, A. (2007): "Estimation of average treatment effects with misclassification," *Econometrica*, 75, 537–551. [4], [8], [9], [10]

LIN, A., D. TOMMASI, AND L. ZHANG (2021): "Bounding Heterogeneous Treatment Effects With Endogenous and Misreported Treatment," Tech. rep., Working paper. [1], [4], [22], [6]

MAHAJAN, A. (2006): "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74, 631–665. [4]

MANSKI, C. (2007): *Identification for Prediction and Decision*, Harvard University Press. [5]

MANSKI, C. F. (1997): "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1311–1334. [14]

MANSKI, C. F. AND J. V. PEPPER (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. [14]

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12, S200–S216. [14]

MARCHI, S. D. AND J. T. HAMILTON (2006): "Assessing the Accuracy of Self-Reported Data: an Evaluation of the Toxics Release Inventory," *Journal of Risk and Uncertainty*, 32, 57–76. [26]

MEYER, B. D. AND N. MITTAG (2019a): "Misreporting of Government Transfers: How Important are Survey Design and Geography?" *Southern Economic Journal*. [2], [3]

——— (2019b): "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net," *American Economic Journal: Applied Economics*, 11, 176–204. [2], [3]

MEYER, B. D., N. MITTAG, AND R. M. GEORGE (2020): "Errors in survey reporting and imputation and their effects on estimates of food stamp program participation," *Journal of Human Resources*, 0818–9704R2. [2], [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], [26]

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]

MITCHELL, C. (2010): "Are Divorce Studies Trustworthy? The Effects of Survey Nonresponse and Response Errors," *Journal of Marriage and Family*, 72, 893–905. [25]

MOGSTAD, M., A. SANTOS, AND A. TORGOVITSKY (2018): "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 86, 1589–1619. [5]

MOGSTAD, M., A. TORGOVITSKY, AND C. R. WALTERS (2020a): "Policy evaluation with multiple instrumental variables," Tech. rep., National Bureau of Economic Research. [5], [6]

——— (2020b): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*. [3], [6]

MOLINARI, F. (2010): "Missing Treatments," *Journal of Business & Economic Statistics*, 28, 82–95. [3], [25]

——— (2020): "Microeconometrics with partial identification," *Handbook of econometrics*, 7, 355–486. [13]

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

O'CONNELL, M. (2006): "The Visible Hand: Editing Marital-History Data from Census Bureau Surveys," in *Handbook of Measurement Issues in Family Research*, ed. by S. L. Hofferth and L. M. Casper, The address of the publisher: Mahwah, NJ: Lawrence Erlbaum, chap. 9. [25]

PALERMO, T., J. BLECK, AND A. PETERMAN (2014): "Tip of the iceberg: reporting and gender-based violence in developing countries," *American journal of epidemiology*, 179, 602–612. [25]

POSSEBOM, V. (2021): "Crime and Mismeasured Punishment: Marginal Treatment Effect with Misclassification," . [5]

POTERBA, J. M., S. F. VENTI, AND D. A. WISE (1995): "Do 401(k) contributions crowd out other personal saving?" *Journal of Public Economics*, 58, 1 – 32. [26]

SŁOCZYŃSKI, T. (2020): "When Should We (Not) Interpret Linear IV Estimands as LATE?" *arXiv preprint arXiv:2011.06695*. [6]

STEPHENS JR, M. AND T. UNAYAMA (2019): "Estimating the impacts of program benefits: Using instrumental variables with underreported and imputed data," *Review of Economics and Statistics*, 101, 468–475. [8], [9]

TOMMASI, D. AND L. ZHANG (2020): "Bounding Program Benefits When Participation Is Misreported," *IZA Discussion Paper*. [20], [24]

URA, T. (2018): "Heterogeneous treatment effects with mismeasured endogenous treatment," *Quantitative Economics*, 9, 1335–1370. [2], [4], [11], [12], [13], [17], [21], [26], [28], [29], [38], [39]

VUONG, Q. AND H. XU (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. [5]

WOSSEN, T., T. ABDOULAYE, A. ALENE, P. NGUIMKEU, S. FELEKE, I. Y. RABBI, M. G. HAILE, AND V. MANYONG (2018): "Estimating the Productivity Impacts of Technology Adoption in the Presence of Misclassification," *American Journal of Agricultural Economics*, 101, 1–16. [26]

YANAGI, T. (2019): "Inference on local average treatment effects for misclassified treatment," *Econometric Reviews*, 38, 938–960. [2]

ZHANG, Z., W. LIU, B. ZHANG, L. TANG, AND J. ZHANG (2016): "Causal inference with missing exposure information: Methods and applications to an obstetric study," *Statistical methods in medical research*, 25. [26]

A Appendix

A.1 Proof of Theorem 2.1

Proof of Theorem 2.1. Assumption 2.2-(ii) guarantees that the denominator of α^{Mis} is nonzero and thus α^{Mis} is well-defined. Consider the denominator of α^{Mis} in equation (2),

$$
\begin{aligned}
&\mathbb{E}[T(g(Z)-\mathbb{E}[g(Z)])] \\
=&\sum_{l=0}^{K}\mathbb{E}\Big[T\Big|Z=z_l\Big](g(z_l)-\mathbb{E}[g(Z)])\,\pi_l \\
=&\sum_{l=0}^{K}\Big[\mathbb{E}(T|Z=z_0)+\mathbb{E}(T|Z=z_1)-\mathbb{E}(T|Z=z_0)+\ldots \\
&\qquad +\mathbb{E}(T|Z=z_l)-\mathbb{E}(T|Z=z_{l-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\,\pi_l \\
=&\sum_{l=0}^{K}\mathbb{E}(T|Z=z_0)(g(z_l)-\mathbb{E}[g(Z)])\pi_l+\sum_{l=0}^{K}\sum_{k=1}^{l}\Big[\mathbb{E}(T|Z=z_k)-\mathbb{E}(T|Z=z_{k-1})\Big](g(z_l)-\mathbb{E}[g(Z)])\pi_l \\
=&\sum_{k=1}^{K}\Big[\mathbb{E}(T|Z=z_k)-\mathbb{E}(T|Z=z_{k-1})\Big]\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l. \qquad \text{(A1)}
\end{aligned}
$$

For any $z_l, z_w \in \Omega_Z$, by the definition of T and Assumption 2.2-(i) (extended unconfoundedness),

$$
\begin{aligned}
&\mathbb{E}(T|Z=z_l)-\mathbb{E}(T|Z=z_w) \\
=&\mathbb{E}[T_0+D_l(T_1-T_0)|Z=z_l]-\mathbb{E}[T_0+D_w(T_1-T_0)|Z=z_w] \\
=&\mathbb{E}[(D_l-D_w)(T_1-T_0)] \\
=&\mathbb{E}[T_1-T_0|D_l-D_w=1]\Pr(D_l-D_w=1)-\mathbb{E}[T_1-T_0|D_l-D_w=-1]\Pr(D_l-D_w=-1).
\end{aligned}
$$

Due to Assumption 2.1-(iv) (monotonicity), it is either that $D_l \geq D_w$ and $\Pr(D_l-D_w=-1)=0$, or $D_l \leq D_w$ and $\Pr(D_l-D_w=1)=0$. Since z_k is ordered such that $P(z_{k-1}) \leq P(z_k)$, we have that $D_{k-1} \leq D_k$[37]. Therefore,

$$
\mathbb{E}(T|Z=z_k)-\mathbb{E}(T|Z=z_{k-1})=\mathbb{E}(T_1-T_0|C_k)\Pr(C_k). \qquad \text{(A2)}
$$

Plug (A2) into (A1), we get

$$
\mathbb{E}[T(g(Z)-\mathbb{E}[g(Z)])]=\sum_{k=1}^{K}\mathbb{E}(T_1-T_0|C_k)\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l. \qquad \text{(A3)}
$$

For the numerator of equation (2), using the same proof of Imbens and Angrist (1994), we have

$$
\mathbb{E}[Y(g(Z)-\mathbb{E}[g(Z)])]=\sum_{k=1}^{K}\alpha_{k,k-1}\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l. \qquad \text{(A4)}
$$

Thus, based on equation (A3) and (A4), the mismeasured LATE is:

$$
\alpha^{Mis}=\frac{\sum_{k=1}^{K}\alpha_{k,k-1}\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l}{\sum_{k=1}^{K}\mathbb{E}(T_1-T_0|C_k)\Pr(C_k)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l}=\sum_{k=1}^{K}\gamma_k^{Mis}\alpha_{k,k-1}. \qquad \text{(A5)}
$$

[37] In discrete IV setting, $\Pr(z_{k-1}) \leq \Pr(z_k)$ implies $D_{k-1} \leq D_k$ can be simply proved by contradiction under Assumption 2.1-(ii) that $Z \perp (Y_1, Y_0, D_k)$ for $k=0,1,\ldots,K$.

□

A.2 Proof of Corollary 2.1

Proof of Corollary 2.1. For $\forall k$, by definitions of γ_k^{IV} and γ_k^{Mis} we have:

$$\frac{\gamma_k^{IV}}{\gamma_k^{Mis}} = \sum_{k=1}^{K} \frac{\Pr(C_k)\sum_{l=k}^{K}\pi_l\left(g(z_l)-\mathbb{E}[g(Z)]\right)}{\sum_{m=1}^{K}\Pr(C_m)\sum_{l=m}^{K}\pi_l\left(g(z_l)-\mathbb{E}[g(Z)]\right)} \times (p_{1,k}-p_{0,k}) = \sum_{k=1}^{K}\gamma_k^{IV}(p_{1,k}-p_{0,k}). \tag{A6}$$

□

A.3 Proof of Lemma 2.1

Proof of Lemma 2.1. Based on Assumption 2.1-(iv) and the fact that Ω_Z is finite, we know that there always exists an order of the possible values of Z, ordered as $z_0, z_1, ..., z_K$, such that $D_0 \leq D_1 \leq ... \leq D_K$ holds for all individuals. It then follows that $\Pr(z_0) \leq \Pr(z_1) \leq ... \leq \Pr(z_K)$. In the rest of the proof, we maintain such an order of $z_0, z_1, ..., z_K$. It suffices to show rest of Lemma 2.1 by verifying $\mathbb{E}(T|Z=z_l) \leq \mathbb{E}(T|Z=z_w)$ implies $\Pr(z_l) \leq \Pr(z_w)$ for $\forall l, w \in \{0,1,...,K\}$.

First, we show that Assumption 2.3 implies $\mathbb{E}[T_1 - T_0|C_k] > 0$ for all k. Conditional on C_k, we have that $D = 1[Z \geq z_k]$. Since Z is independent of $(\{D_k\}_{k=0}^K, T_1, T_0)$, we have for all k,

$$\begin{aligned} p_{1,k} - p_{0,k} = \mathbb{E}[T_1 - T_0|C_k] =& \mathbb{E}[T_1|C_k, Z=z_k] - \mathbb{E}[T_0|C_k, Z=z_{k-1}] \\ =& \mathbb{E}[T_1|C_k, D=1] - \mathbb{E}[T_0|C_k, D=0] \\ =& \mathbb{E}[T|C_k, D=1] - \mathbb{E}[T|C_k, D=0] \\ =& \Pr(T=1|C_k, D=1) - \Pr(T=1|C_k, D=0) \\ >& 0, \end{aligned}$$

where the last line is due to Assumption 2.3.

Next, we show that for any $l, w \in \{0,1,...,K\}$ and $w \neq l$, $\mathbb{E}[T_1 - T_0|C_k] > 0$ implies $\mathbb{E}[T_1 - T_0|D_w - D_l = 1] > 0$ if $w > l$ and $\mathbb{E}[T_1 - T_0|D_w - D_l = -1] > 0$ if $w < l$. If given $D_w - D_l = 1$, it is apparent that $D_w = 1$, $D_l = 0$ $(l < w)$ and this individual belongs to one of the complier groups $C_{l+1}, C_{l+2}, ..., C_w$; on the other hand, if conditional on $D_w - D_l = -1$, then we have $D_w = 0$, $D_l = 1$ $(l > w)$ and this individual belongs to one of the complier groups $C_{w+1}, C_{w+2}, ..., C_l$. For the case $w > l$, by the law of iterated expectation,

$$\begin{aligned} \mathbb{E}[T_1 - T_0|D_w - D_l = 1] =& \sum_{k=l+1}^{w} \mathbb{E}[T_1 - T_0|C_k, D_w - D_l = 1]\Pr(C_k|D_w - D_l = 1) \\ =& \sum_{k=l+1}^{w} \mathbb{E}[T_1 - T_0|C_k]\Pr(C_k|D_w - D_l = 1) \\ >& 0, \end{aligned}$$

where the last inequality is due that there exists at least one $k \in \{l+1, ..., w\}$ such that $\Pr(C_k|D_w - D_l = 1) > 0$. Similar arguments can be applied to show that $\mathbb{E}[T_1 - T_0|D_w - D_l = -1] > 0$ if $w < l$.

In the rest of this proof, we show that $\mathbb{E}(T|Z=z_l) \leq \mathbb{E}(T|Z=z_w)$ implies $\Pr(z_l) \leq \Pr(z_w)$. For any $l, w \in \{0,1,...,K\}$, because of the monotonicity assumption, we have

$$\mathbb{E}(T|Z=z_w) - \mathbb{E}(T|Z=z_l) = \begin{cases} \mathbb{E}[T_1 - T_0|D_w - D_l = 1]\Pr(D_w - D_l = 1), & \text{if } w > l \\ -\mathbb{E}[T_1 - T_0|D_w - D_l = -1]\Pr(D_w - D_l = -1), & \text{if } w < l \end{cases} \tag{A7}$$

and

$$\Pr(z_w) - \Pr(z_l) = \begin{cases} \Pr(D_w - D_l = 1), & \text{if } w > l \\ -\Pr(D_w - D_l = -1), & \text{if } w < l \end{cases}. \tag{A8}$$

Because $\mathbb{E}[T_1 - T_0 | D_w - D_l = 1] > 0$ if $w > l$ and $\mathbb{E}[T_1 - T_0 | D_w - D_l = -1] > 0$ if $w < l$. Then, if $\mathbb{E}(T|Z = z_w) - \mathbb{E}(T|Z = z_l) > 0$, from (A7) we know that $w > l$ and $\Pr(z_w) - \Pr(z_l) > 0$. Similar result holds if $\mathbb{E}(T|Z = z_w) - \mathbb{E}(T|Z = z_l) < 0$. At last, when $\mathbb{E}(T|Z = z_w) - \mathbb{E}(T|Z = z_l) = 0$, we can get $\Pr(D_w - D_l = 1) = 0$ and $\Pr(D_w - D_l = -1) = 0$. Therefore, we can conclude that the sign of $\mathbb{E}(T|Z = z_w) - \mathbb{E}(T|Z = z_l)$ is the same with the sign of $\Pr(z_w) - \Pr(z_l)$. □

A.4 Proof of Lemma 2.2

Proof of Lemma 2.2. By law of iterated expectation and the independence of instrument Z,

$$\begin{aligned} f_{(Y,T)|Z=z_k} =& f_{(Y,T)|C_k, Z=z_k} \Pr(C_k) + f_{(Y,T)|D_{k-1}=0, D_k=0, Z=z_k} \Pr(D_{k-1} = 0, D_k = 0) \\ &+ f_{(Y,T)|D_{k-1}=1, D_k=1, Z=z_k} \Pr(D_{k-1} = 1, D_k = 1) \\ =& f_{(Y_1,T_1)|C_k} \Pr(C_k) + f_{(Y_0,T_0)|D_{k-1}=0, D_k=0} \Pr(D_{k-1} = 0, D_k = 0) \\ &+ f_{(Y_1,T_1)|D_{k-1}=1, D_k=1} \Pr(D_{k-1} = 1, D_k = 1). \end{aligned}$$

Similarly,

$$\begin{aligned} f_{(Y,T)|Z=z_{k-1}} =& f_{(Y_0,T_0)|C_k} \Pr(C_k) + f_{(Y_0,T_0)|D_{k-1}=0, D_k=0} \Pr(D_{k-1} = 0, D_k = 0) \\ &+ f_{(Y_1,T_1)|D_{k-1}=1, D_k=1} \Pr(D_{k-1} = 1, D_k = 1). \end{aligned}$$

Therefore, we can get that

$$\begin{aligned} TV_{(Y,T),k} =& \frac{1}{2} \sum_{t=0,1} \int |f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)| d\mu_Y(y) \\ =& \frac{1}{2} \sum_{t=0,1} \int \left| f_{(Y_1,T_1)|C_k}(y,t) - f_{(Y_0,T_0)|C_k}(y,t) \right| d\mu_Y(y) Pr(C_k) \\ \leq& \frac{1}{2} \sum_{t=0,1} \int \left[f_{(Y_1,T_1)|C_k}(y,t) + f_{(Y_0,T_0)|C_k}(y,t) \right] d\mu_Y(y) Pr(C_k) \\ =& \Pr(C_k). \end{aligned}$$

By the monotonicity assumption, we know that compliers groups are mutually exclusive. Then,

$$\begin{aligned} \Pr(C_k) =& 1 - \sum_{k' \neq k} \Pr(C_{k'}) - \Pr(D_0 = D_1 = ... = D_K = 0) - \Pr(D_0 = D_1 = ... = D_K = 1) \\ \leq& 1 - \sum_{k' \neq k} \Pr(C_{k'}) \leq 1 - \sum_{k' \neq k} TV_{(Y,T),k'}, \end{aligned}$$

where the last inequality is due that $TV_{(Y,T),k} \leq \Pr(C_k)$ for all $k = 1, 2, ..., K$. □

A.5 Proof of Lemma 2.3

The proofs of Lemma 2.3 are similar to the proof of Theorem 17 in Ura (2018), but with nontrivial adjustments to deal with the multi-valued instrument setting of our paper. In order to prove this

Lemma, we need to introduce Lemma A.1 below. In what follows, we first prove Lemma A.1 and then proceed to the proof of Lemma 2.3.

