I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13430

Bounding Program Benefits When Participation Is Misreported

Denni Tommasi
Lina Zhang

JUNE 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13430

Bounding Program Benefits When Participation Is Misreported

Denni Tommasi
Monash University and IZA

Lina Zhang
Monash University

JUNE 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

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

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

www.iza.org

ABSTRACT

Bounding Program Benefits When Participation Is Misreported*

In empirical research, measuring correctly the benefits of welfare interventions is incredibly relevant for policymakers as well as academic researchers. Unfortunately, the endogenous program participation is often misreported in survey data and standard instrumental variable techniques are not sufficient to point identify and consistently estimate the effects of interest. In this paper, we focus on the weighted average of local average treatment effects (LATE) and (i) derive a simple relationship between the causal and the identifiable parameter that can be recovered from the observed data, (ii) provide an instrumental variable method to partially identify the heterogeneous treatment effects, (iii) formalize a strategy to combine administrative data on the misclassification probabilities of treated individuals to further tighten the bounds. Finally, we use our method to reassess the benefits of participating to the 401(k) pension plan on savings.

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

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

Corresponding author:
Denni Tommasi
Department of Econometrics and Business Statistics
Monash University
Building H, Caulfield Campus
900 Dandenong Road
Melbourne VIC 3145
Australia

E-mail: denni.tommasi@monash.edu

* We would like to thank Umair Khalil, Arthur Lewbel, Whitney Newey, Didier Nibbering, Mervyn Silvapulle and Tatsushi Oka for valuable comments. All errors are our own.

1 Introduction

There is increasing evidence that the endogenous participation to social programs is substantially misreported in survey data (Meyer et al., 2015). Since participation is binary, attempting to evaluate the benefits of a program using a standard instrumental variable method would lead to biased estimates.[1] As shown by Millimet (2011), this is a problem of primary importance because, even with infrequent arbitrary errors, the bias can be severe. In this paper, we develop an instrumental variable method to partially identify the heterogenous treatment effect when the endogenous treatment variable is misclassified. Our method requires minimal additional assumptions and can be applied to a wide range of empirical settings. Moreover, we formalize a strategy to combine administrative data on the misclassification probabilities of program receipts which can provide tight bounds.

We focus on the weighted average of local average treatment effects (LATE),[2] which is a parameter that can be estimated to measure the benefits of a program in case of non-compliance (Athey and Imbens, 2017). Here the true treatment is endogenous, the treatment effects are heterogeneous, and binary, discrete or multiple discrete instrument(s) are available. In our setting, the instrumental variable(s) satisfy some familiar conditions: (i) They are independent of the measurement error; (ii) They affect the outcome and the mismeasured treatment only through the true treatment status; and (iii) They affect the true treatment monotonically. Furthermore, we place no restriction on the marginal distributions of the measurement error, nor on the dependence between the measurement error with the potential outcomes and treatments. This means that we allow for a general form of (endogenous and heterogeneous) misclassification error in the recorded treatment status, such as endogenous misreporting and strategic answering, which may be due to stigma or the sensitivity of information collected.[3]

We start by showing the limitations of the standard LATE approach when the observed binary treatment is a mismeasured proxy of the true treatment. We derive a simple relationship between the causal and the identifiable parameter that can be recovered from the observed data, which can be captured by a summary statistic of the weighted average of the treatment misclassification probabilities. The relationship is useful because it can be used by researchers to hypothesize the bias of the benefits of a program for different values of the misclassification probabilities. Then, we show that one endogenous misclassified treatment variable already suffices the main partial identification result of the paper. Two strategies yield the bounds for the weighted average of LATEs: First, via the identified sets of LATEs; Second, via the identified sets of the local average of

[1] In a classical measurement error scenario, an instrumental variable is a standard method to correct for both endogeneity and measurement error of the treatment variable at the same time. However, in case of a binary treatment, measurement error is always nonclassical because the only way to misclassify a true treated is downwards, as a control, whereas the only way to misclassify a true control is upwards, as a treated. This creates a negative correlation between the true treatment status and the error term, which leads to severe bias.

[2] The LATE parameter of Imbens and Angrist (1994) is the average effect of the treatment for compliers when the scalar instrument is binary. The presence of discrete, or similarly, multiple, instruments, gives rise to a different parameter which is the weighted average of local average treatment effects. In our paper, we provide results valid for both parameters, hence, to avoid any confusion, throughout the text one should be careful to distinguish when we refer to LATE and when to the weighted average of LATEs.

[3] Indeed, according to Meyer et al. (2015), among the reasons for the misreporting of transfer benefits in household surveys there is: "[...] 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." (page 219). Our method can deal with all these scenarios.

treatment misclassifications (LATMs). We also provide sufficient conditions under which the bounds of LATEs and LATMs are sharp. We call this method P-LATE, for partially identified weighted average of LATEs.

We proceed by developing a strategy to improve the partial identification of the parameter. Our approach is based on the idea that administrative records of program receipts can provide accurate information about the extent of misreporting in survey data. We formalize a way to incorporate such information into the framework and show that, potentially, it can lead to tight bounds of program benefits. The idea is motivated by the observation that, in the literature, an increasing number of studies report misclassification probabilities for a wide range of programs. For example, using data from the Survey of Income and Program Participation (SIPP) merged with information from tax records, Dushi and Iams (2010) find that the participation rate in defined contribution (DC) pension plans is about 11% higher when using tax records rather than survey reports. In a similar vein, Meyer et al. (2018) use administrative data on Food Stamp Program (SNAP) participation and link them to the American Community Survey (ACS), the Current Population Survey (CPS), and SIPP. They find that 23% of true food stamp recipient households do not report receipt in the SIPP, 35% in the ACS, and 50% in the CPS. Misclassification probabilities of other US government transfer programs are reported in Meyer and Mittag (2019a,b). As more researchers gain access to linked administrative data, similar information can be obtained for other countries and be utilized to improve the partial identification results.

Regarding inference, we construct confidence intervals for the identified sets with uniformly and asymptotically size control. They are built based on a two-step bootstrap procedure following the seminal work by Chernozhukov et al. (2019). The confidence intervals of the weighted average of LATEs are computed depending on how its identified set is built. Specifically, it is either a union of the confidence intervals of the LATEs, or a union of the confidence intervals of the LATMs. Additional information about the accuracy of the measurement can also be taken into account. 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. Firstly, in order to further improve the identified set on the parameter of interest, we show the benefits of having multiple treatment indicators, or repeated measures of the same treatment, in case of discrete and multiple instrument(s). Importantly, we do not restrict the dependence among our treatments, thus the extra measures might be endogenous. Secondly, 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 show how to use the propensity score index to include covariates in the analysis. Furthermore, we use our method to reassess the benefits of participating to the 401(k) pension plan on savings.

Overall, our article shows that researchers measuring the benefits of a program can obtain bounds of the weighted average of LATEs if the binary treatment is misclassified. These bounds can potentially be tight, provided that information about the extent of misreporting in survey data can be found. In many applications, these information are readily available from studies of administrative records of program receipts. In other applications, one could also rely on small validation studies, repeated measurements of the same individual, as well as economic theory. We propose a method that is applicable as the leading identification strategy in any setting where the practitioner knows that the endogenous binary treatment is not well measured. Alternatively, one could also apply our method as the leading robustness check either (i) in case misreporting is only suspected, or (ii) to assess the sensitivity of program benefits under different assumptions of the misclassification probabilities. Although our method is primarily motivated by the program evaluation literature, it is not limited to applications within this context. Indeed, it can be applied to any setting where the endogenous binary treatment is misclassified by endogenous measurement error. The weighted average of LATEs could also be used to extrapolate to the average treatment effects, or other parameters of interest, which commonly require additional assumptions (Imbens, 2010).