Lemma A.1. *Under Assumptions 2.1-(ii)-(iv), 2.2-(i) and 2.3, we have that for* $\forall k = 1, 2, ..., K$,

(i) $TV_{(Y,T),k} \geq |\Delta_k \mathbb{E}(T|Z)|$;

(ii) $|\Delta_k \mathbb{E}(Y|Z)| > 0 \;\Rightarrow\; TV_{(Y,T),k} > 0$.

Proof of Lemma A.1. (i) This is a multi-valued IV version of the proof of Lemma 5 in Ura (2018).

$$\begin{aligned}
TV_{(Y,T),k} =& \frac{1}{2}\sum_{t=0,1}\int |f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)| d\mu_Y(y) \\
\geq& \frac{1}{2}\sum_{t=0,1}\left|\int f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t) d\mu_Y(y)\right| \\
=& \frac{1}{2}\sum_{t=0,1}\left|f_{T|Z=z_k}(t) - f_{T|Z=z_{k-1}}(t)\right| \\
=& \frac{1}{2}\left[\left|f_{T|Z=z_k}(1) - f_{T|Z=z_{k-1}}(1)\right| + \left|f_{T|Z=z_k}(0) - f_{T|Z=z_{k-1}}(0)\right|\right] \\
=& \left|f_{T|Z=z_k}(1) - f_{T|Z=z_{k-1}}(1)\right| \\
=& |\Delta_k \mathbb{E}(T|Z)|.
\end{aligned}$$

(ii) We prove (ii) by verifying $\Delta_k \mathbb{E}(Y|Z) \neq 0$ implies that

$$\Pr\left[\left\{(y,t) \in \Omega_Y \times \{0,1\} : \left|f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)\right| \neq 0\right\}\right] > 0. \tag{A9}$$

It can be verified by proof by contradiction as below. Suppose $\Delta_k \mathbb{E}(Y|Z) \neq 0$ but the probability in (A9) is zero. It means with probability one that

$$\begin{aligned}
& \left|f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)\right| = 0 \\
\Leftrightarrow\;& f_{(Y,T)|Z=z_k}(y,t) = f_{(Y,T)|Z=z_{k-1}}(y,t), \text{ for both } t = 0,1 \\
\Leftrightarrow\;& \sum_{t=0,1} f_{(Y,T)|Z=z_k}(y,t) = \sum_{t=0,1} f_{(Y,T)|Z=z_{k-1}}(y,t) \\
\Leftrightarrow\;& f_{Y|Z=z_k}(y) = f_{Y|Z=z_{k-1}}(y) \\
\Leftrightarrow\;& \Delta_k \mathbb{E}(Y|Z) = \int y[f_{Y|Z=z_k}(y) - f_{Y|Z=z_{k-1}}(y)] d\mu_Y(y) = 0,
\end{aligned} \tag{A10}$$

which contradicts $\Delta_k \mathbb{E}(Y|Z) \neq 0$. Therefore, $|\Delta_k \mathbb{E}(Y|Z)| > 0$ implies (A9), and we have that $TV_{(Y,T),k} > 0$ by definition. □

Now we can proceed to the proof of Lemma 2.3.

Proof of Lemma 2.3. (i) If $TV_{(Y,T),k} = 0$, then by Lemma A.1 $\Delta_k \mathbb{E}(Y|Z) = 0$, and any $\alpha_{k,k-1} \in \Theta$ satisfies the inequalities (6), (7) and (8). If $TV_{(Y,T),k} > 0$, we have $1 - \sum_{k' \neq k} TV_{(Y,T),k'} > 0$ and

$$\frac{|\Delta_k \mathbb{E}(Y|Z)|}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \leq |\alpha_{k,k-1}| \leq \frac{|\Delta_k \mathbb{E}(Y|Z)|}{TV_{(Y,T),k}},$$

and the sign of $\alpha_{k,k-1}$ is identified by the sign of $\Delta_k \mathbb{E}(Y|Z)$.

(ii) The sharpness can be proved by the same proof of Lemma 2.4(ii), via replacing $\Delta_k \mathbb{E}(T|Z)$ and Δp_k in the proof of Lemma 2.4(ii) by $\Delta_k \mathbb{E}(Y|Z)$ and $\alpha_{k,k-1}$ respectively. □

A.6 Proof of Lemma 2.4

Proof of Lemma 2.4. (i) If $TV_{(Y,T),k} = 0$, $\Delta_k \mathbb{E}(T|Z) = 0$ by Lemma A.1, and any $\Delta p_k \in [-1,1]$ satisfies the inequalities (10), (11) and (12). If $TV_{(Y,T),k} > 0$, we have $1 - \sum_{k' \neq k} TV_{(Y,T),k'} > 0$ and

$$\frac{|\Delta_k \mathbb{E}(T|Z)|}{1 - \sum_{k' \neq k} TV_{(Y,T),k'}} \leq |\Delta p_k| \leq \frac{|\Delta_k \mathbb{E}(T|Z)|}{TV_{(Y,T),k}},$$

and the sign of Δp_k is identified by the sign of $\Delta_k \mathbb{E}(T|Z)$.

(ii) The proof of sharpness can be implemented in two steps.

In **Step 1**, we show that if $\max_{0 \leq m \leq K} VT_{(Y,T),m} = 0$, which means all $VT_{(Y,T),m} = 0$ for $\forall m = 1, ..., K$, the sharp identified set for Δp_k is $[-1, 1]$. In **Step 2**, we show that if $TV_{(Y,T),k} > 0$ and $TV_{(Y,T),k'} = 0$ with all $k' \neq k$, then any point lies in $\Theta_k^p(\mathbf{P})$ equals to Δp_k under some DGP which generates (Y, T, Z).

Step 1. Since $VT_{(Y,T),m} = 0$ for all m, we know

$$f_{(Y,T)|Z=z_0}(y,t) = f_{(Y,T)|Z=z_1}(y,t) = ... = f_{(Y,T)|Z=z_K}(y,t) = f_{(Y,T)}(y,t) \tag{A11}$$

almost sure for all $(y,t) \in \Omega_Y \times \{0,1\}$.

Denote f_1, f_0 to be any arbitrary pair of well-defined probability functions with support $[0,1]$, satisfying $0 \leq f_1, f_0 \leq 1$ and $\sum_{t=0,1} f_1 = \sum_{t=0,1} f_0 = 1$. Define a data generate process $P^*_{f_1,f_0}$ based on f_1, f_0 as below:

$$\begin{aligned} Z \sim f_Z, \quad D_k|_Z = 1 \text{ for all } k = 0, 1, ..., K, \\ (Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)} \sim \begin{cases} f_{(Y,T)}, & \text{if all } D_k \text{ are equal,} \\ f_Y f_1, & \text{if at least one } D_k \neq D_{k-1}. \end{cases} \\ (Y_0, T_0)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_Y f_0, \end{aligned}$$

where f_Z, f_Y and $f_{(Y,T)}$ are the true marginal distributions of the observable (Y, T, Z). In what follows, we denote f^* as any density function associated with the DGP P^*. Next, we show that for any arbitrary pair f_1, f_0 described above:

(a) $P^*_{f_1,f_0}$ satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i).

(b) $P^*_{f_1,f_0}$ generates the data (Y, T, Z).

(c) Under $P^*_{f_1,f_0}$, we have that $\Delta p_k = f_1(1) - f_0(1)$ for all $k = 1, 2, ..., K$.

(a) The DGP $P^*_{f_{T_1},f_{T_0}}$ above shows that $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, T_1, T_0)$, and $D_l \geq D_w$ almost surely and $P(z_l) \geq P(z_w)$ for $l > w$.

(b) Denote f^* as the distribution function under $P^*_{f_1,f_0}$, e.g. f^*_Y is the distribution of Y generated by the DGP $P^*_{f_1,f_0}$. Then, for $\forall k = 0, 1, ..., K$

$$\begin{aligned} f^*_{(Y,T)|Z=z_k}(y,t) &= f^*_{(Y,T)|D_0=1,D_1=1,...,D_K=1,Z=z_k}(y,t) \\ &= f^*_{(Y_1,T_1)|D_0=1,D_1=1,...,D_K=1,Z=z_k}(y,t) \\ &= f_{(Y,T)}(y,t) \\ &= f_{(Y,T)|Z=z_k}(y,t) \end{aligned}$$

where the third equality is due that $(Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_{(Y,T)}$ if all D_k are equal, and the last equality is because of (A11). Thus, $P^*_{f_1,f_0}$ generates (Y, T, Z), since $f^*_{(Y,T)|Z=z_k} = f_{(Y,T)|Z=z_k}$.

(c) Under $P^*_{f_1,f_0}$, we have the independence of Z to $(T_1, T_0, \{D_k\}_{k=0}^K)$,

$$\begin{aligned}\Delta p_k = \mathbb{E}_{P^*_{f_1,f_0}}[T_1 - T_0|C_k] &= \mathbb{E}_{P^*_{f_1,f_0}}[T_1 - T_0|C_k, Z] \\ &= f^*_{T_1|C_k,Z}(1) - f^*_{T_0|C_k,Z}(1) \\ &= \int f^*_{Y_1,T_1|C_k,Z}(y,1)d\mu_Y(y) - \int f^*_{Y_0,T_0|C_k,Z}(y,1)d\mu_Y(y) \\ &= f_1(1)\int f_Y(y)d\mu_Y(y) - f_0(1)\int f_Y(y)d\mu_Y(y) \\ &= f_1(1) - f_0(1).\end{aligned}$$

Given that $P^*_{f_1,f_0}$ with any pair of (f_1, f_0) satisfies (a)-(c), it fulfills the proof of **Step 1**.

Step 2. We prove the statement in **Step 2** above in three sub-steps.

(a) There exists a DGP P^*_L that satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i), generates (Y, T, Z) and $\Delta p_k = \Delta_k \mathbb{E}(T|Z)$ under P^*_L.

(b) There exists a DGP P^*_U that satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i), generates (Y, T, Z) and $\Delta p_k = \frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}}$ under P^*_U.

(c) For some constant $\psi \in [0, 1]$, the mixture $\psi P^*_L + (1-\psi)P^*_U$ satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i), generates (Y, T, Z) and $\Delta p_k = \psi \Delta_k \mathbb{E}(T|Z) + (1-\psi)\frac{\Delta_k \mathbb{E}(T|Z)}{TV_{(Y,T),k}}$.

(a) Given $TV_{(Y,T),k} > 0$ and $TV_{(Y,T),k'} = 0$ with all $k' \neq k$, define a DGP P^*_L as below:

$$\begin{aligned}Z \sim f_Z, \;\; (D_{k-1}, D_k)|_Z = (0,1), \;\; D_l \leq D_w \text{ if } l < w \\ (Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_{(Y,T)|Z=z_k} \\ (Y_0, T_0)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_{(Y,T)|Z=z_{k-1}}.\end{aligned}$$

It is easy to see that under P^*_L, $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, T_1, T_0)'$, $D_l \geq D_w$ almost surely and $Pr(z_l) \geq Pr(z_w)$ for $l > w$. Denote f^{*L} as the distribution functions under P^*_L. Then, for $\forall m \leq k-1$,

$$f^{*L}_{(Y,T)|Z=z_m}(y,t) = f^{*L}_{(Y,T)|D=0,Z=z_m}(y,t) = f^{*L}_{(Y_0,T_0)|D_m=0,Z=z_m}(y,t) = f_{(Y,T)|Z=z_{k-1}} = f_{(Y,T)|Z=z_m},$$

where the last equality is due to $TV_{(Y,T),k'} = 0$ for all $k' \neq k$, implying $f_{(Y,T)|Z=z_m} = f_{(Y,T)|Z=z_{k-1}}$ for all $m \leq k-1$. Furthermore, for $\forall m \geq k$,

$$f^{*L}_{(Y,T)|Z=z_m}(y,t) = f^{*L}_{(Y,T)|D=1,Z=z_m}(y,t) = f^{*L}_{(Y_1,T_1)|D_m=1,Z=z_m}(y,t) = f_{(Y,T)|Z=z_k} = f_{(Y,T)|Z=z_m},$$

where the last equality is due to $TV_{(Y,T),k'} = 0$ for all $k' \neq k$, implying $f_{(Y,T)|Z=z_m} = f_{(Y,T)|Z=z_k}$ for all $m \geq k$. Hence, we have shown that the DGP P^*_L generates (Y, T, Z).

Next, consider Δp_k under P^*_L:

$$\begin{aligned}\Delta p_k = \mathbb{E}_{P^*_L}[T_1 - T_0|C_k] &= \mathbb{E}_{P^*_L}[T_1|C_k, Z = z_k] - \mathbb{E}_{P^*_L}[T_0|C_k, Z = z_{k-1}] \\ &= f_{T|Z=z_k}(1) - f_{T|Z=z_{k-1}}(1) \\ &= \mathbb{E}[T|Z = z_k] - \mathbb{E}[T|Z = z_{k-1}] \\ &= \Delta_k \mathbb{E}[T|Z].\end{aligned}$$

(b) Given $TV_{(Y,T),k} > 0$ and $TV_{(Y,T),k'} = 0$ with all $k' \neq k$, we first define a random variable $H = 0.5 \times \text{sign}(\Delta_k f_{(Y,T)|Z}(Y,T))$, where $\Delta_k f_{(Y,T)|Z}(Y,T) = f_{(Y,T)|Z=z_k}(Y,T) - f_{(Y,T)|Z=z_{k-1}}(Y,T)$. Then,

let us define a DGP P_U^* as follows

$$
\begin{aligned}
Z &\sim f_Z, \\
(D_{k-1}, D_k)|_Z &= \begin{cases} (0,1), & D_l \le D_w \text{ if } l < w, \quad \text{with probability } \Delta_k \mathbb{E}[H|Z], \\ (0,0), & D_l = D_w \text{ for all } l, w, \quad \text{with probability } Pr(H = -0.5|Z = z_k), \\ (1,1), & D_l = D_w \text{ for all } l, w, \quad \text{with probability } Pr(H = 0.5|Z = z_{k-1}). \end{cases} \\
(Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)} &\sim \begin{cases} \frac{\Delta_k f_{(Y,T,H)|Z}(y,t,0.5)}{\Delta_k \mathbb{E}[H|Z]}, & \text{if } D_{k-1} < D_k, \\ f_{(Y,T)|H=0.5, Z=z_{k-1}}(y,t), & \text{if } D_{k-1} = D_k \end{cases} \\
(Y_0, T_0)|_{(\{D_k\}_{k=0}^K, Z)} &\sim \begin{cases} -\frac{\Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)}{\Delta_k \mathbb{E}[H|Z]}, & \text{if } D_{k-1} < D_k, \\ f_{(Y,T)|H=-0.5, Z=z_k}(y,t), & \text{if } D_{k-1} = D_k \end{cases}
\end{aligned}
$$

First of all, noticing that

$$
\begin{aligned}
\Delta_k \mathbb{E}[H|Z] &= \mathbb{E}[H|Z = z_k] - \mathbb{E}[H|Z = z_{k-1}] \\
&= \frac{1}{2} \sum_{t=0,1} \Big[\int \text{sign}(\Delta_k f_{(Y,T)|Z}(y,t)) f_{(Y,T)|Z=z_k}(y,t) d\mu_Y(y) \\
&\qquad - \int \text{sign}(\Delta_k f_{(Y,T)|Z}(y,t)) f_{(Y,T)|Z=z_{k-1}}(y,t) d\mu_Y(y) \Big] \\
&= \frac{1}{2} \sum_{t=0,1} \left[\int \text{sign}(\Delta_k f_{(Y,T)|Z}(y,t)) \Delta_k f_{(Y,T)|Z}(y,t) d\mu_Y(y) \right] \\
&= \frac{1}{2} \sum_{t=0,1} \int \left| \Delta_k f_{(Y,T)|Z}(y,t) \right| d\mu_Y(y) \\
&= TV_{(Y,T),k}.
\end{aligned}
$$

It's easy to check that DGP P_U^* satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i). Denote f^{*U} as the distribution functions under P_U^*. We first show that f^{*U} is well-defined: (b.1) the summation of the probabilities of all possible values for $\{D_k\}_{k=0}^K$ is one, (b.2) the density functions of $(Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)}$ and $(Y_0, T_0)|_{(\{D_k\}_{k=0}^K, Z)}$ under P_U^* are nonnegative, and (b.3) their integrals are one.