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents our framework and the main results. Section 4 develops an inference procedure for the parameter of interest. Section 5 discusses extensions, how to use P-LATE in practice, simulations and an application. Concluding remarks are in Section 6. Proofs and additional material are in the Appendix.

2 Related Literature

Our paper is primarily motivated by the extensive literature documenting misclassification error in the observed treatment (Bollinger, 1996; Angrist and Krueger, 1999; Kane et al., 1999; Card, 2001; Hernandez et al., 2007). Particularly Bound et al. (2001) and Black et al. (2003) argue that measurement error is likely to be endogenous in some applications, such as educational attainment. Recent works by Meyer et al. (2015, 2018) and Meyer and Mittag (2019a,b) document extensive and increasing endogenous measurement error also in the participation to social programs. Importantly, Kreider (2010) shows how severe the identification problem is in a binary regressor model given the presence of even infrequent arbitrary errors. Similarly, Millimet (2011) studies the performance of several estimators, commonly used in the treatment effects literature, in the presence of measurement error in the binary treatment indicator, and emphasizes the importance of not ignoring the measurement error in this case.

Using an instrumental variable (IV) is a standard approach to correct for both endogeneity and measurement error of the treatment variable at the same time.[4] In the context of endogenous treatment and endogenous measurement error, Nguimkeu et al. (2018) provide point identification of homogeneous treatment effects under the availability of two instrumental variables and strong

[4] In the context of an exogenous treatment subject to exogenous misclassification, many authors, such as Aigner (1973), Kane et al. (1999), Black et al. (2000) and Frazis and Loewenstein (2003), use instrumental variables techniques to estimate homogeneous (constant) treatment effects of a mismeasured binary regressor. More recently, Mahajan (2006), Lewbel (2007) and Hu (2008), also use instruments to point identify average treatment effects, without assuming that treatment effects are homogeneous, in the case of an exogenous and mismeasured binary (or discrete) treatment indicator. The partial identification treatment effects literature has also been active. Indeed, under more general conditions, bounds on average treatment effects with misclassified treatment are provided by Klepper (1988), Bollinger (1996), Kreider and Pepper (2007), Molinari (2008) and Imai and Yamamoto (2010).

parametric and distributional assumptions. Differently from this paper, we focus on nonparametric heterogeneous treatment effects and provide a partial identification method that does not require such strong assumptions.

We aim to estimate the weighted average of local average treatment effects (LATE) of Imbens and Angrist (1994).[5] For this reason, our paper is primarily related to Ura (2018), who investigates the identifying power of a binary instrumental variable and provides the conditions to set identify the effect of an endogenous and misclassified treatment variable in a heterogeneous treatment effect framework.[6]

With respect to the latter, our contribution is to generalize his results in three main directions. First, we derive a simple relationship between the weighted average of LATEs and the identifiable parameter that can be recovered from the observed data. Similar formal relationships in an IV context are provided by Frazis and Loewenstein (2003), Lewbel (2007), Battistin and Sianesi (2011), Calvi, Lewbel, and Tommasi (2018) and Stephens Jr. and Unayama (2020). Second, we generalize his method to discrete and multiple instrumental variable settings by using all information contained in the IVs. In particular, we gain identification power by making use of multiple total variation distances, which capture the distributional effect of the instrument(s) on observable variables, and form a tighter bound for the proportion of the compliers. We find that when the instrumental variable(s) are discrete, there are two sources of information that can be extracted from the observable data to conduct partial identification for the weighted average of LATEs. One is directly used to build the sets for LATEs. The other one contains two key factors: the relationship between the causal and the identifiable parameter that we discovered is this paper, and the identified sets of LATMs. Finally, similarly to Lewbel (2007), Kreider and Pepper (2007), Molinari (2008), Battistin and Sianesi (2011), Kreider et al. (2012) and Battistin et al. (2014), we formalize an approach to potentially further tighten the bounds which is based on the use of external information about misclassification probabilities.