(b.1) Consider the following summation.

$$
\begin{aligned}
&\Delta_k \mathbb{E}[H|Z] + \Pr(H = -0.5|Z = z_k) + \Pr(H = 0.5|Z = z_{k-1}) \\
=&0.5\Pr(H = 0.5|Z = z_k) - 0.5\Pr(H = -0.5|Z = z_k) - 0.5\Pr(H = 0.5|Z = z_{k-1}) \\
&\quad + 0.5\Pr(H = -0.5|Z = z_{k-1}) + \Pr(H = -0.5|Z = z_k) + \Pr(H = 0.5|Z = z_{k-1}) \\
=&0.5\Pr(H = 0.5|Z = z_k) + 0.5\Pr(H = -0.5|Z = z_k) + 0.5\Pr(H = 0.5|Z = z_{k-1}) \\
&\quad + 0.5\Pr(H = -0.5|Z = z_{k-1}) \\
=&0.5 + 0.5 = 1.
\end{aligned}
$$

(b.2) We show that the density functions of $(Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)}$ and $(Y_1, T_1)|_{(\{D_k\}_{k=0}^K, Z)}$ under P_U^* are nonnegative and integral are one.

$$
\begin{aligned}
\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) &= f_{(Y,T,H)|Z=z_k}(y,t,0.5) - f_{(Y,T,H)|Z=z_{k-1}}(y,t,0.5) \\
&= f_{(Y,T)|Z=z_k}(y,t) 1[\Delta_k f_{(Y,T)|Z}(y,t) \ge 0] - f_{(Y,T)|Z=z_{k-1}}(y,t) 1[\Delta_k f_{(Y,T)|Z}(y,t) \ge 0] \\
&= \Delta_k f_{(Y,T)|Z}(y,t) 1[\Delta_k f_{(Y,T)|Z}(y,t) \ge 0] \ge 0. \qquad \text{(A12)}
\end{aligned}
$$

Moreover,

$$\begin{aligned}\Delta_k f_{(Y,T,H)|Z}(y,t,-0.5) &= f_{(Y,T,H)|Z=z_k}(y,t,-0.5) - f_{(Y,T,H)|Z=z_{k-1}}(y,t,-0.5)\\ &= f_{(Y,T)|Z=z_k}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t)<0] - f_{(Y,T)|Z=z_{k-1}}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t)<0]\\ &= \Delta_k f_{(Y,T)|Z}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t)<0] \le 0. \end{aligned} \tag{A13}$$

Since $\Delta_k \mathbb{E}[H|Z] = TV_{(Y,T),k} > 0$, the density functions are both nonnegative.

(b.3) From (A12) and (A13) we have that

$$\begin{aligned}&\sum_{t=0,1}\int \left[\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) + \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)\right] d\mu_Y(y)\\ =&\sum_{t=0,1}\int \Delta_k f_{(Y,T)|Z}(y,t) d\mu_Y(y) = 0, \end{aligned} \tag{A14}$$

and

$$\begin{aligned}&\sum_{t=0,1}\int \left[\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) - \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)\right] d\mu_Y(y)\\ =&\sum_{t=0,1}\int \Delta_k f_{(Y,T)|Z}(y,t)\text{sign}(\Delta_k f_{(Y,T)|Z}(y,t)) d\mu_Y(y)\\ =&\sum_{t=0,1}\int |\Delta_k f_{(Y,T)|Z}(y,t)| d\mu_Y(y)\\ =&2TV_{(Y,T),k}. \end{aligned} \tag{A15}$$

Based on (A14) and (A15), we get that

$$\sum_{t=0,1}\int \Delta_k f_{(Y,T,H)|Z}(y,t,0.5) d\mu_Y(y) = TV_{(Y,T),k}, \tag{A16}$$

$$\sum_{t=0,1}\int \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5) d\mu_Y(y) = -TV_{(Y,T),k}. \tag{A17}$$

Given (A16) and (A17), it is clear that the integrals of the density functions are all one.

Next, we show that P_U^* generates the data (Y,T,Z). For $\forall m \le k-1$,

$$\begin{aligned} f^{*U}_{(Y,T)|Z=z_m}(y,t) =& f^{*U}_{(Y,T)|D_0=\ldots=D_K=0,Z=z_m}(y,t)\Pr(H=-0.5|Z=z_k)\\ &+ f^{*U}_{(Y,T)|D_0=\ldots=D_K=1,Z=z_m}(y,t)\Pr(H=0.5|Z=z_{k-1})\\ &+ f^{*U}_{(Y,T)|D_0=0,\ldots,=D_{k-1}=0,D_k=1,\ldots,D_K=1,Z=z_m}(y,t)\Delta_k\mathbb{E}[H|Z]\\ =& f_{(Y,T)|H=-0.5,Z=z_k}(y,t)\Pr(H=-0.5|Z=z_k)\\ &+ f_{(Y,T)|H=0.5,Z=z_{k-1}}(y,t)\Pr(H=0.5|Z=z_{k-1})\\ &- \Delta_k\mathbb{E}[H|Z]\frac{\Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)}{\Delta_k\mathbb{E}[H|Z]}\\ =& f_{(Y,T,H)|Z=z_{k-1}}(y,t,0.5) + f_{(Y,T,H)|Z=z_{k-1}}(y,t,-0.5)\\ =& f_{(Y,T)|Z=z_{k-1}}(y,t)\\ =& f_{(Y,T)|Z=z_m}(y,t), \end{aligned}$$

where the last equality is because $TV_{(Y,T),m} = 0$ for all $m \leq k-1$. Moreover, we have for $m \geq k$,

$$\begin{aligned}
f^{*U}_{(Y,T)|Z=z_m}(y,t) =& f^{*U}_{(Y,T)|D_0=\ldots=D_K=0,Z=z_m}(y,t)\Pr(H=-0.5|Z=z_k) \\
&+ f^{*U}_{(Y,T)|D_0=\ldots=D_K=1,Z=z_m}(y,t)\Pr(H=0.5|Z=z_{k-1}) \\
&+ f^{*U}_{(Y,T)|D_0=0,\ldots,=D_{k-1}=0,D_k=1,\ldots,D_K=1,Z=z_m}(y,t)\Delta_k\mathbb{E}[H|Z] \\
=& f_{(Y,T)|H=-0.5,Z=z_k}(y,t)\Pr(H=-0.5|Z=z_k) \\
&+ f_{(Y,T)|H=0.5,Z=z_{k-1}}(y,t)\Pr(H=0.5|Z=z_{k-1}) \\
&+ \Delta_k\mathbb{E}[H|Z]\frac{\Delta_k f_{(Y,T,H)|Z}(y,t,0.5)}{\Delta_k\mathbb{E}[H|Z]} \\
=& f_{(Y,T,H)|Z=z_k}(y,t,-0.5) + f_{(Y,T,H)|Z=z_k}(y,t,0.5) \\
=& f_{(Y,T)|Z=z_k}(y,t) \\
=& f_{(Y,T)|Z=z_m}(y,t),
\end{aligned}$$

where the last equality is because of $TV_{(Y,T),m} = 0$ for all $m \geq k$. Thus, so far we have shown that P^*_U generates the data (Y,T,Z).

The last step in (b) is to prove that under P^*_U, $\Delta p_k = \frac{\Delta_k\mathbb{E}(T|Z)}{TV_{(Y,T),k}}$:

$$\begin{aligned}
\Delta p_k = \mathbb{E}_{P^*_U}[T_1 - T_0|C_k] =& \mathbb{E}_{P^*_U}[T_1|C_k,Z] - \mathbb{E}_{P^*_U}[T_0|C_k,Z] \\
=& \int \frac{\Delta_k f_{(Y,T,H)|Z}(y,1,0.5)}{\Delta_k\mathbb{E}[H|Z]}d\mu_Y(y) + \int \frac{\Delta_k f_{(Y,T,H)|Z}(y,1,-0.5)}{\Delta_k\mathbb{E}[H|Z]}d\mu_Y(y) \\
=& \int \frac{\Delta_k f_{(Y,T)|Z}(y,1)}{\Delta_k\mathbb{E}[H|Z]}d\mu_Y(y) \\
=& \frac{\Delta_k\mathbb{E}[T|Z]}{TV_{(Y,T),k}}.
\end{aligned}$$

(c) For any $\psi \in [0,1]$, denote the mixture DGP $P^*_{mix} := \psi P^*_L + (1-\psi)P^*_U$, which means with probability ψ the data (Y,T,Z) is generated from P^*_L and with probability $1-\psi$ the data (Y,T,Z) is generated from P^*_U. Given the results in **Steps 1 and 2**, we have that if $TV_{(Y,T),k} > 0$ and $TV_{(Y,T),k'} = 0$ with all $k' \neq k$, the DGP P^*_{mix} satisfies Assumptions 2.1-(ii)-(iv) and 2.2-(i); P^*_{mix} generates the data (Y,T,Z); and under P^*_{mix}, $\Delta p_k = \psi\Delta_k\mathbb{E}(T|Z) + (1-\psi)\frac{\Delta_k\mathbb{E}(T|Z)}{TV_{(Y,T),k}}$. □

A.7 Proof of Theorem 2.2

Proof of Theorem 2.2. From (14), we have $\min_{k\in\{1,2,\ldots,K\}}\{\alpha_{k,k-1}\} \leq \alpha^{IV} \leq \max_{k\in\{1,2,\ldots,K\}}\{\alpha_{k,k-1}\}$. Because each LATE is partially identified by $\Theta^\alpha_k(\mathbf{P})$, based on which we have $\alpha^{IV} \in \bigcup_{k=1,2,\ldots,K}\Theta^\alpha_k(\mathbf{P})$. □

A.8 Proof of Theorem 2.3

Proof of Theorem 2.3. Since $\xi = \sum_{k=1}^K \gamma^{IV}_k \Delta p_k$, we know that $\min_{k=1,2,\ldots,K}\{\Delta p_k\} \leq \xi \leq \max_{k=1,2,\ldots,K}\{\Delta p_k\}$. Thus, $\xi \in \bigcup_{k=1,2,\ldots,K}\Theta^p_k(\mathbf{P})$ leads to the result. □

A.9 Proof of Corollary 2.4

Proof of Corollary 2.4. By definition of $\Delta_k\mathbb{E}(Y|Z)$ and $\Delta_k\mathbb{E}(T|Z)$, we know $\Delta_k\mathbb{E}(Y|Z)/\Delta_k\mathbb{E}(T|Z) = \alpha_{k,k-1}/\Delta p_k$. From Theorem 2.2, we have that

$$\begin{aligned}\Theta^{\alpha}(\mathbf{P}) &= \bigcup_{k=1,2,\ldots,K} \Theta_k^{\alpha}(\mathbf{P}) = \bigcup_{k=1,2,\ldots,K} \left\{ \frac{\Delta_k\mathbb{E}(Y|Z)}{pc} : \ pc \in \left[TV_{(Y,T),k}, 1 - \sum_{k' \neq k} TV_{(Y,T),k'} \right] \right\} \\ &= \bigcup_{k=1,2,\ldots,K} \left\{ \frac{\alpha_{k,k-1}}{\Delta p_k} \times \frac{\Delta_k\mathbb{E}(T|Z)}{pc} : \ pc \in \left[TV_{(Y,T),k}, 1 - \sum_{k' \neq k} TV_{(Y,T),k'} \right] \right\} \\ &= \bigcup_{k=1,2,\ldots,K} \left\{ \frac{\alpha_{k,k-1}}{\Delta p_k} \times \Delta p : \ \Delta p \in \Theta_k^{p}(\mathbf{P}) \right\},\end{aligned}$$

where the last equality is due to the definition of $\Theta_k^p(\mathbf{P})$. Similarly, from Theorem 2.3 and (3)

$$\begin{aligned}\Theta^{p}(\mathbf{P}) &= \left\{ \alpha^{Mis} \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,\ldots,K} \Theta_k^{p}(\mathbf{P}) \right\} = \left\{ \frac{\alpha^{IV}}{\xi} \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,\ldots,K} \Theta_k^{p}(\mathbf{P}) \right\} \\ &= \bigcup_{k=1,2,\ldots,K} \left\{ \frac{\alpha^{IV}}{\xi} \times \Delta p : \ \Delta p \in \Theta_k^{p}(\mathbf{P}) \right\}.\end{aligned}$$

□

A.10 Proof of Theorem 2.4

Proof of Theorem 2.4. Since $0 < \underline{\xi} \leq \xi \leq \bar{\xi} \leq 1$, it yields from $\alpha^{IV} = \xi\alpha^{Mis}$ that α^{IV} is between $\underline{\xi}\alpha^{Mis}$ and $\overline{\xi}\alpha^{Mis}$, and its sign is determined by the sign of α^{Mis}. □

A Online Appendix

This Online Appendix contains three sections with proofs, additional material and analysis. The information are organized as follows. Appendix A.1 provides the details to construct the confidence intervals of the partially identified α^{IV}. Appendix A.2 presents the details about the partial identification results using multiple treatment proxies. Appendix A.3 discusses the details about how to use our partial identification strategies in empirical applications with covariates.

Contents

A.1 Inference: Details

In this section, we provide the details to construct the confidence interval of α^{IV}. Since, the partial identification of α^{IV} is based on the union of either $\Theta_k^{\alpha}(\mathbf{P})$ or $\Theta_k^{p}(\mathbf{P})$, we proceed with the estimation in three steps. First, we construct the moment inequalities representations of the sets $\Theta_k^{\alpha}(\mathbf{P})$ and $\Theta_k^{p}(\mathbf{P})$. Second, we construct the confidence intervals for $\alpha_{k,k-1}$ and Δp_k. Third, depending on the chosen identification strategy, we construct the appropriate confidence intervals of α^{IV} by taking the unions of the confidence intervals of either $\alpha_{k,k-1}$ or Δp_k.

A.1.1 Moment Inequalities Representations

The Lemma below shows that $\Theta_k^{\alpha}(\mathbf{P})$ and $\Theta_k^{p}(\mathbf{P})$ have equivalent expressions in terms of unconditional moment inequalities.

Lemma A.1. *Let Assumption 2.1, 2.2 and 2.3 hold. Denote a random variable*

$$\varphi_k = \frac{1[Z = z_k]\pi_{k-1} - 1[Z = z_{k-1}]\pi_k}{\pi_k \pi_{k-1}}$$

for $k = 1, 2, ..., K$. Then, $\Theta_k^{\alpha}(\mathbf{P})$ can be characterized by the following moment inequalities:

$$\mathbb{E}\left[-\varphi_k sign(\alpha_{k,k-1})Y\right] \leq 0, \tag{A1}$$

$$\mathbb{E}\left\{\varphi_k\left[|\alpha_{k,k-1}|h(Y,T) - sign(\alpha_{k,k-1})Y\right]\right\} \leq 0, \ \forall h \in \mathbf{H} \tag{A2}$$

$$\mathbb{E}\left[\varphi_k sign(\alpha_{k,k-1})Y - |\alpha_{k,k-1}|\left(1 - \sum_{k' \neq k} \varphi_{k'} h_{k'}(Y,T)\right)\right] \leq 0, \ \forall h_{k'} \in \mathbf{H}. \tag{A3}$$

Moreover, $\Theta_k^{p}(\mathbf{P})$ can be characterized by the following moment inequalities:

$$\mathbb{E}\left[-\varphi_k sign(\Delta p_k)T\right] \leq 0, \tag{A4}$$

$$\mathbb{E}\left\{\varphi_k\left[|\Delta p_k|h(Y,T) - sign(\Delta p_k)T\right]\right\} \leq 0, \ \forall h \in \mathbf{H} \tag{A5}$$

$$\mathbb{E}\left[\varphi_k sign(\Delta p_k)T - |\Delta p_k|\left(1 - \sum_{k' \neq k} \varphi_{k'} h_{k'}(Y,T)\right)\right] \leq 0, \ \forall h_{k'} \in \mathbf{H}, \tag{A6}$$

where $\pi_k = Pr(Z = z_k)$, $\mathbf{H}$ is a set of measurable functions mapping $(y,t) \in \Omega_Y \times \{0,1\}$ to $\{-0.5, 0.5\}$ and $sign(x) = 1[x \geq 0] - 1[x < 0]$.

Proof of Lemma A.1. See Appendix A.3.1.[38] □

Lemma A.1 is based on the facts that $\Delta_k \mathbb{E}[h(Y,T)|Z]$ with $h \in \mathbf{H}$ can bound the total variation distance:

$$\Delta_k \mathbb{E}[h(Y,T)|Z] \leq TV_{(Y,T),k},$$

and φ_k helps rewrite the conditional moments to unconditional ones: for $Q \in \{Y, T, h(Y,T)\}$,

$$\Delta_k \mathbb{E}[Q|Z] = \mathbb{E}[\varphi_k Q].$$

Next, we introduce some regularity conditions on the data generating process. Denote $\pi = (\pi_0, \pi_1, ..., \pi_K)'$ and its parameter space as $\mathbf{\Pi} \subset [0,1]^{(K+1)}$. Suppose $(1 - \eta_\pi)$-confidence interval

[38] We use subscript k' to distinguish different $h_{k'}$, because each $\varphi_{k'}$ can be multiplied by different $h_{k'}$ and it is not necessarily the same with h. For simplicity, we do not distinguish h and $h_{k'}$ elsewhere if it is not necessary, and we use h to denote any generic function in $\mathbf{H}$.

for all π_k, denoted as $\mathscr{C}_{\pi_k}(\eta_\pi)$, and $(1-\eta_{\alpha^{Mis}})$-confidence interval for α^{Mis}, denoted as $\mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})$, are available.