3 Theoretical framework

This section proceeds in four acts. First, we describe our theoretical framework and show the limitations of the standard LATE approach when the treatment variable is contaminated by measurement error. This leads to a simple relationship between the true and mismeasured parameter, which can be captured by a summary statistic of the weighted average of the misclassification probabilities. Second, we provide the conditions to obtain the identified sets of the LATEs and of the misclassification probabilities. We do not restrict the dependence between the misclassification and the potential outcomes, nor between the misclassification and the potential treatments. This means

[5] See Huber and Wuthrich (2018) for a recent review of methodological advancements in the evaluation of heterogeneous treatment effect models based on instrumental variable (IV) methods.

[6] Whereas Calvi, Lewbel, and Tommasi (2018) address the problem of exogenous misclassification of the true treatment status in a setting with a binary treatment and a binary instrument. For a recent application of their estimator, see Tommasi (2019). For a recent extension, see Hoagland (2019). Other related papers, albeit not necessarily focused on LATE, are Battistin et al. (2014), Chalak (2017), DiTraglia and García-Jimeno (2018), Jiang and Ding (2019), Kasahara and Shimotsu (2019), Kedagni (2019) and Yanagi (2019).

that we can allow for endogenous and heterogeneous misclassification error, such as endogenous misreporting and strategic answering. Third, we show that one endogenous binary misclassified treatment indicator already suffices the main partial identification result. Moreover, we provide two main strategies based on different sources of information to partially identify the parameter of interest. Fourth, we show that the resulting bounds can be tightened by making use of external information regarding the extent of the misclassification probabilities. This constitutes our third main partial identification strategy.

3.1 Set up and limitations of standard LATE approach

We introduce some notations 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.

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, each of which is unconfounded (e.g., randomized), correlated with D, and satisfies the standard Imbens and Angrist (1994) assumptions of instruments for LATE estimation. Let $\Omega_Z = \{z_0, z_1, ..., z_K\}$ be the support of Z with $z_k \in \mathbb{R}^h$. Suppose that the random binary variables D_k, for $k = 0, 1, ..., K$, are potential treatments corresponding to possible realizations 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 $Pr(z_k) = \mathbb{E}(D = 1|Z = z_k)$ the propensity score. Let Y be an observed outcome of interest and let random variables Y_1 and Y_0 be the potential outcomes Y_d with $d \in \{0, 1\}$ for possible realizations d of D. Denote by $\mathbf{Y} \subset \mathbb{R}$ 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 $g : \Omega_Z \mapsto \mathbb{R}$. In a standard instrumental variables approach, $g(z)$ can be an estimate of $\Pr(z)$ or other known functions.[7]

Assumption 3.1. *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; $0 < \pi_k = Pr(Z = z_k) < 1$, $k = 0, 1, ..., K$;

[7] If Z is a single binary instrument, $g(z) = z$ is the special case considered by Ura (2018). Moreover, note that if Z consists of a single discrete instrument, we can simply set $g(z)$ to be an identity function. 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.

(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 ensures no defiers. Throughout the paper, we denote compliers $(D_{k-1} = 0, D_k = 1)$ as C_k. If D was observed, under the conditions listed in Assumption 3.1, the Imbens and Angrist (1994)'s weighted average of local average treatment effect (LATE) would be identified by the instrumental 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 γ_k^{IV} are the weights and $\alpha_{k,k-1}$ is the local average treatment effect $\mathbb{E}[Y_1 - Y_0 | C_k]$ 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 we do not observe D, we cannot implement this standard approach.

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 error. Define random variables T_0 and T_1 as potential observed treatments so that T_d with $d \in \{0, 1\}$ is for possible realizations d 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 classification errors: if $T_0 = 1$, then a true $D = 0$ is misclassified as treated, and if $T_1 = 0$, then a true $D = 1$ is misclassified as untreated.

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

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

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

Assumption 3.2-(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 would be sufficient to make 3.2-(i) hold.[8] Assumption 3.2-(ii) is a minimal relevance condition saying that T, although suffers from potential misclassification error, still provides some information regarding D. Hence T is correlated with $g(Z)$.

[8] Notice that the definitions used for Y and T assume implicitly that, once the true treatment D is controlled for, there is no direct effect of Z on Y, nor of Z on T. That is, Z satisfies the individual-level exclusion restriction. See Swanson et al. (2018) for a recent and comprehensive review of the various versions of exclusion restriction.

Using the proxy T in place of D leads to the identification of a new parameter, which is useful to characterize. Let $p_{d,k} = E(T_d \mid C_k)$ 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 3.1. *Let Assumption 3.1 and 3.2 hold for T. 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 γ_k^{Mis} are the weights for each subgroup of compliers C_k.

Proof of Theorem 3.1. See Appendix A.1.1. □

Intuitively, α^{Mis} denotes the parameter that we estimate if we ignore the misclassification error and use a mismeasured treatment indicator T in place of the true treatment D. 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 $k = 1,2,...,K$ (no misclassification error). There is a simple relationship between α^{IV} and α^{Mis} which can be captured by a summary statistic of the weighted average of the misclassification probabilities. This relationship will become useful later on in conducting partial identification of $\alpha^I V$.

Corollary 3.1. *Let Assumption 3.1 and 3.2 hold for T 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:*

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

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 3.1. See Appendix A.1.2. □

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. and unobserved in practice. Corollary 3.1 demonstrates that, in case of misclassification of the binary treatment variable, the estimated weighted average of LATEs exceeds α^{IV} by a factor $1/\xi$.

A final remark is in order. α^{Mis} generalizes the B-LATE (for Biased LATE) estimator of Calvi, Lewbel, and Tommasi (2018) to a multiple instruments setting. The latter paper is among the first to show that, in the standard LATE framework, if T is misclassified, the estimated LATE exceeds the true LATE parameter by a factor $1/p$, where p is the fraction of individuals correctly reporting their treatment status. Our factor $1/\xi$ becomes $1/p$ when a scalar binary instrument is used. Similar calculations in an IV context are provided by Frazis and Loewenstein (2003), Lewbel (2007), Battistin and Sianesi (2011) and Stephens Jr. and Unayama (2020). Note that all these papers are related to one another and benefited from the result by Hausman et al. (1998). First, similarly to Frazis and Loewenstein (2003) and Stephens Jr. and Unayama (2020), but differently from Lewbel (2007) and Battistin and Sianesi (2011), we assume 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. Finally, differently from Frazis and Loewenstein (2003), we assume a monotone IV. This allows us to derive the misclassification probabilities in terms of the compilers.

3.2 Identified sets of the LATEs and of the misclassification probabilities

We introduce an additional assumption needed for partial identification of α^{IV} (besides Assumptions 3.1 and 3.2).

Assumption 3.3. *(Ascending order) 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 $Pr(z_l) \leq Pr(z_w)$, and this order is known.*

Assumption 3.3 says that, even though the propensity scores cannot be recovered from the observed data (because D is unobserved in practice), the ascending order of them in Z is still known. This can be seen as a structural restriction imposed on the true treatment D to recover the sign of $\alpha_{k,k-1}$. Therefore, the ascending order is itself informative.[9] As noticed by Abadie et al. (2002), an example of sufficient condition for Assumption 3.3 is a constant-coefficient latent-index model. That is, suppose the treatment is generated by $D = 1(\gamma Z > \eta)$, where γ is a parameter and η is an error term independent of Z. Then, the order of $\Pr(z_k)$, which is required in Assumption 3.3, is determined by the sign of γ. It is plausible in many applications that the sign of γ can be retrieved from economic theory.[10] For example, in the study of the returns to schooling, distance to college is often used as an instrument for completed college education (e.g. Card (2001) among others). In this specific example, the parameter γ is negative.