Assumption A.1. *The parameter space $\Theta \times \mathbf{\Pi} \times \mathscr{P}_0$ satisfies the following conditions:*

(i) Θ is bounded. $\max\{\mathbb{E}[Y^3]^{2/3}, \mathbb{E}[Y^4]^{1/2}\} < M$ *for some constant M.*

(ii) All random variables inside $\mathbb{E}[\cdot]$ in Lemma A.1 have nonzero variance for $\forall h \in \mathbf{H}$, $\forall \alpha_{k,k-1} \in \Theta$ and $\forall \Delta p_k \in [-1,1]$.

(iii) $\liminf_{n\to\infty} \inf_{\mathbf{P}\in\mathscr{P}_0} Pr[\pi_k \in \mathscr{C}_{\pi_k}(\eta_\pi)] \geq 1-\eta_\pi$ *for $k = 1,2,...,K$.*

(iv) $\liminf_{n\to\infty} \inf_{\mathbf{P}\in\mathscr{P}_0} Pr[\alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})] \geq 1-\eta_{\alpha^{Mis}}$.

The number of the moment inequalities in Lemma A.1 can be either finite or infinite, depending on the support of Y. If Y is discrete, the number of possible $h \in \mathbf{H}$ is finite, so as the total number of the moment inequalities in Lemma A.1. When Y is continuous, the number of elements in $\mathbf{H}$ will be infinite and we are then facing an infinite number of moment inequalities. To deal with the potential uncountable infinite moment inequalities, we consider a sequence of sets $\mathbf{H}_n$, which converges to $\mathbf{H}$ in the sense defined in Assumption A.2 (When $\mathbf{H}$ has finite dimension, we can simply let $\mathbf{H}_n = \mathbf{H}$). The key in forming $\mathbf{H}_n$ is the partition $\Omega_Y \times \{0,1\} = \bigcup_{l=1,2,...,L_n} I_{n,l}$, in which L_n is the number of the partitions $\{I_{n,l}\}$, and L_n may grow with the sample size n. Denote by $h_{n,j}$, $j = 1,2,...,2^{L_n}$, the function that maps $\Omega_Y \times \{0,1\}$ into $\{-0.5, 0.5\}$, which is a constant over each $I_{n,l}$, $l = 1,2,...,L_n$. We can then define $\mathbf{H}_n = \{h_{n,1}, h_{n,2}, ..., h_{n,2^{L_n}}\}$ to be the collection of all such functions.

By construction, $\mathbf{H}_n$ is a subset of $\mathbf{H}$. Replacing $\mathbf{H}$ by $\mathbf{H}_n$ in the moment inequalities in Lemma A.1 yields two sets, denoted by $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$. They cover and converge to $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^p(\mathbf{P})$, respectively, as the sample size increases. Their convergence will be formally described in Lemma A.2 below. Thus, the confidence intervals considered later will be based on the moment inequalities that characterize $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$.

Let $\kappa_n = 2^{L_n}$ be the number of functions in $\mathbf{H}_n$, and denote by p_n the number of moment inequalities under $\mathbf{H}_n$. Then, $p_n = 1 + \kappa_n + \kappa_n^{K-1}$. Assumption A.2 below outlines the sufficient assumptions on the DGP and the partition $\{I_{n,l}\}_{l=1}^{L_n}$ that ensure the convergence of $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$ to $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^p(\mathbf{P})$.

Assumption A.2. *The following assumptions hold:*

(i) The density function $f_{(Y,T)|Z=z_k}(y,t)$ is Hölder continuous in $(y,t) \in \Omega_Y \times \{0,1\}$ with the Hölder constant M_0 and exponent m.

(ii) The partition $I_{n+1,1}, I_{n+1,2}, ..., I_{n+1,L_{n+1}}$ is a refinement of the partition $I_{n,1}, I_{n,2}, ..., I_{n,L_n}$.

(iii) There is a positive constant M_1 such that $I_{n,l}$ is a subset of some open ball with radius M_1/L_n in $\Omega_Y \times \{0,1\}$.

(iv) There exist some constants $c_1 \in (0, 1/2)$ and $C_1 > 0$ such that p_n satisfies

$$\log^{7/2}(p_n n) \leq C_1 n^{1/2-c_1}, \ \log^{1/2} p_n \leq C_1 n^{1/2-c_1}, \ \log^{3/2} p_n \leq C_1 n.$$

Assumption A.2-(i) restricts the smoothness of the density function of observable (Y,T) and A.2-(ii) implies the sequence $\{\mathbf{H}_n\}$ satisfying $\mathbf{H}_n \subset \mathbf{H}_{n+1} \subset \cdots \subset \mathbf{H}$. Assumption A.2-(iii) is used to make sure that the partition becomes finer as sample size increases. Assumption A.2-(iv) is borrowed from Chernozhukov et al. (2019) for the asymptotic performance of the confidence interval.

If $\mathbf{H}_n = \{h_{n,1}, h_{n,2}, ..., h_{n,\kappa_n}\}$ based on partition $\{I_{n,l}\}_{l=1}^{L_n}$ satisfies Assumption A.2, we have the following convergence results.

Lemma A.2. *Let Assumption 2.1, 2.2, 2.3, A.1 and A.2 hold. Then,* $\Theta_k^{\alpha}(\mathbf{P}) \subset \widetilde{\Theta}_k^{\alpha}(\mathbf{P})$ *and* $\Theta_k^{p}(\mathbf{P}) \subset \widetilde{\Theta}_k^{p}(\mathbf{P})$. *As sample size increases, the convergence below hold uniformly over* $(\pi, \mathbf{P}) \in \mathbf{\Pi} \times \mathscr{P}_0$.

$$\sup_{h \in \mathbf{H}} \mathbb{E}\left[\varphi_k h(Y,T)\right] - \max_{h \in \mathbf{H}_n} \mathbb{E}\left[\varphi_k h(Y,T)\right] \to 0,$$

$$\inf_{\{h_{k'}\} \in \mathbf{H}^{K-1}} \left[1 - \sum_{k' \neq k} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)]\right] - \min_{\{h_{k'}\} \in \mathbf{H}_n^{K-1}} \left[1 - \sum_{k' \neq k} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)]\right] \to 0.$$

Proof of Lemma A.2. See Appendix A.3.2. □

Given the convergence in Lemma A.2, we can now proceed to the inference stage.

A.1.2 Confidence Intervals of the LATEs and the LATMs

For simplicity, hereafter we use θ_k to represent $\alpha_{k,k-1}$ or Δp_k, and use $\Theta_k^{\theta}(\mathbf{P})$ to represent $\Theta_k^{\alpha}(\mathbf{P})$ (when $\theta_k = \alpha_{k,k-1}$) or $\Theta_k^{p}(\mathbf{P})$ (when $\theta_k = \Delta p_k$). In addition, and with slight abuse of notation, we also use Θ to represent the parameter space of θ_k, and $\Theta = [-1,1]$ when $\theta_k = \Delta p_k$.

Given $\eta \in (0, 0.5)$ and $\eta_\pi \in (0, \eta/2)$, the $(1-\eta-2\eta_\pi)$-confidence interval of θ_k is:

$$\mathscr{C}_{\theta_k}(\eta + 2\eta_\pi) := \bigcup_{\pi_k \in \mathscr{C}_{\pi_k}(\eta_\pi),\ \pi_{k-1} \in \mathscr{C}_{\pi_{k-1}}(\eta_\pi)} \left\{\theta_k \in \Theta :\ \tau(\theta_k, \pi_k, \pi_{k-1}) \leq c_k(\eta)\right\}, \tag{A7}$$

where the test statistic $\tau(\theta_k, \pi_k, \pi_{k-1})$ and the critical value $c_k(\eta)$ are defined in the two-step multiplier bootstrap procedure of Chernozhukov et al. (2019) described in our Appendix A.1.4. The testing procedure is for the p_n moment inequalities in Lemma A.1 under $\mathbf{H}_n$.[39] The following Theorem holds for both $\theta_k = \alpha_{k,k-1}$ and $\theta_k = \Delta p_k$.

Theorem A.1. *Let Assumption 2.1, 2.2, 2.3, A.1 and A.2 hold. Construct the test statistic* $\tau(\theta_k, \pi_k, \pi_{k-1})$ *and the critical value* $c_k(\eta)$ *by the two-step multiplier bootstrap described in Appendix A.1.4.*

(i) The confidence interval $\mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)$ *controls the asymptotic size uniformly over* $\mathscr{P}_0$,

$$\liminf_{n \to \infty} \inf_{\mathbf{P} \in \mathscr{P}_0,\ \theta_k \in \Theta_k^{\theta}(\mathbf{P})} Pr\left[\theta_k \in \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)\right] \geq 1 - \eta - 2\eta_\pi.$$

(ii) Given $\pi_k^0 = Pr(Z = z_k)$ *and* $\pi_{k-1}^0 = Pr(Z = z_{k-1})$, *for any fixed alternative* $\theta_k \notin \Theta_k^{\theta}(\mathbf{P})$,

$$\lim_{n \to \infty} Pr\left[\tau\left(\theta_k, \pi_k^0, \pi_{k-1}^0\right) \leq c_k(\eta)\right] = 0.$$

Proof of Theorem A.1. See Appendix A.3.3. □

Theorem A.1-(i) shows that the confidence interval $\mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)$ defined in (A7) covers any point in $\Theta_k^{\theta}(\mathbf{P})$ with probability at least $(1 - \eta - 2\eta_\pi)$ uniformly over $\mathscr{P}_0$. In addition, Theorem A.1-(ii) tells us that the confidence interval, evaluated at the true π_{k-1}^0, π_k^0, will exclude any point outside $\Theta_k^{\theta}(\mathbf{P})$ with probability going to one. Hence, it is reasonable to expect that $\mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)$ will not be too conservative for large enough sample sizes, as long as the standard $\sqrt{n}$-consistent estimator of π and its associated confidence interval are used to construct the confidence interval.

In practice, a simpler version of the confidence interval of θ_k, denoted by $\hat{\mathscr{C}}_{\theta_k}(\eta)$, can be implemented as below:

$$\hat{\mathscr{C}}_{\theta_k}(\eta) := \left\{\theta_k \in \Theta :\ \tau(\theta_k, \hat{\pi}_k, \hat{\pi}_{k-1}) \leq \hat{c}_k(\eta)\right\}, \tag{A8}$$

[39] The critical value $c_k(\eta)$ also depends on $(\theta_k, \pi_k, \pi_{k-1})$. For notation simplicity, we simplify it to be $c_k(\eta)$.

where $(\hat{\pi}_k, \hat{\pi}_{k-1})$ are $\sqrt{n}$-consistent estimators of (π_k, π_{k-1}), and $\hat{c}_k(\eta)$ is obtained in the two-step multiplier bootstrap using $\hat{\pi}_k, \hat{\pi}_{k-1}$. The asymptotic properties of the confidence interval, constructed by testing the moment inequalities with estimated nuisance parameters, are considered in Appendix B.2 of Chernozhukov et al. (2019).[40] Moreover, the simpler version confidence interval $\hat{\mathscr{C}}_{\theta_k}(\eta)$ in (A8) can also be applied to construct $\mathscr{C}^{\alpha}(\beta^{\alpha})$ or $\mathscr{C}^{p}(\beta^{p})$ for practical purpose.

A.1.3 Confidence Intervals of α^{IV}

Given the confidence intervals of $\alpha_{k,k-1}$ and Δp_k, we can now move on to construct confidence intervals of α^{IV}. These are the only details that are reported in the main text in section 3.

A.1.4 Two-Step Multiplier Bootstrap: Details

In this final sub-section, we provide further details of the two-step multiplier bootstrap method proposed in Chernozhukov et al. (2019) and how to use it to construct confidence intervals in our paper. For the sake of notation consistency, we use θ_k and π_k, π_{k-1} to denote the parameter of interest and the nuisance parameters, respectively.

Denote the data as $\{V_i\}_{i=1}^n = \{Y_i, T_i, S_i, Z_i\}_{i=1}^n$ and $V = \{Y, T, S, Z\}$. As slight abuse of notations, for $h_k, h_{k'} \in \mathbf{H}_n$ and $Q = \{Y, T\}$, define moment functions in Lemma A.1 as

$$
\begin{aligned}
&g_1(V, \theta_k, \pi_k, \pi_{k-1}) = -\varphi_k \text{sign}(\theta_k) Q, \\
&g_j(V, \theta_k, \pi_k, \pi_{k-1}) = \varphi_k \left[|\theta_k| h_k(Y, T) - \text{sign}(\theta_k) Q \right], \text{ for } j = 2, ..., \kappa_n + 1, \\
&g_j(V, \theta_k, \pi_k, \pi_{k-1}) = \varphi_k \text{sign}(\theta_k) Q - |\theta_k| \left(1 - \sum_{k' \neq k} \varphi_{k'} h_{k'}(Y, T) \right), \text{ for } j = \kappa_n + 2, ..., p_n,
\end{aligned}
$$

where $Q = Y$ when $\theta_k = \alpha_{k,k-1}$ and $Q = T$ when $\theta_k = \Delta p_k$. Denote

$$
\begin{aligned}
\hat{m}_j(\theta_k, \pi_k, \pi_{k-1}) &= \frac{1}{n} \sum_{i=1}^{n} g_j(V_i, \theta_k, \pi_k, \pi_{k-1}), \\
\hat{\sigma}_j^2(\theta_k, \pi_k, \pi_{k-1}) &= \frac{1}{n} \sum_{i=1}^{n} \left[g_j(V_i, \theta_k, \pi_k, \pi_{k-1}) - \hat{m}_j(\theta_k, \pi_k, \pi_{k-1}) \right]^2.
\end{aligned}
$$

The test statistic for $H_0 : \mathbb{E}[g_j(\theta_k, \pi_k, \pi_{k-1})] \leq 0$ for all $j = 1, 2, ..., p_n$ is defined as

$$
\tau(\theta_k, \pi_k, \pi_{k-1}) = \max_{1 \leq j \leq p_n} \frac{\sqrt{n} \hat{m}_j(\theta_k, \pi_k, \pi_{k-1})}{\hat{\sigma}_j(\theta_k, \pi_k, \pi_{k-1})}
$$

Given the above test statistic, $\tau(\theta_k, \pi_k, \pi_{k-1})$, its critical value $c_k(\eta)$ can be calculated by the two-step multiplier bootstrap procedure, including two main steps: moment inequalities selection and approximating the distribution of the test statistic by bootstrapping. For selecting inequalities, we use $\beta = \beta_n$ as size and follow Chernozhukov et al. (2019) that β_n satisfies $\beta_n \leq \eta/3$ and $\log(1/\beta_n) \leq C_1 \log(n)$. Detailed algorithm of calculating critical value is given below.

[40] Although our moment inequalities in Lemma A.1 fail to satisfy the necessary condition of the uniform size control for the simpler $\hat{\mathscr{C}}_{\theta_k}(\eta)$ (Comment B.2 of Chernozhukov et al. (2019)), simulation results in (Tommasi and Zhang, 2020) show that $\hat{\mathscr{C}}_{\theta_k}(\eta)$ still performs good in terms of achieving the desired coverage rates and of indicating the sign and the true value of the treatment effect in all the DGP designs considered in this paper. Therefore, practitioners may apply the simpler version for practical purpose, because it is less computational consuming and less conservative.

A.1.5 Algorithm

(1) Generate i.i.d. standard normal random variables $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ that are independent of $\{V_i\}_{i=1}^n$.

(2) Construct the multiplier bootstrap test statistic,

$$\tau^{B,1}(\theta_k, \pi_k, \pi_{k-1}) = \max_{1\le j\le p_n} \frac{\sqrt{n}\hat{m}_j^B(\theta_k, \pi_k, \pi_{k-1})}{\hat{\sigma}_j(\theta_k, \pi_k, \pi_{k-1})},$$

where $\hat{m}_j^B(\theta_k, \pi_k, \pi_{k-1}) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \big[g_j(V_i, \theta_k, \pi_k, \pi_{k-1}) - \hat{m}_j(\theta_k, \pi_k, \pi_{k-1})\big]$. Repeat the process in (1)-(2) N^B times, and get the conditional $(1-\beta_n)$-quantile of $\tau^{B,1}(\theta_k, \pi_k, \pi_{k-1})$ given $\{V_i\}_{i=1}^n$, denoted as $c_k^{B,1}(\beta_n)$.

(3) Select inequalities and define the set $\hat{J}_k$ by

$$\hat{J}_k = \left\{ j = 1, 2, ..., p_n : \frac{\sqrt{n}\hat{m}_j(\theta_k, \pi_k, \pi_{k-1})}{\hat{\sigma}_j(\theta_k, \pi_k, \pi_{k-1})} > -2c_k^{B,1}(\beta_n) \right\}.$$

(4) Calculate the critical value $c_k(\eta)$ for the test statistic $\tau(\theta_k, \pi_k, \pi_{k-1})$ as follows. Construct the multiplier bootstrap test statistic,

$$\tau^{B,2}(\theta_k, \pi_k, \pi_{k-1}) = \max_{j\in\hat{J}_k} \frac{\sqrt{n}\hat{m}_j^B(\theta_k, \pi_k, \pi_{k-1})}{\hat{\sigma}_j(\theta_k, \pi_k, \pi_{k-1})},$$

where $\tau^{B,2}(\theta_k, \pi_k, \pi_{k-1}) = 0$ if $\hat{J}_k$ is empty. The critical value $c_k(\eta)$ is the conditional $(1-\eta+2\beta_n)$-quantile of $\tau^{B,2}(\theta_k, \pi_k, \pi_{k-1})$ given $\{V_i\}_{i=1}^n$.

Further details about the algorithm we use to construct the confidence intervals in Stata are given in Lin, Tommasi, and Zhang (2021).

A.1.6 Proof of Corollary 3.1

Proof of Corollary 3.1. (i) Consider $\mathscr{C}^\alpha(\beta^\alpha)$. Denote the set $\mathscr{H}_{0,n}^\alpha = \{(\theta, \mathbf{P}) \in \Theta^\alpha(\mathbf{P}) \times \mathscr{P}_0\}$. Since $\Theta^\alpha(\mathbf{P}) = \bigcup_{k=1,2,..,K} \Theta_k^\alpha(\mathbf{P})$, for $\forall\theta \in \Theta^\alpha(\mathbf{P})$, there exists a k^* such that $\theta \in \Theta_{k^*}^\alpha(\mathbf{P})$. Now, for $\forall\theta \in \Theta^\alpha(\mathbf{P})$, the probability such a θ does not lie in $\mathscr{C}^\alpha(\beta^\alpha)$ is

$$\Pr[\theta \notin \mathscr{C}^\alpha(\beta^\alpha)] \le \Pr\Big[\theta \notin \mathscr{C}_{\alpha_{k^*,k^*-1}}(\eta + 2\eta_\pi)\Big],$$

where the inequality is due that $\theta \notin \mathscr{C}^\alpha(\beta^\alpha)$ implies such θ not in any $\mathscr{C}_{\alpha_{k,k-1}}(\eta + 2\eta_\pi)$ for $k = 1, 2, ..., K$. It yields from the above inequality and Theorem A.1-(i) that

$$\begin{aligned}\liminf_{n\to\infty} \inf_{(\theta,\mathbf{P})\in\mathscr{H}_{0,n}^\alpha} \Pr[\theta \in \mathscr{C}^\alpha(\beta^\alpha)] &\ge \liminf_{n\to\infty} \inf_{\theta\in\Theta_{k^*}^\alpha(\mathbf{P}),\mathbf{P}\in\mathscr{P}_0} \Pr\Big[\theta \in \mathscr{C}_{\alpha_{k^*,k^*-1}}(\eta + 2\eta_\pi)\Big] \\ &\ge 1 - (\eta + 2\eta_\pi).\end{aligned}$$

(ii) Consider $\mathscr{C}^p(\beta^p)$. Denote set $\mathscr{H}_{0,n}^p = \{(\theta, \mathbf{P}) \in \Theta^p(\mathbf{P}) \times \mathscr{P}_0\}$. Recall $\alpha^{Mis} = \frac{\mathrm{Cov}(Y, g(Z))}{\mathrm{Cov}(T, g(Z))}$. For $\forall\theta \in \Theta^p(\mathbf{P})$, there exists a Δp such that $\theta = \alpha^{Mis} \times \Delta p$ and $\Delta p \in \bigcup_{k=1,2,..,K} \Theta_k^p(\mathbf{P})$. Then, there exists a $k^* \in \{1, 2, ..., K\}$ such that $\Delta p \in \Theta_{k^*}^p(\mathbf{P})$. Hence, for $\forall\theta \in \Theta^p(\mathbf{P})$, probability such a θ does not lie in $\mathscr{C}^p(\beta^p)$ is

$$\Pr(\theta \notin \mathscr{C}^p(\beta^p))$$

$$
\begin{aligned}
=&\Pr\left[\theta \notin \mathscr{C}^p(\beta^p), \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\theta \notin \mathscr{C}^p(\beta^p), \alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq&\Pr\left[\Delta p \notin \bigcup_{k=1,2,..,K} \mathscr{C}_{\Delta p_k}(\eta + 2\eta_\pi), \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq&\Pr\left[\Delta p \notin \bigcup_{k=1,2,..,K} \mathscr{C}_{\Delta p_k}(\eta + 2\eta_\pi)\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq&\Pr\left[\Delta p \notin \mathscr{C}_{p_{1,k^*} - p_{0,k^*}}(\eta + 2\eta_\pi)\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right],
\end{aligned}
$$

where the last inequality is due that Δp does not lie in any $\mathscr{C}_{\Delta p_k}(\eta + 2\eta_\pi)$ for $\forall k = 1, 2, ..., K$, which implies $\Delta p \notin \mathscr{C}_{p_{1,k^*} - p_{0,k^*}}(\eta + 2\eta_\pi)$. By Theorem A.1 and Assumption A.1,

$$
\begin{aligned}
\liminf_{n\to\infty} \inf_{(\theta,\mathbf{P}) \in \mathscr{H}^p_{0,n}} \Pr[\theta \in \mathscr{C}^p(\beta^p)] \geq& \liminf_{n\to\infty} \inf_{(\theta,\mathbf{P}) \in \mathscr{H}^p_{0,n}} \Pr\left[\Delta p \in \mathscr{C}_{p_{1,k^*} - p_{0,k^*}}(\eta + 2\eta_\pi)\right] \\
&- \liminf_{n\to\infty} \sup_{\mathbf{P} \in \mathscr{P}_0} \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\geq& 1 - (\eta + 2\eta_\pi + \eta_{\alpha^{Mis}}).
\end{aligned}
$$

(iii) Similarly for $\mathscr{C}^\xi(\beta^\xi)$, let $\mathscr{H}^\xi_{0,n}$ be the set $\mathscr{H}^\xi_{0,n} = \left\{(\theta, \mathbf{P}) \in \Theta^\xi(\mathbf{P}) \times \mathscr{P}_0\right\}$. Then, for $\forall \theta \in \Theta^\xi(\mathbf{P})$, there is a Δp such that $\theta = \alpha^{Mis} \times \Delta p$ and $\Delta p \in [\underline{\xi}, \overline{\xi}]$. Now, for $\forall \theta \in \Theta^\xi(\mathbf{P})$, the probability such a θ does not lie in $\mathscr{C}^\xi(\beta^\xi)$ is

$$
\begin{aligned}
\Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi)\right] =& \Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi), \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi), \alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq& \Pr\left[\Delta p \notin \left[\underline{\xi}, \overline{\xi}\right], \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq& \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right],
\end{aligned}
$$

where the last inequality is because $\Pr\left[\Delta p \notin \left[\underline{\xi}, \overline{\xi}\right], \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] = 0$ for $\theta \in \Theta^\xi(\mathbf{P})$. Then

$$
\liminf_{n\to\infty} \inf_{(\theta,\mathbf{P}) \in \mathscr{H}^\xi_{0,n}} \Pr\left[\theta \in \mathscr{C}^\xi(\beta^\xi)\right] \geq 1 - \eta_{\alpha^{Mis}}.
$$

In addition, if what we have is a η_ξ-confidence interval of $[\underline{\xi}, \overline{\xi}]$ (or, of ξ), denoted by $\mathscr{C}_\xi(\eta_\xi)$, and construct the confidence interval for α^{IV} as $\mathscr{C}^\xi(\beta^\xi) = \bigcup_{\alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \{\alpha \times \Delta p : \Delta p \in \mathscr{C}_\xi(\eta_\xi)\}$, then for $\forall \theta \in \Theta^\xi(\mathbf{P})$, the probability such a θ does not lie in $\mathscr{C}^\xi(\beta^\xi)$ is

$$
\begin{aligned}
\Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi)\right] =& \Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi), \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\theta \notin \mathscr{C}^\xi(\beta^\xi), \alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq& \Pr\left[\Delta p \notin \mathscr{C}_\xi(\eta_\xi), \alpha^{Mis} \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
\leq& \Pr\left[\Delta p \notin \mathscr{C}_\xi(\eta_\xi)\right] + \Pr\left[\alpha^{Mis} \notin \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})\right] \\
=& \eta_\xi + \eta_{\alpha^{Mis}}.
\end{aligned}
$$

It leads to that $\beta^\xi = \eta_\xi + \eta_{\alpha^{Mis}}$ and $\liminf_{n\to\infty} \inf_{(\theta,\mathbf{P}) \in \mathscr{H}^\xi_{0,n}} \Pr\left[\theta \in \mathscr{C}^\xi(\beta^\xi)\right] \geq 1 - (\eta_\xi + \eta_{\alpha^{Mis}})$. □

A.2 Extension: Partial Identification of α^{IV} using Multiple Treatment Proxies

Consider two treatment proxies T and S, where T is the binary indicator used in Section 2.2 and S is a discrete or continuous variable (hence, for the moment, we do not restrict the support of S). The extension to multiple treatment measurements is straightforward, hence we do not discuss it here. The bounds of $\alpha_{k,k-1}$ and Δp_k under two treatment measures T and S, and their sharpness results, are summarized by Lemma A.3. Denote $\Theta_k^{p^W}(\mathbf{P})$ as $\Theta_k^p(\mathbf{P})$ associated with $W \in \{T,S\}$.

Lemma A.3. *Let Assumption 2.1-(ii)-(iv) and 2.3 hold, and suppose Assumption 2.2-(i) is satisfied by both T and S.*

(i) For $\forall k = 1,2,...,K$,

(1) *if $TV_{(Y,T,S),k} = 0$, then $\Theta_k^\alpha(\mathbf{P}) = \Theta$; if $TV_{(Y,T,S),k} > 0$, then*

$$\Theta_k^\alpha(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) < 0. \end{cases} \tag{A9}$$

(2) *if $\max_{0\leq m\leq K} TV_{(Y,T,S),m} = 0$, then $\Theta_k^\alpha(\mathbf{P}) = \Theta$ is the sharp identified set of $\alpha_{k,k-1}$; if $TV_{(Y,T,S),k} > 0$ and $TV_{(Y,T,S),k'} = 0$ for all $k' \neq k$, then $\Theta_k^\alpha(\mathbf{P})$ in (A9) is the sharp identified set of $\alpha_{k,k-1}$.*

(ii) For $\forall k = 1,2,...,K$ and $\forall W \in \{T,S\}$,

(1) *if $TV_{(Y,T,S),k} = 0$, then $\Theta_k^{p^W}(\mathbf{P}) = [-1,1]$; if $TV_{(Y,T,S),k} > 0$, then*

$$\Theta_k^{p^W}(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(W|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}, \frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}}\right], & \text{if } \Delta_k \mathbb{E}(W|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(W|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}}, \frac{\Delta_k \mathbb{E}(W|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}\right], & \text{if } \Delta_k \mathbb{E}(W|Z) < 0. \end{cases} \tag{A10}$$

(2) *if $\max_{0\leq m\leq K} TV_{(Y,T,S),m} = 0$, then $\Theta_k^{p^W}(\mathbf{P}) = [-1,1]$ is the sharp identified set of Δp_k^W; if $TV_{(Y,T,S),k} > 0$ and $TV_{(Y,T,S),k'} = 0$ for all $k' \neq k$, then $\Theta_k^{p^W}(\mathbf{P})$ in (A10) is the sharp identified set of Δp_k^W.*

Proof of Lemma A.3. The proof of Lemma A.3-(i) is similar to the proof of Lemma A.3-(ii), so we only consider (ii). For (ii), we use Δp_k^T as an example, analogue proof can deliver the results for Δp_k^S. The same proof of Lemma 2.4, together with Lemma A.4, can be used to get the results for Δp_k^T, with $H = 0.5 \times \text{sign}(\Delta_k f_{Y,T,S|Z}(Y,T,S))$, and change $P^*_{f_1,f_0}$, P^*_L, P^*_U as follows. $P^*_{f_1,f_0}$ becomes to

$$\begin{aligned} Z \sim f_Z, \;\; D_k|_Z = 1 \text{ for all } k = 0,1,...,K, \\ (Y_1,T_1,S_1)|_{(\{D_k\}_{k=0}^K,Z)} \sim \begin{cases} f_{(Y,T,S)}, & \text{if all } D_k \text{ are equal,} \\ f_{Y,S} f_1, & \text{if at least one } D_k \neq D_{k-1}. \end{cases} \\ (Y_0,T_0,S_0)|_{(\{D_k\}_{k=0}^K,Z)} \sim f_{Y,S} f_0, \end{aligned}$$

P^*_L is constructed as

$$Z \sim f_Z, \;\; (D_{k-1},D_k)|_Z = (0,1), \;\; D_l \leq D_w \text{ if } l < w$$

$$(Y_1, T_1, S_1)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_{(Y,T,S)|Z=z_k}$$
$$(Y_0, T_0, S_0)|_{(\{D_k\}_{k=0}^K, Z)} \sim f_{(Y,T,S)|Z=z_{k-1}}.$$

P_U^* should be changed to

$$Z \sim f_Z,$$
$$(D_{k-1}, D_k)|_Z = \begin{cases} (0,1), & D_l \le D_w \text{ if } l < w, \quad \text{with probability } \Delta_k \mathbb{E}[H|Z], \\ (0,0), & D_l = D_w \text{ for all } l, w, \quad \text{with probability } \Pr(H = -0.5|Z = z_k), \\ (1,1), & D_l = D_w \text{ for all } l, w, \quad \text{with probability } \Pr(H = 0.5|Z = z_{k-1}). \end{cases}$$
$$(Y_1, T_1, S_1)|_{(\{D_k\}_{k=0}^K, Z)} \sim \begin{cases} \frac{\Delta_k f_{(Y,T,S,H)|Z}(y,t,s,0.5)}{\Delta_k \mathbb{E}[H|Z]}, & \text{if } D_{k-1} < D_k, \\ f_{(Y,T,S)|H=0.5, Z=z_{k-1}}(y,t,s), & \text{if } D_{k-1} = D_k \end{cases}$$
$$(Y_0, T_0, S_0)|_{(\{D_k\}_{k=0}^K, Z)} \sim \begin{cases} -\frac{\Delta_k f_{(Y,T,S,H)|Z}(y,t,s,-0.5)}{\Delta_k \mathbb{E}[H|Z]}, & \text{if } D_{k-1} < D_k, \\ f_{(Y,T,S)|H=-0.5, Z=z_k}(y,t,s), & \text{if } D_{k-1} = D_k. \end{cases}$$

□

We then introduce the Lemma below, which shows that, when multiple proxies are available, the identified set of compliers' probability can be improved.

Lemma A.4. *Let Assumption 2.1-(ii)-(iv) and 2.3 hold, and suppose Assumption 2.2-(i) is satisfied by both T and S. For $k = 1, 2, ..., K$,*

$$TV_{(Y,T),k} \le TV_{(Y,T,S),k} \le Pr(C_k) \le 1 - \sum_{k' \ne k} TV_{(Y,T,S),k'} \le 1 - \sum_{k' \ne k} TV_{(Y,T),k'}.$$

Proof of Lemma A.4. Consider two treatment proxies T and S. If S is discrete, we can simply replace the second integral in the equation below by a summation over the support of S. By the triangle inequality, we have that for $k = 1, 2, ..., K$

$$\begin{aligned} TV_{(Y,T,S),k} &= \frac{1}{2} \sum_{t=0,1} \iint \left| f_{(Y,T,S)|Z=z_k}(y,t,s) - f_{(Y,T,S)|Z=z_{k-1}}(y,t,s) \right| d\mu_Y(y) d\mu_S(s) \\ &\ge \frac{1}{2} \sum_{t=0,1} \int \left| \int f_{(Y,T,S)|Z=z_k}(y,t,s) - f_{(Y,T,S)|Z=z_{k-1}}(y,t,s) d\mu_S(s) \right| d\mu_Y(y) \\ &= \frac{1}{2} \sum_{t=0,1} \int \left| f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t) \right| d\mu_Y(y) \\ &= TV_{(Y,T),k}. \end{aligned}$$

In addition, we can get

$$f_{(Y,T,S)|Z=z_k}(y,t,s) - f_{(Y,T,S)|Z=z_{k-1}}(y,t,s) = \Pr(C_k) \left[f_{(Y_1,T_1,S_1)|C_k}(y,t,s) - f_{(Y_0,T_0,S_0)|C_k}(y,t,s) \right].$$

Then, $TV_{(Y,T,S),k} \le \Pr(C_k) \le 1 - \sum_{k' \ne k} TV_{(Y,T,S),k'}$ follows from same proof of Lemma 2.2. □

Lemma A.4 says that the bounds of $\Pr(C_k)$ shrink from $[TV_{(Y,T),k}, 1 - \sum_{k' \ne k} TV_{(Y,T),k'}]$, when only a single proxy T is used, to $[TV_{(Y,T,S),k}, 1 - \sum_{k' \ne k} TV_{(Y,T,S),k'}]$, when both T and S are used. The improved bound of $\Pr(C_k)$ also leads to narrower bounds of (i) LATEs, (ii) LATMs, and (iii) the IV estimand α^{IV}. In what follows, we focus on refining our partial identification strategies for α^{IV} when multiple treatment proxies are available.