Given Assumption 3.3, we use the concept of total variation (TV) distance introduced by Ura (2018), and adapt it to our general framework. 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:

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

[9] Indeed, if the ascending order of Ω_Z is known, even without an observable treatment D, we can still identify the sign of $\alpha_{k,k-1}$, for those k satisfying (i) $\Pr(z_k) \neq \Pr(z_{k-1})$ and (ii) $\mathbb{E}(Y|Z = z_k)$ and $\mathbb{E}(Y|Z = z_{k-1})$ are finite. The fact that the sign of $\alpha_{k,k-1}$ cannot be identified if $\Pr(z_k) = \Pr(z_{k-1})$ is immaterial, because in this case we would have $\gamma_k^{IV} = 0$, indicating that $\alpha_{k,k-1}$ contributes nothing to and will not affect the estimand α^{IV}. When $\Pr(z_k) = \Pr(z_{k-1})$, we also have $\gamma_k^{Mis} = 0$ and therefore $\alpha_{k,k-1}$ does not affect α^{Mis} as well.

[10] A similar structural restriction on treatment D is employed by Chalak (2017), when recovering treatment effects with instrument suffering from measurement error. A sufficient condition is a threshold crossing in D with a monotonic latent function. See his Assumption 4 and the discussion below it.

where μ_A denotes a dominating measure for the distribution of A.[11] 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 $TV_{Y,k}$ is the distribution version of the "intent-to-treat" effect. The TV will play a crucial role in characterizing the identification set of the probability of compliers.[12]

Lemma 3.1. *Let Assumption 3.1-(ii) to (iv), 3.2-(i) and 3.3 hold for T. 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 3.1. See Appendix A.1.4. □

Lemma 3.1 provides an intuitive and simple identified set for the probability of the compliers when the actual treatment D is unobservable.[13] The width of the bound in Lemma 3.1 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.

Given Lemma 3.1, we can now proceed to consider the identified sets for each LATE, $\alpha_{k,k-1}$, and for 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):

$$LATM = \Delta p_k = \mathbb{E}(T_1 - T_0 | C_k)$$

because the conditional expectation is analogous to the LATE if we replace $Y_1 - Y_0$ by $T_1 - T_0$. Let $\mathbf{P}$ be an arbitrary data generating process of (Y, T, Z). Denote the class of data generating processes 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 $\Theta = \mathbb{R}$ if outcome Y is continuous with infinite support.[14] For notational simplicity, we denote $\Delta_k \mathbb{E}(Y|Z) = \mathbb{E}(Y|Z = z_k) - \mathbb{E}(Y|Z = z_{k-1})$. Theorem 1 in Imbens and Angrist (1994) says that under Assumption 3.1 in this paper, we have:

$$\Delta_k \mathbb{E}(Y|Z) = \alpha_{k,k-1} P(C_k). \tag{4}$$

Multiplying both sides of (4) by $\alpha_{k,k-1}$, we obtain that:

$$\alpha_{k,k-1} \Delta_k \mathbb{E}(Y|Z) = \alpha_{k,k-1}^2 Pr(C_k) \geq 0. \tag{5}$$

[11] 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).

[12] Similar identification strategies have been introduced in the partial identification literature. See e.g. the integrated envelope in Kitagawa (2009).

[13] Throughout the paper, the "identified set" is referred to as a set which includes the collection of possible values of the parameter of interest, and those values are all compatible with the data and the assumptions. The identified set here may not be sharp. Chesher (2010) uses "outer region" to refer the "identified set" in this paper. Throughout the paper, we also refer to "bound" as the two extreme values of a identified set when the set is a interval.

[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 3.1 to the absolute value of (4), we have:

$$|\Delta_k \mathbb{E}(Y|Z)| \leq |\alpha_{k,k-1}| \left[1 - \sum_{k' \neq k} TV_{(Y,T),k'}\right], \tag{6}$$

$$|\Delta_k \mathbb{E}(Y|Z)| \geq |\alpha_{k,k-1}| TV_{(Y,T),k}. \tag{7}$$

Thus, under Assumptions 3.1, 3.2 and 3.3, each LATE $\alpha_{k,k-1}$ satisfies the inequalities (5)-(7). Inequality (5) 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 (6) and (7) give the lower and upper bounds of $|\alpha_{k,k-1}|$, respectively. Denote the set of $\alpha_{k,k-1}$, characterized by (5)-(7), as $\Theta_k^\alpha(\mathbf{P}) \subset \Theta$.

Similar to equation (4), we also have $\Delta_k \mathbb{E}(T|Z) = \Delta p_k \Pr(C_k)$, and similar arguments can be applied to obtain the inequalities (8)-(10) below, satisfied by each Δp_k:

$$\Delta p_k \Delta_k \mathbb{E}(T|Z) \geq 0, \tag{8}$$

$$|\Delta_k \mathbb{E}(T|Z)| \leq |\Delta p_k| \left[1 - \sum_{k' \neq k} TV_{(Y,T),k'}\right], \tag{9}$$

$$|\Delta_k \mathbb{E}(T|Z)| \geq |\Delta p_k| TV_{(Y,T),k}. \tag{10}$$

Denote the set of Δp_k, characterized by (8)-(10), as $\Theta_k^p(\mathbf{P}) \subset [-1,1]$. In the next Lemma, we derive the explicit expression for $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^p(\mathbf{P})$ under different conditions. We also give sufficient conditions under which the identified sets are sharp.

Lemma 3.2. *Let Assumption 3.1-(ii)-(iv), 3.2-(i) and 3.3 hold for T. 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:

$$\Theta_k^\alpha(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T),k}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) < 0; \end{cases} \tag{11}$$

(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 (11) is the sharp identified set of $\alpha_{k,k-1}$.

Proof of Lemma 3.2. See Appendix A.1.5. □

Lemma 3.2 shows that, if $TV_{(Y,T),k} = 0$, then no useful information can be extracted from the observable data, so that $\Theta_k^\alpha(\mathbf{P})$ excludes no values from the parameter space of $\alpha_{k,k-1}$, i.e. $\Theta_k^\alpha(\mathbf{P}) = \Theta$. It also indicates that the instrument variation from z_{k-1} to z_k has no identification power. Once $TV_{(Y,T),k} > 0$, the instrument has nontrivial identification power, and an explicit expression of the identified set can be derived for $\alpha_{k,k-1}$. 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 $\Delta_k\mathbb{E}(Y|Z)$. In addition, Lemma 3.2 provides the sufficient condition for sharp identified set of each LATE in case of multiple or multi-valued instrumental variable(s).[15]

Similarly, the Lemma below gives the identified set of Δp_k, as well as the sufficient conditions for the sharpness of the identified set.

Lemma 3.3. *Let Assumption 3.1-(ii)-(iv), 3.2-(i) and 3.3 hold for T. 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:

$$\Theta_k^p(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k\mathbb{E}(T|Z)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}}, \frac{\Delta_k\mathbb{E}(T|Z)}{TV_{(Y,T),k}}\right], & \text{if } \Delta_k\mathbb{E}(T|Z) > 0, \\ \{0\}, & \text{if } \Delta_k\mathbb{E}(T|Z) = 0, \\ \left[\frac{\Delta_k\mathbb{E}(T|Z)}{TV_{(Y,T),k}}, \frac{\Delta_k\mathbb{E}(T|Z)}{1-\sum_{k'\neq k} TV_{(Y,T),k'}}\right], & \text{if } \Delta_k\mathbb{E}(T|Z) < 0; \end{cases} \tag{12}$$

(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 (12) is the sharp identified set of Δp_k.