Lemma A.5. *Suppose both T and S are binary variables. Define another treatment proxy as $TS = 1[T = 1, S = 1]$. Then, $TV_{(Y,TS),k} \leq TV_{(Y,T,S),k}$ for all k.*

Proof of Lemma A.5.

$$\begin{aligned}
TV_{(Y,TS),k} =& \frac{1}{2}\sum_{ts=0,1}\int \left|f_{(Y,TS)|Z=z_k}(y,ts) - f_{(Y,TS)|Z=z_{k-1}}(y,ts)\right| d\mu_Y(y) \\
=& \frac{1}{2}\sum_{ts=0,1}\int \left|\sum_{s=0,1}\left[f_{(Y,TS,S)|Z=z_k}(y,ts,s) - f_{(Y,TS,S)|Z=z_{k-1}}(y,ts,s)\right]\right| d\mu_Y(y) \\
\leq& \frac{1}{2}\sum_{ts=0,1}\sum_{s=0,1}\int \left|f_{(Y,TS,S)|Z=z_k}(y,ts,s) - f_{(Y,TS,S)|Z=z_{k-1}}(y,ts,s)\right| d\mu_Y(y) \\
=& \frac{1}{2}\int \left|f_{(Y,TS,S)|Z=z_k}(y,0,0) - f_{(Y,TS,S)|Z=z_{k-1}}(y,0,0)\right| \\
&\quad + \left|f_{(Y,TS,S)|Z=z_k}(y,0,1) - f_{(Y,TS,S)|Z=z_{k-1}}(y,0,1)\right| \\
&\quad + \left|f_{(Y,TS,S)|Z=z_k}(y,1,0) - f_{(Y,TS,S)|Z=z_{k-1}}(y,1,0)\right| \\
&\quad + \left|f_{(Y,TS,S)|Z=z_k}(y,1,1) - f_{(Y,TS,S)|Z=z_{k-1}}(y,1,1)\right| d\mu_Y(y).
\end{aligned}$$

By definition of TS, we have

$$\begin{aligned}
f_{(Y,TS,S)|Z=z_k}(y,0,0) - f_{(Y,TS,S)|Z=z_{k-1}}(y,0,0) =& f_{(Y,T,S)|Z=z_k}(y,0,0) - f_{(Y,T,S)|Z=z_{k-1}}(y,0,0) \\
&+ f_{(Y,T,S)|Z=z_k}(y,1,0) - f_{(Y,T,S)|Z=z_{k-1}}(y,1,0) \\
f_{(Y,TS,S)|Z=z_k}(y,0,1) - f_{(Y,TS,S)|Z=z_{k-1}}(y,0,1) =& f_{(Y,T,S)|Z=z_k}(y,0,1) - f_{(Y,T,S)|Z=z_{k-1}}(y,0,1) \\
f_{(Y,TS,S)|Z=z_k}(y,1,0) - f_{(Y,TS,S)|Z=z_{k-1}}(y,1,0) =& 0 \\
f_{(Y,TS,S)|Z=z_k}(y,1,1) - f_{(Y,TS,S)|Z=z_{k-1}}(y,1,1) =& f_{(Y,T,S)|Z=z_k}(y,1,1) - f_{(Y,T,S)|Z=z_{k-1}}(y,1,1),
\end{aligned}$$

thus,

$$TV_{(Y,TS),k} \leq \frac{1}{2}\sum_{t,s=0,1}\int \left|f_{(Y,T,S)|Z=z_k}(y,t,s) - f_{(Y,T,S)|Z=z_{k-1}}(y,t,s)\right| d\mu_Y(y) = TV_{(Y,T,S),k}.$$

□

First strategy. For the multiple treatment proxies case, the Corollary below gives the sign of α^{IV} and the expression of $\Theta^{\alpha}(\mathbf{P})$.

Corollary A.1. *Let Assumption 2.1, 2.3 hold. Suppose T and S satisfy Assumption 2.2.*

(i) If $\Delta_k\mathbb{E}(Y|Z) > 0$ for all $k = 1, 2, ..., K$, then $\alpha^{IV} > 0$ and

$$\Theta^{\alpha}(\mathbf{P}) = \left[\min_{k\in\{1,2,...,K\}}\left\{\frac{\Delta_k\mathbb{E}(Y|Z)}{1-\sum_{k'\neq k}TV_{(Y,T,S),k'}}\right\}, \max_{k\in\{1,2,...,K\}}\left\{\frac{\Delta_k\mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}\right\}\right].$$

(ii) If $\Delta_k\mathbb{E}(Y|Z) < 0$ for all $k = 1, 2, ..., K$, then $\alpha^{IV} < 0$ and

$$\Theta^{\alpha}(\mathbf{P}) = \left[\min_{k\in\{1,2,...,K\}}\left\{\frac{\Delta_k\mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}\right\}, \max_{k\in\{1,2,...,K\}}\left\{\frac{\Delta_k\mathbb{E}(Y|Z)}{1-\sum_{k'\neq k}TV_{(Y,T,S),k'}}\right\}\right].$$

Proof of Corollary A.1. The proof follows from Theorem 2.2, Lemma A.4, and Lemmas A.1 and A.3 in Appendix A. □

If we relax Corollary A.1-(i) by $\Delta_k \mathbb{E}(Y|Z) \geq 0$, while keeping $TV_{(Y,T,S),k} > 0$ for all k, the expression of $\Theta^{\alpha}(\mathbf{P})$ is still valid and $\alpha^{IV} \geq 0$. Similar arguments apply for A.1-(ii). If the direction consistency of LATEs does not hold, the general form of $\Theta^{\alpha}(\mathbf{P})$ will simply be the union of the $\{\Theta_k^{\alpha}(\mathbf{P})\}_{k=1}^{K}$ under multiple proxies given in Lemma A.3, while we fail to recover the sign of α^{IV}. The identification gains of $\Theta^{\alpha}(\mathbf{P})$ in Corollary A.1 are only due to the improvement of the possible region for $\Pr(C_k)$, from

$$[TV_{(Y,T),k}, 1 - \sum_{k' \neq k} TV_{(Y,T),k'}]$$

with single proxy T, to

$$[TV_{(Y,T,S),k}, 1 - \sum_{k' \neq k} TV_{(Y,T,S),k'}]$$

with multiple proxies (T,S).

Second and Third strategy. For our second and third partial identification strategies with multiple treatment proxies, we require all proxies to be binary. This is because both strategies rely on the existence of the LATMs, Δp_k, for all available proxies.

Denote the estimand α^{Mis} associated with $W \in \{T,S\}$, as $\alpha^{Mis,W}$. Furthermore, denote the LATMs, for $W \in \{T,S\}$, as Δp_k^W. The sign of α^{IV} and $\Theta^{p}(\mathbf{P})$, using our second identification strategy with multiple treatment proxies, are characterized by the following Corollary.

Corollary A.2. *Let Assumption 2.1, 2.3 hold. T and S are both binary and satisfy Assumption 2.2. Suppose $\Delta_k \mathbb{E}(W|Z) \geq 0$ for $\forall k = 1,2,...,K$ and $W \in \{T,S\}$.*

(i) For $W \in \{T,S\}$, if $\alpha^{Mis,W} \geq 0$, then

$$\Theta^{p}(\boldsymbol{P}) = \Bigg[\max_{W \in \{T,S\}} \min_{k \in \{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(W|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T,S),k'}} \times \alpha^{Mis,W} \right\}, \min_{W \in \{T,S\}} \max_{k \in \{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}} \times \alpha^{Mis,W} \right\} \Bigg].$$

(ii) For $W \in \{T,S\}$, if $\alpha^{Mis,W} < 0$, then

$$\Theta^{p}(\boldsymbol{P}) = \Bigg[\max_{W \in \{T,S\}} \min_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}} \times \alpha^{Mis,W} \right\}, \min_{W \in \{T,S\}} \max_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(W|Z)}{1 - \sum_{k' \neq k} TV_{(Y,T,S),k'}} \times \alpha^{Mis,W} \right\} \Bigg].$$

Proof of Corollary A.2. The proof follows from Theorem 2.3, Lemma A.4, and Lemmas A.1 and A.3 in Appendix A. □

Corollary A.2 states that, by employing multiple treatment measures, there are two sources of gains compared to Corollary 2.3. Firstly, we narrow down the range of $\Pr(C_k)$ by using multiple measurements (T,S). Secondly, we shrink the bound of α^{IV} by intersecting its bounds associated with T and S, respectively. The intersection contributes to tightening the bound of α^{IV}, as long as S contains additional information about the true treatment D, other than those contained in T, leading to different values of $\Delta_k \mathbb{E}(W|Z)$ and $\alpha^{Mis,W}$ with $W \in \{T,S\}$.

Next, for our third partial identification strategy with multiple treatment proxies, denote by $(\underline{\xi}^W, \overline{\xi}^W)$ the lower and upper bounds of the LATMs Δp_k^W, for $W \in \{T,S\}$. Like before, these bounds may come from external sources of information.

Corollary A.3. *Let Assumption 2.1, 2.3 hold. T and S are both binary and satisfy Assumption 2.2. Suppose $0 < \underline{\xi}^T \leq \overline{\xi}^T \leq 1$ and $0 < \underline{\xi}^S \leq \overline{\xi}^S \leq 1$.*

(i) If $\alpha^{Mis,T} \geq 0$ and $\alpha^{Mis,S} \geq 0$, then $\alpha^{IV} \geq 0$ and

$$\Theta^{\xi}(\mathbf{P}) = \left[\max\left\{\underline{\xi}^T \alpha^{Mis,T}, \underline{\xi}^S \alpha^{Mis,S}\right\}, \min\left\{\overline{\xi}^T \alpha^{Mis,T}, \overline{\xi}^S \alpha^{Mis,S}\right\}\right].$$

(ii) If $\alpha^{Mis,T} \leq 0$ and $\alpha^{Mis,S} \leq 0$, then $\alpha^{IV} \leq 0$ and

$$\Theta^{\xi}(\mathbf{P}) = \left[\max\left\{\overline{\xi}^T \alpha^{Mis,T}, \overline{\xi}^S \alpha^{Mis,S}\right\}, \min\left\{\underline{\xi}^T \alpha^{Mis,T}, \underline{\xi}^S \alpha^{Mis,S}\right\}\right].$$

Proof of Corollary A.3. The proof follows directly from Theorem 2.4 and the fact that α^{IV} lies in both of the bounds derived using T and S. □

We can summarize the results of this subsection as follows. When multiple treatment proxies are available, the improvements are, in general, nontrivial compared to those with one binary proxy. This is because different proxies may provide different and relevant information about the true treatment. Again, by intersecting the bounds we can obtain, potentially, even tighter bounds of α^{IV}.

A.3 Extension: Partial Identification Strategies with Covariates

Let X be a vector of observable covariates with support Ω_X. For $\forall x \in \Omega_X$, denote $\pi_k(x) = \Pr(Z = z_k|X = x)$ with $k = 0, 1, ..., K$ and $\Pr(z, x) = \mathbb{E}(D|Z = z, X = x)$. The analysis in this section can be directly extended to conditional on $X \in A$ or $e(X) \in A$, where $e : \Omega_X \to \mathbb{R}$ is a scalar function and $A \subseteq \mathbb{R}$ denotes a set.

Assumption A.3. *(Covariates) Y, D, T, Z and X satisfy the following assumptions:*

(i) (i.i.d.) $(Y_1, Y_0, \{D_k\}_{k=0}^K, T_1, T_0, Z, X)$ are independent and identically distributed across all individuals and have finite first and second moments;

(ii) (Unconfoundedness) $Z \perp (Y_1, Y_0, \{D_k\}_{k=0}^K, T_1, T_0)|X$. For $\forall x \in \Omega_X$, $Pr(z, x)$ with $z \in \Omega_Z$ is a nontrivial function of z and $0 < \pi_k(x) < 1$, $k = 0, 1, ..., K$;

(iii) (First stage) For $\forall x \in \Omega_X$, $Cov(D, g(Z)|X = x) \neq 0$ and $Cov(T, g(Z)|X = x) \neq 0$;

(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 $z_l, z_w \in \Omega_Z$ and all $x \in \Omega_X$, either $Pr(z_l, x) \leq Pr(z_w, x)$ implies $g(z_l) \leq g(z_w)$, or $Pr(z_l, x) \leq Pr(z_w, x)$ implies $g(z_l) \geq g(z_w)$;

Assumption A.4. *(Conditional Informative Treatment Proxy) For all $k = 1, 2, ..., K$, $Pr(T = d|C_k, D = d, X = x) > Pr(T = d|C_k, D = 1 - d, X = x)$, $d = \{0, 1\}$.*

Assumption A.3 and A.4 extend Assumption 2.1, 2.2 and 2.3 to accommodate covariates. They are sufficient to obtain the desired partial identification results.

For $\forall x \in \Omega_X$, define the conditional LATE as $\alpha_{k,k-1}(x) = \mathbb{E}[Y_1 - Y_0|C_k, X = x]$ and the conditional LATM as $\Delta p_k(x) = \mathbb{E}[T_1 - T_0|C_k, X = x] = p_{1,k}(x) - p_{0,k}(x)$, where $p_{d,k}(x) = \Pr(T_d = 1|C_k, X = x)$ and $d = \{0, 1\}$. Our identification target is the conditional IV estimand $\alpha^{IV}(x)$, which can be expressed as a weighted average of the conditional LATEs:

$$\alpha^{IV}(x) := \frac{\text{Cov}(Y, g(Z)|X = x)}{\text{Cov}(D, g(Z)|X = x)} = \sum_{k=1}^{K} \gamma_k^{IV}(x)\alpha_{k,k-1}(x), \tag{A11}$$

with weights

$$\gamma_k^{IV}(x) := \frac{\Pr(C_k|X = x) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)|X = x])\, \pi_l(x)}{\sum_{m=1}^{K} \Pr(C_m|X = x) \sum_{l=m}^{K} (g(z_l) - \mathbb{E}[g(Z)|X = x])\, \pi_l(x)},$$

where $\Pr(C_k|X = x)$ is the conditional probability of compliers group.

Instead of D, suppose we can observe a binary treatment indicator T. In this case, we can obtain the biased conditional IV estimand $\alpha^{Mis}(x)$:

$$\alpha^{Mis}(x) := \frac{\text{Cov}(Y, g(Z)|X = x)}{\text{Cov}(T, g(Z)|X = x)} = \sum_{k=1}^{K} \gamma_k^{Mis}(x)\alpha_{k,k-1}(x), \tag{A12}$$

with weights

$$\gamma_k^{Mis}(x) := \frac{\Pr(C_k|X = x) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)|X = x])\, \pi_l(x)}{\sum_{m=1}^{K} \Delta p_m(x) \Pr(C_m|X = x) \sum_{l=m}^{K} (g(z_l) - \mathbb{E}[g(Z)|X = x])\, \pi_l(x)}.$$

Derivations of (A11) and (A12) can be obtained under Assumption A.3 by applying similar arguments used in the proof of Theorem 2.1 when conditional on X. The relationship between the actual and the biased conditional IV estimands can be summarized by the theorem below.

Theorem A.2. *Let Assumption A.3 hold. Then:*

$$\alpha^{IV}(x) = \xi(x)\alpha^{Mis}(x),$$

where $\xi(x) = \sum_{k=1}^{K} \gamma_k^{IV}(x)\Delta p_k(x)$ *is the weighted average of the conditional LATMs.*

Proof of Theorem A.2. Let $\xi(x) = \gamma_k^{IV}(x)/\gamma_k^{Mis}(x)$. The proof follows directly from the expressions of $\gamma_k^{IV}(x)$, $\gamma_k^{Mis}(x)$ and $\alpha^{Mis}(x)$. □

Given Assumption A.4, without loss of generality, suppose for any given $x \in \Omega_X$, the support of $\Omega_Z = \{z_0, z_1, ..., z_K\}$ is ordered in such a way that $\forall l, w = 0, 1, ..., K, l < w$ implies $\Pr(z_l, x) \leq \Pr(z_w, x)$. Define the conditional total variation distance for any generic random variable A as below:

$$TV_{A,k}(x) = \frac{1}{2}\int |f_{A|Z=z_k,X=x}(a) - f_{A|Z=z_{k-1},X=x}(a)| d\mu_A(a),$$

which bounds the conditional probability of compliers as shown by the lemma below.

Lemma A.6. *Under Assumption A.3 and A.4, for* $k = 1, 2, ..., K$ *and* $\forall x \in \Omega_X$,

$$TV_{(Y,T),k}(x) \leq Pr(C_k|X = x) \leq 1 - \sum_{k' \neq k} TV_{(Y,T),k'}(x).$$

Proof of Lemma A.6. This proof is a direct extension of the proof of Lemma 2.2 conditional on X. □

From the expressions of $\alpha^{IV}(x)$, $\alpha^{Mis}(x)$ and their relationship in Theorem A.2, it is clear that the partial identification for $\alpha_k^{IV}(x)$ relies on the bounds of $\{\alpha_{k,k-1}(x)\}_{k=1}^K$ or of $\{\Delta p_k(x)\}_{k=1}^K$. For notational simplicity, let $\Delta_k \mathbb{E}(A|Z, X = x) = \mathbb{E}(A|Z = z_k, X = x) - \mathbb{E}(A|Z = z_{k-1}, X = x)$. Under Assumption A.3, we have that the conditional LATE satisfies

$$\Delta_k \mathbb{E}(Y|Z, X = x) = \alpha_{k,k-1}(x)P(C_k|X = x). \tag{A13}$$

Similarly, the following equation holds for each $\Delta p_k(x)$:

$$\Delta_k \mathbb{E}(T|Z, X = x) = \Delta p_k(x)P(C_k|X = x). \tag{A14}$$

Given (A13), (A14) and Lemma A.6, we can obtain the following Lemmas that establish analytic bounds of $\alpha_{k,k-1}(x)$ and $\Delta p_k(x)$, denoted by $\Theta_k^{\alpha}(\mathbf{P}, x) \subset \Theta$ and $\Theta_k^{p}(\mathbf{P}, x) \subset [-1, 1]$, respectively.

Lemma A.7. *Let Assumption A.3 and A.4 hold. The results below hold for* $\forall k = 1, 2, ..., K$ *and* $\forall x \in \Omega_X$.

(i) If $TV_{(Y,T),k}(x) = 0$, *then* $\Theta_k^{\alpha}(\mathbf{P}, x) = \Theta$. *Whereas if* $TV_{(Y,T),k}(x) > 0$, *then:*

$$\Theta_k^{\alpha}(\mathbf{P}, x) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}, \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k}(x)}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z, X = x) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z, X = x) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k(x)}}, \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z, X = x) < 0; \end{cases} \tag{A15}$$

(ii) If $\max_{0 \leq m \leq K} TV_{(Y,T),m}(x) = 0$, *then* $\Theta_k^{\alpha}(\mathbf{P}, x) = \Theta$ *is the sharp identified set of* $\alpha_{k,k-1}(x)$. *Whereas, if* $TV_{(Y,T),k}(x) > 0$ *and* $TV_{(Y,T),k'}(x) = 0$ *for all* $k' \neq k$, *then* $\Theta_k^{\alpha}(\mathbf{P}, x)$ *in (A15) is the sharp identified set of* $\alpha_{k,k-1}(x)$.

Proof of Lemma A.7. The proof is a direct extension of the proof of Lemma 2.3 conditional on X. □

The analytic bound for $\Delta p_k(x)$ can be defined in an analogous manner by replacing Y by T, and replacing Θ by $[-1,1]$.

For $x \in \Omega_X$, the bound of $\alpha^{IV}(x)$ can be constructed using either the bounds of $\{\alpha_{k,k-1}(x)\}_{k=1}^K$ or $\{\Delta p_k(x)\}_{k=1}^K$ or external information. The same logic of partial identification Strategies 1, 2 and 3 still holds, thus can be extended straightforwardly to conditional on covariates.

Strategy 1 with covariates. Let Assumption A.3 and A.4 hold. For $\forall x \in \Omega_X$:

(i) Denote $\Theta^{\alpha}(\mathbf{P},x) = \bigcup_{k\in\{1,2,...,K\}} \Theta_k^{\alpha}(\mathbf{P},x)$, then $\alpha^{IV}(x) \in \Theta^{\alpha}(\mathbf{P},x)$;

(ii) If $\Delta_k \mathbb{E}(Y|Z,X=x) > 0$ for all $k = 1,2,...,K$, then $\alpha^{IV}(x) > 0$ and

$$\Theta^{\alpha}(\mathbf{P},x) = \left[\min_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}(x)} \right\}, \max_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k}(x)} \right\} \right].$$

(iii) If $\Delta_k \mathbb{E}(Y|Z,X=x) < 0$ for all $k = 1,2,...,K$, then $\alpha^{IV}(x) < 0$ and

$$\Theta^{\alpha}(\mathbf{P},x) = \left[\min_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k}(x)} \right\}, \max_{k\in\{1,2,...,K\}} \left\{ \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}(x)} \right\} \right].$$

Strategy 2 with covariates. Let Assumption A.3 and A.4 hold. Under the ascending order, all $\Delta_k \mathbb{E}(T|Z,X=x) \geq 0$. For $\forall x \in \Omega_X$:

(i) Denote $\Theta^{p}(\mathbf{P},x) = \left\{ \alpha^{Mis}(x) \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,...,K} \Theta_k^{p}(\mathbf{P},x) \right\}$, where Δp represents any generic value in $\bigcup_{k=1,2,...,K} \Theta_k^{p}(\mathbf{P},x)$. Then, $\alpha^{IV}(x) \in \Theta^{p}(\mathbf{P},x)$.

(ii) If $\alpha^{Mis}(x) \geq 0$, then $\alpha^{IV}(x) \geq 0$ and

$$\Theta^{p}(\mathbf{P},x) = \alpha^{Mis}(x) \times \left[\min_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}(x)} \right\}, \max_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{TV_{(Y,T),k}(x)} \right\} \right],$$

(iii) If $\alpha^{Mis}(x) < 0$, then $\alpha^{IV}(x) < 0$ and

$$\Theta^{p}(\mathbf{P},x) = \alpha^{Mis}(x) \times \left[\min_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{TV_{(Y,T),k}(x)} \right\}, \max_{k=1,2,...,K} \left\{ \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}(x)} \right\} \right].$$

Strategy 3 with covariates. Let Assumption A.3 hold. Suppose there exist two known constants $\underline{\xi}(x) \leq \overline{\xi}(x)$ and $\underline{\xi}(x), \overline{\xi}(x) \in (0,1]$, such that $\underline{\xi}(x) \leq \xi(x) \leq \overline{\xi}(x)$. Then:

(i) If $\alpha^{Mis}(x) \geq 0$, denote $\Theta^{\xi}(\mathbf{P},x) = \left[\underline{\xi}(x)\alpha^{Mis}(x), \overline{\xi}(x)\alpha^{Mis}(x) \right]$. Then, $\alpha^{IV}(x) \geq 0$ and $\alpha^{IV}(x) \in \Theta^{\xi}(\mathbf{P},x)$.

(ii) If $\alpha^{Mis}(x) \leq 0$, denote $\Theta^{\xi}(\mathbf{P},x) = \left[\overline{\xi}(x)\alpha^{Mis}(x), \underline{\xi}(x)\alpha^{Mis}(x) \right]$. Then, $\alpha^{IV}(x) \leq 0$ and $\alpha^{IV}(x) \in \Theta^{\xi}(\mathbf{P},x)$.

where values of $\underline{\xi}(x)$ and $\overline{\xi}(x)$ may be obtained from external sources of information.

Targeting the unconditional IV estimand in the presence of covariates. We conclude this section by showing the technical challenge one would face if, in the case of a discrete or multiple-discrete instrument(s) and when covariates are included, the identification target was a variant of the unconditional IV estimand:

$$\tilde{\alpha}^{IV} = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(D, g(Z))} = \frac{\mathbb{E}\Big\{Y\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big\}}{\mathbb{E}\Big\{D\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big\}} = \frac{\mathbb{E}_X\Big\{\mathbb{E}\Big[Y\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big|X\Big]\Big\}}{\mathbb{E}_X\Big\{\mathbb{E}\Big[D\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big|X\Big]\Big\}}.$$

For the numerator, we have

$$\begin{aligned}
\mathbb{E}\Big[Y\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big|X\Big] =& \sum_{k=0}^{K} \mathbb{E}[Y|Z = z_0, X]\,(g(z_l) - \mathbb{E}[g(Z)])\,\pi_l(X) \\
&+ \sum_{k=1}^{K} \Pr(C_k|X)\alpha_{k,k-1}(X) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l(X) \\
=& \mathbb{E}[Y|Z = z_0, X](\mathbb{E}[g(Z)|X] - \mathbb{E}[g(Z)]) \\
&+ \sum_{k=1}^{K} \Pr(C_k|X)\alpha_{k,k-1}(X) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l(X),
\end{aligned}$$

where if without further restrictions, the first term in the above equation is nonzero. Thus, we impose restrictions that Z is independent to X, and consider two questions: (1) if $\tilde{\alpha}^{IV}$ and $\mathbb{E}_X\left[\alpha^{IV}(X)\right]$ are the same estimand; (2) if the answer is no in (1), whether $\tilde{\alpha}^{IV}$ is a meaningful estimand. The assumption $Z \perp X$ is satisfied in randomized experimental settings where Z is randomly allocated treatment assignment or incentives, while such an independence assumption may be infeasible in other empirical studies.

First, under $Z \perp X$, we know that $\pi_l(X)$ reduces to π_l and

$$\begin{aligned}
\mathbb{E}\Big[Y\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big|X\Big] &= \sum_{k=1}^{K} \Pr(C_k|X)\alpha_{k,k-1}(X) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l, \\
\mathbb{E}\Big[D\Big(g(Z) - \mathbb{E}[g(Z)]\Big)\Big|X\Big] &= \sum_{k=1}^{K} \Pr(C_k|X) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l.
\end{aligned} \tag{A16}$$

Moreover, we have

$$\begin{aligned}
\tilde{\alpha}^{IV} &= \sum_{k=1}^{K} \frac{\mathbb{E}_X\left[\Pr(C_k|X)\alpha_{k,k-1}(X)\right] \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l}{\sum_{k=1}^{K} \mathbb{E}_X\left[\Pr(C_k|X)\right] \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l} \\
&= \sum_{k=1}^{K} \frac{\Pr(C_k) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l}{\sum_{k=1}^{K} \Pr(C_k) \sum_{l=k}^{K} (g(z_l) - \mathbb{E}[g(Z)])\,\pi_l} \times \alpha_{k,k-1} \\
&= \sum_{k=1}^{K} \gamma_k^{IV} \alpha_{k,k-1},
\end{aligned} \tag{A17}$$

where the second line follows from Frölich (2007) that

$$\mathbb{E}_X\left[\Pr(C_k|X)\alpha_{k,k-1}(X)\right] = \mathbb{E}_X\left[\Pr(C_k|X)\right]\mathbb{E}_X\left[\alpha_{k,k-1}(X)\right] = \Pr(C_k)\alpha_{k,k-1}.$$

However,

$$\begin{aligned}\mathbb{E}_X\left[\alpha^{IV}(X)\right] &= \sum_{k=1}^{K}\mathbb{E}_X[\gamma_k^{IV}(X)\alpha_{k,k-1}(X)] \\ &= \sum_{k=1}^{K}\mathbb{E}_X\left[\frac{\Pr(C_k|X=x)\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l}{\sum_{m=1}^{K}\Pr(C_m|X=x)\sum_{l=m}^{K}(g(z_l)-\mathbb{E}[g(Z)])\,\pi_l}\times\alpha_{k,k-1}(X)\right] \\ &\neq \sum_{k=1}^{K}\gamma_k^{IV}\alpha_{k,k-1},\end{aligned}$$

this is because, for two random variables A and B, $\mathbb{E}[AB] \neq \mathbb{E}[A]\mathbb{E}[B]$ and the equality holds if $\Pr(C_k|X=x)$ is invariant with X. One sufficient condition for $\Pr(C_k|X=x) = \Pr(C_k)$ is that $(D,Z) \perp X$, which, however, is quite restrictive and infeasible in many scenarios such as observational studies. Thus, in general $\tilde{\alpha}^{IV}$ and $\mathbb{E}_X\left[\alpha^{IV}(X)\right]$ are not the same.

Second, under the assumption $Z \perp X$, $\tilde{\alpha}^{IV}$ is still a meaningful estimand because it is also a weighted average of unconditional LATEs $\alpha_{k,k-1}$. Given the expression in (A17), by Frölich (2007),

$$\alpha_{k,k-1} = \mathbb{E}_X[\alpha_{k,k-1}(X)] = \frac{\mathbb{E}_X[\Delta_k\mathbb{E}(Y|X,Z)]}{\mathbb{E}_X[\Pr(C_k|X)]},$$

where the numerator is identifiable and the denominator can be bounded via the conditional version of our method. We can then conduct the partial identification for $\tilde{\alpha}^{IV}$ using a method analogue to Strategies 1, 2 and 3. One final remark is, for studies where the assumption $Z \perp X$ is not applicable, we suggest practitioners follow our proposed method in Section 4.2 to incorporate covariates in the presence of discrete or multiple-discrete IV(s).

A.3.1 Proof of Lemma A.1

Proof of Lemma A.1. Recall $\pi_k = \Pr(Z = z_k)$ and $\varphi_k = \frac{1[Z=z_k]\pi_{k-1} - 1[Z=z_{k-1}]\pi_k}{\pi_k\pi_{k-1}}$. Similar to Abadie (2003)'s binary instrument case, we can get for any generic random variable Q and $\forall k$,

$$\begin{aligned}\Delta_k\mathbb{E}[Q|Z] &= \mathbb{E}[Q|Z=z_k] - \mathbb{E}[Q|Z=z_{k-1}] \\ &= \frac{1}{\pi_k}\mathbb{E}\left[\pi_k Q|Z=z_k\right] - \frac{1}{\pi_{k-1}}\mathbb{E}\left[\pi_{k-1}Q|Z=z_{k-1}\right] \\ &= \mathbb{E}\left[\frac{Q\times 1[Z=z_k]}{\pi_k}\right] - \mathbb{E}\left[\frac{Q\times 1[Z=z_{k-1}]}{\pi_{k-1}}\right] \\ &= \mathbb{E}\left[\frac{1[Z=z_k]\pi_{k-1} - 1[Z=z_{k-1}]\pi_k}{\pi_k\pi_{k-1}}Q\right] \\ &= \mathbb{E}[\varphi_k Q], \end{aligned} \tag{A18}$$

where $1[\cdot]$ is the indicator function. In addition, it holds that for $\forall h \in \mathbf{H}$,

$$\begin{aligned}\Delta_k\mathbb{E}[h(Y,T)|Z] &= \sum_{t=0,1}\int h(y,t)\Delta_k f_{(Y,T)|Z}(y,t)d\mu_Y(y) \\ &\leq \frac{1}{2}\sum_{t=0,1}\int \left|\Delta_k f_{(Y,T)|Z}(y,t)\right| d\mu_Y(y) = TV_{(Y,T),k},\end{aligned} \tag{A19}$$

where the inequality is by definition of $h(y,t) \in \{-0.5, 0.5\}$ and the last equality holds if and only if h is such that for all $(y,t) \in \Omega_Y \times \{0,1\}$, $h(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \geq 0$. Moreover, (A19) also implies

that for $\forall h \in \mathbf{H}$,

$$1 - \sum_{k' \neq k} TV_{(Y,T),k'} \leq 1 - \sum_{k' \neq k} \Delta_{k'} \mathbb{E}[h_{k'}(Y,T)|Z], \tag{A20}$$

where the equality holds if and only if $h_{k'}$ is such that $\forall (y,t) \in \Omega_Y \times \{0,1\}$, $h_{k'}(y,t)\Delta_{k'} f_{(Y,T)|Z}(y,t) \geq 0$ for all $k' \neq k$. Given (A19) and (A20) above, we can then rewrite Θ_k^{α} as

$$\begin{aligned}
&-\text{sign}(\alpha_{k,k-1})\Delta_k \mathbb{E}[Y|Z] \leq 0, \\
&|\alpha_{k,k-1}|\Delta_k \mathbb{E}[h(Y,T)|Z] \leq \text{sign}(\alpha_{k,k-1})\Delta_k \mathbb{E}[Y|Z], \text{ for all } h \in \mathbf{H} \\
&\text{sign}(\alpha_{k,k-1})\Delta_k \mathbb{E}[Y|Z] \leq |\alpha_{k,k-1}| \left[1 - \sum_{k' \neq k} \Delta_{k'} \mathbb{E}[h_{k'}(Y,T)|Z] \right], \text{ for all } h_{k'} \in \mathbf{H}.
\end{aligned}$$

Applying (A18) to the above inequalities gives us the desired results. Same arguments can be applied to prove the results for Θ_k^{p}. □

A.3.2 Proof of Lemma A.2

Proof of Lemma A.2. Assumption A.2(ii) implies $\mathbf{H}_n \subset \mathbf{H}_{n+1} \subset \cdots \subset \mathbf{H}$. Thus, it is straightforward that $\widetilde{\Theta}_k^{\alpha}(\mathbf{P})$ and $\widetilde{\Theta}_k^{p}(\mathbf{P})$ cover $\Theta_k^{\alpha}(\mathbf{P})$ and $\Theta_k^{p}(\mathbf{P})$, respectively.

Define $h_k^*(y,t) = 0.5 \times \text{sign}(\Delta_k f_{(Y,T)|Z}(y,t))$ and $h_{k,n}^* = \arg\max_{h \in \mathbf{H}_n} \Delta_k \mathbb{E}[h(Y,T)|Z]$. Since (A19) holds for $\forall h \in \mathbf{H}$ and the equality holds if and only if h is such that for all $(y,t) \in \Omega_Y \times \{0,1\}$, $h(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \geq 0$, we know that

$$TV_{(Y,T),k} = \sup_{h \in \mathbf{H}} \Delta_k \mathbb{E}[h(Y,T)|Z]. \tag{A21}$$

In addition, because $H = h_k^*(Y,T)$, where the random variable H is defined in the proof of Lemma 2.4 and we have shown that $\Delta_k \mathbb{E}[H|Z] = TV_{(Y,T),k}$, it yields that

$$h_k^* = \arg\sup_{h \in \mathbf{H}} \Delta_k \mathbb{E}[h(Y,T)|Z].$$

Due that $f_{(Y,T)|Z=z_k}$ is Hölder continuous, for $l = 1, 2, ..., L_n$

$$\begin{aligned}
&\max_{(y,t) \in I_{n,l}} \Delta_k f_{(Y,T)|Z}(y,t) - \min_{(y',t') \in I_{n,l}} \Delta_k f_{(Y,T)|Z}(y',t') \\
\leq &\max_{(y,t) \in I_{n,l}} f_{(Y,T)|Z=z_k}(y,t) - \min_{(y,t) \in I_{n,l}} f_{(Y,T)|Z=z_{k-1}}(y,t) - \min_{(y',t') \in I_{n,l}} f_{(Y,T)|Z=z_k}(y',t') \\
&+ \max_{(y',t') \in I_{n,l}} f_{(Y,T)|Z=z_{k-1}}(y',t') \\
\leq &2M_0 \left(\frac{2M_1}{L_n}\right)^m.
\end{aligned} \tag{A22}$$

Denote $M_n = 2M_0(\frac{2M_1}{L_n})^m$. If $\max_{(y,t) \in I_{n,l}} |\Delta_k f_{(Y,T)|Z}(y,t)| > M_n$, from (A22) it has to be the maximum and minimum of $\Delta_k f_{(Y,T)|Z}(y,t)$ over $(y,t) \in I_{n,l}$ both stand on one side of zero. Thus, $\text{sign}(\Delta_k f_{(Y,T)|Z}(y,t))$ is a constant over $I_{n,l}$, and $h_k^* = h_{k,n}^*$ for those $I_{n,l}$. Therefore, for each $I_{n,l}$, we have either $h_k^* = h_{k,n}^*$ and $|\Delta_k f_{(Y,T)|Z}(y,t)| > M_n$, or $|\Delta_k f_{(Y,T)|Z}(y,t)| \leq M_n$. Now, consider the following three cases. Firstly, $h_k^* = h_{k,n}^*$ and $|\Delta_k f_{(Y,T)|Z}(y,t)| > M_n$. Then,

$$h_k^*(y,t)\Delta_k f_{(Y,T)|Z}(y,t) - h_{k,n}^*(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \leq M_n, \tag{A23}$$

since the left hand side of (A23) is zero and $M_n \geq 0$. Secondly, for (y,t) such that $h_k^*(y,t) =$

$h^*_{k,n}(y,t)$ and $|\Delta_k f_{(Y,T)|Z}(y,t)| \leq M_n$, (A23) still holds. Lastly, for (y,t) such that $h^*_k(y,t) = -h^*_{k,n}(y,t)$ and $|\Delta_k f_{(Y,T)|Z}(y,t)| \leq M_n$, we have

$$\begin{aligned}
&h^*_k(y,t)\Delta_k f_{(Y,T)|Z}(y,t) - h^*_{k,n}(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \\
=&2h^*_k(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \\
=&2h^*_k(y,t)sign(\Delta_k f_{(Y,T)|Z}(y,t))|\Delta_k f_{(Y,T)|Z}(y,t)| \\
=&2 \times 0.5|\Delta_k f_{(Y,T)|Z}(y,t)| \\
\leq& M_n,
\end{aligned}$$

where the third equality is because $h^*_k(y,t) = 0.5 \times sign(\Delta_k f_{(Y,T)|Z}(y,t))$, and the last inequality is due to $|\Delta_k f_{(Y,T)|Z}(y,t)| \leq M_n$. Therefore, (A23) holds for $\forall (y,t) \in \Omega_Y \times \{0,1\}$.

$$\begin{aligned}
0 \leq & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \sup_{h\in\mathbf{H}} \mathbb{E}[\varphi_k h(Y,T)] - \max_{h\in\mathbf{H}_n} \mathbb{E}[\varphi_k h(Y,T)] \right\} \\
= & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \mathbb{E}\left[\varphi_k h^*_k(Y,T)\right] - \mathbb{E}\left[\varphi_k h^*_{k,n}(Y,T)\right] \right\} \\
= & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \sum_{t=0,1} \int h^*_k(y,t)\Delta_k f_{(Y,T)|Z} d\mu_Y(y) - \sum_{t=0,1} \int h^*_{k,n}(y,t)\Delta_k f_{(Y,T)|Z} d\mu_Y(y) \right\} \\
= & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \sum_{l=1}^{L_n} \int_{I_{n,l}} \left[h^*_k(y,t) - h^*_{k,n}(y,t) \right] \Delta_k f_{(Y,T)|Z} d\mu_Y(y) d\mu_T(t) \right\} \\
\leq & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} M_n \to 0,
\end{aligned}$$

which gives the first convergence in Lemma A.2, since $M_n \to 0$ uniformly over $\mathbf{\Pi} \times \mathscr{P}_0$.

Moreover, recall that (A20) is satisfied by $\forall h_{k'} \in \mathbf{H}$, and the equality holds if and only if $h_{k'}$ is such that $\forall (y,t) \in \Omega_Y \times \{0,1\}$, $h_{k'}(y,t)\Delta_{k'} f_{(Y,T)|Z}(y,t) \geq 0$ for all $k' \neq k$. Thus,

$$1 - \sum_{k'\neq k} TV_{(Y,T),k'} = \inf_{\{h_{k'}\}\in\mathbf{H}^{K-1}} \left[1 - \sum_{k'\neq k} \Delta_{k'} \mathbb{E}[h_{k'}(Y,T)|Z] \right] = 1 - \sum_{k'\neq k} \sup_{h_{k'}\in\mathbf{H}} \Delta_{k'} \mathbb{E}[h_{k'}(Y,T)|Z]. \quad \text{(A24)}$$

Hence, it follows from (A23) and (A24) that

$$\begin{aligned}
0 \leq & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \min_{\{h_{k'}\}\in\mathbf{H}_n^{K-1}} \left[1 - \sum_{k'\neq k} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)] \right] - \inf_{\{h_{k'}\}\in\mathbf{H}^{K-1}} \left[1 - \sum_{k'\neq k} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)] \right] \right\} \\
= & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \left\{ \sum_{k'\neq k} \left[\sup_{h_{k'}\in\mathbf{H}} \Delta_{k'} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)] - \max_{h_{k'}\in\mathbf{H}_n} \mathbb{E}[\varphi_{k'} h_{k'}(Y,T)] \right] \right\} \\
\leq & \sup_{(\pi,\mathbf{P})\in\mathbf{\Pi}\times\in\mathscr{P}_0} \sum_{k'\neq k} M_n = (K-1)M_n \to 0.
\end{aligned}$$

□

A.3.3 Proof of Theorem A.1

By abuse of notation, denote by Θ the parameter space of θ_k, and denote by $\widetilde{\Theta}^{\theta}_k(\mathbf{P})$ the partially identified set of θ_k under $\mathbf{H}_n$. Before we proceed to the proof of Theorem A.1, let us introduce Corollary 5.1 and Theorem 6.1 in Chernozhukov et al. (2019) of the asymptotic size and power of

the test statistic $\tau(\theta_k, \pi_k, \pi_{k-1})$.

Theorem A.3. *By abuse of notation, we denote $\pi_k^0 = Pr(Z = z_k)$ to emphasis that it is the true probability. Given a sequence of $\varepsilon_n > 0$ with $\varepsilon_n \to 0$ and $\varepsilon_n\sqrt{\log(p_n)} \to \infty$, denote $\mathscr{H}_{0,n}$ as*

$$\mathscr{H}_{0,n} = \left\{(\theta_k, \pi, \mathbf{P}) \in \Theta \times \mathbf{\Pi} \times \mathscr{P}_0 : \ \theta_k \in \widetilde{\Theta}_k^{\theta}(\mathbf{P}),\ (\pi_{k-1}, \pi_k) = (\pi_{k-1}^0, \pi_k^0)\right\}.$$

Denote $\mathscr{H}_{1,n}$ as

$$\mathscr{H}_{1,n} = \left\{(\theta_k, \pi, \mathbf{P}) \in \Theta \times \mathbf{\Pi} \times \mathscr{P}_0 : \max_{j=1,\ldots,p_n} \frac{m_j(\theta_k, \pi_k, \pi_{k-1})}{\sigma_j(\theta_k, \pi_k, \pi_{k-1})} \geq (1+\varepsilon_n)\sqrt{\frac{2\log(p_n)}{n}},\ \textit{and}\right.$$
$$\left.(\pi_{k-1}, \pi_k) = (\pi_{k-1}^0, \pi_k^0)\right\}.$$

Under assumptions in Theorem A.1,

(i) $\liminf_{n\to\infty} \inf_{(\theta_k,\pi,\mathbf{P})\in\mathscr{H}_{0,n}} Pr\left[\tau(\theta_k, \pi_k, \pi_{k-1}) \leq c_k(\eta)\right] \geq 1-\eta.$

(ii) $\lim_{n\to\infty} \sup_{(\theta_k,\pi,\mathbf{P})\in\mathscr{H}_{1,n}} Pr\left[\tau(\theta_k, \pi_k, \pi_{k-1}) \leq c_k(\eta)\right] = 0.$

Now, the proof of Theorem A.1 can be shown as below.

Proof of Theorem A.1. Denote the event A as $A = \{\pi_k^0 \in \mathscr{C}_{\pi_k}(\eta_\pi), \pi_{k-1}^0 \in \mathscr{C}_{\pi_{k-1}}(\eta_\pi)\}$ and its complement as A^C.

(i) Under assumptions in Theorem A.1, for any $\mathbf{P} \in \mathscr{P}_0$ such that $\theta_k \in \widetilde{\Theta}_k^{\theta}(\mathbf{P})$, we can get

$$\begin{aligned}
&\Pr\left[\theta_k \notin \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)\right] \\
=&\Pr\left[\theta_k \notin \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi), A\right] + \Pr\left[\theta_k \notin \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi), A^C\right] \\
\leq&\Pr\left[\theta_k \notin \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi), A\right] + \Pr\left[A^C\right] \\
\leq&\Pr\left[\theta_k \notin \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi), A\right] + \Pr\left[\pi_k^0 \notin \mathscr{C}_{\pi_k}(\eta_\pi)\right] + \Pr\left[\pi_{k-1}^0 \notin \mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right] \\
\leq&\Pr\left[\tau(\theta_k, \pi_k^0, \pi_{k-1}^0) > c_k(\eta), A\right] + \Pr\left[\pi_k^0 \notin \mathscr{C}_{\pi_k}(\eta_\pi)\right] + \Pr\left[\pi_{k-1}^0 \notin \mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right] \\
\leq&\Pr\left[\tau(\theta_k, \pi_k^0, \pi_{k-1}^0) > c_k(\eta)\right] + \Pr\left[\pi_k^0 \notin \mathscr{C}_{\pi_k}(\eta_\pi)\right] + \Pr\left[\pi_{k-1}^0 \notin \mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right], \qquad \text{(A25)}
\end{aligned}$$

where the second last inequality is by definition of $\mathscr{C}_{\theta_k}$. Therefore, it follows from Theorem A.3-(i), Assumption A.1 and the convergence of $\widetilde{\Theta}_k^{\theta}(\mathbf{P})$ to $\Theta_k^{\theta}(\mathbf{P})$ in Lemma A.2, that

$$\liminf_{n\to\infty} \inf_{\mathbf{P}\in\mathscr{P}_0,\ \theta_k\in\Theta_k^{\theta}(\mathbf{P})} \Pr\left[\theta_k \in \mathscr{C}_{\theta_k}(\eta + 2\eta_\pi)\right] \geq 1 - (\eta + 2\eta_\pi).$$

(ii) Given Theorem A.3-(ii), for $\forall(\theta_k, \pi, \mathbf{P}) \in \Theta \times \mathbf{\Pi} \times \mathscr{P}_0$ such that $(\pi_{k-1}, \pi_k) = (\pi_{k-1}^0, \pi_k^0)$ and $\theta_k \notin \Theta_k^{\theta}(\mathbf{P})$, it suffices to show that the above $(\theta_k, \pi, \mathbf{P}) \in \mathscr{H}_{1,n}$, when n is sufficiently large. Since if so, Theorem A.3-(ii) leads to that for any fixed $\theta_k \notin \Theta_k^{\theta}(\mathbf{P})$, we have $\Pr\left[\tau(\theta_k, \pi_k^0, \pi_{k-1}^0) > c_k(\eta)\right]$ going to one. By Assumption A.1-(i), suppose there exists a constant M_2 such that $\sigma_j(\cdot) < M_2$ for all $j = 1, 2, ..., p_n$. Consider the two cases below.

Case 1. $\theta_k = \alpha_{k,k-1}$. If $\alpha_{k,k-1} \notin \Theta_k^{\alpha}(\mathbf{P})$, at least one of (A1)-(A3) is violated. If (A1) does not hold, then $\mathbb{E}[-\varphi_k sign(\alpha_{k,k-1})Y] > 0$ at $(\pi_{k-1}, \pi_k) = (\pi_{k-1}^0, \pi_k^0)$, which means its sample analogue $\hat{m}_1(\alpha_{k,k-1}, \pi_k^0, \pi_{k-1}^0) > 0$ for large enough n. By the definition of the test statistic $\tau(\alpha_{k,k-1}, \pi_k, \pi_{k-1})$

and the boundedness of $\sigma_j(\cdot)$, we have that

$$\tau(\alpha_{k,k-1}, \pi_k, \pi_{k-1}) = O_p(\sqrt{n}) \to \infty.$$

While, it yields from $\varepsilon_n \to 0$ and the assumption on the rate of p_n that $(1+\varepsilon_n)\sqrt{\frac{2\log(p_n)}{n}}$ goes to zero. Therefore, we know that $(\theta_k, \pi, \mathbf{P}) \in \mathscr{H}_{1,n}$ for large enough n.

If (A2) does not hold, it implies that

$$\sup_{h_k \in \mathbf{H}} \mathbb{E}\left\{\varphi_k\left[|\alpha_{k,k-1}| h_k(Y,T) - \text{sign}(\alpha_{k,k-1})Y\right]\right\} > 0.$$

Based on the first convergence result in Lemma A.2 and the fact that $\varphi_k|\alpha_{k,k-1}|$ is bounded by Assumption A.1, there exists some $h_k \in \mathbf{H}_n$ such that when n is large enough,

$$\mathbb{E}\left\{\varphi_k\left[|\alpha_{k,k-1}| h_k(Y,T) - \text{sign}(\alpha_{k,k-1})Y\right]\right\} > 0. \tag{A26}$$

Let $c > 0$ be the value of the left hand side of (A26). We can then conclude that there exists a $j = 2, ..., \kappa_n+1$ such that $m_j(\theta_k, \pi_k, \pi_{k-1}) \geq c$ when n is sufficiently large, leading to $\tau(\alpha_{k,k-1}, \pi_k, \pi_{k-1}) = O_p(\sqrt{n}) \to \infty$. Thus, $(\theta_k, \pi, \mathbf{P}) \in \mathscr{H}_{1,n}$ is satisfied.

If (A3) does not hold, it implies

$$\sup_{h_k \in \mathbf{H}} \mathbb{E}\left[\varphi_k \text{sign}(\alpha_{k,k-1})Y - |\alpha_{k,k-1}|\left(1 - \sum_{k' \neq k} \varphi_{k'} h_{k'}(Y,T)\right)\right] > 0.$$

The same arguments for (A26) can be applied to arrive the same conclusion, based on the second convergence in Lemma A.2 as well as the fact that $\varphi_{k'}|\alpha_{k,k-1}|$ is bounded. Hence, we can conclude that if $\alpha_{k,k-1} \notin \Theta_k^{\alpha}(\mathbf{P})$, then $(\alpha_{k,k-1}, \pi, \mathbf{P}) \in \mathscr{H}_{1,n}$. The desired result follows directly from Theorem A.3-(ii).

Case 2. $\theta_k = \Delta p_k$. Since $\Delta p_k \notin \Theta_k^{p}(\mathbf{P})$, at least one of Equations (A4)-(A6) is violated. The same arguments for **Case 1** can be applied to achieve the desired results. □