Proof of Lemma 3.3. See Appendix A.1.6. □

We require 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 3.3, the sign and an informative bound for Δp_k can be obtained as long as $TV_{(Y,T),k} > 0$. It is also clear that, in order to construct the identified set of Δp_k, we do not need any prior or external information about how severely the treatment proxy T is contaminated by measurement error. The identified sets of the LATEs, $\{\alpha_{k,k-1}\}_{k=1}^K$, and of the LATMs, $\{\Delta p_k\}_{k=1}^K$, provide the fundamental basis for constructing the identified set of the estimand α^{IV}.

Two final remarks are in order. First, Lemma 3.1 generalizes Lemma 3 of Ura (2018) to accommodate multiple or multi-valued instrument(s). Notice that, with a binary instrument, this author proposes to use only the subpopulation where the instrument takes two values, and construct the identified set as $TV_{(Y,T),k} \leq \Pr(C_k) \leq 1$. However, Lemma 3.1 demonstrates the possible identification power gain of such a strategy, as we can actually bound $\Pr(C_k)$ from above by $1 - \sum_{k'\neq k} TV_{(Y,T),k'}$ instead of 1. This improvement is due to the fact that all groups of compliers $\{C_k\}_{k=1}^K$ are mutually exclusive, and that all $\Pr(C_{k'})$ with $k' \neq k$ can be bounded from below by their corresponding total variation distances. Second, compared to the subpopulation strategy proposed by Ura (2018), Lemma 3.2 improves one side of the bound of $\alpha_{k,k-1}$, from $\Delta_k\mathbb{E}(Y|Z)$ to $\Delta_k\mathbb{E}(Y|Z)/(1-\sum_{k'\neq k} TV_{(Y,T),k'})$. Such an improvement is substantial, especially when there exists at least one $k' \neq k$ such that $TV_{(Y,T),k'} > 0$.

[15] Two further points about Lemma 3.2 are worth noticing. First, if there is only one k such that $TV_{(Y,T),k} > 0$, while $TV_{(Y,T),k'} = 0$ for all $k' \neq k$, then $\Theta_k^\alpha(\mathbf{P})$ in Lemma 3.2-(ii) is sharp, and it is identical to the identified set suggested by Ura (2018) using only a subpopulation. The latter result is intuitive, because $TV_{(Y,T),k'} = 0$ for all $k' \neq k$ implies that C_k is the only compliers group that induces nonzero changes in the potential outcomes or the potential treatment indicators. Thus, only the subpopulation that includes the compliers C_k matters. Second, for more general cases, where more than two total variation distances are nonzero, our identified set $\Theta_k^\alpha(\mathbf{P})$ is going to be more informative about each LATE $\alpha_{k,k-1}$ than the identified set of this author.

3.3 Partial Identification of α^{IV}

We begin by proposing two main 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 $\{\lambda_k^{IV}\}_{k=1}^K$ summing up to one. Hence the first partial identification strategy is based on the identified sets of $\{\alpha_{k,k-1}\}_{k=1}^K$:

$$\min_{k=1,2,\ldots,K}\{\alpha_{k,k-1}\} \leq \alpha^{IV} = \sum_{i=1}^{K} \lambda_k^{IV}\alpha_{k,k-1} \leq \max_{k=1,2,\ldots,K}\{\alpha_{k,k-1}\} \tag{13}$$

Denote our first identified set of α^{IV} as $\Theta^{\alpha}(\mathbf{P})$, where the superscript α means that it is constructed from $\{\Theta_k^{\alpha}(\mathbf{P})\}_{k=1}^K$. Then, $\Theta^{\alpha}(\mathbf{P})$ can be obtained from (13) and the identified sets of LATEs given in Lemma 3.2.

Theorem 3.2. *Let Assumptions 3.1, 3.2, and 3.3 hold for T. Then, $\Theta^{\alpha}(\mathbf{P}) = \bigcup_{k\in\{1,2,\ldots,K\}}\Theta_k^{\alpha}(\mathbf{P})$.*

Proof of Theorem 3.2. See Appendix A.1.7. □

Theorem 3.2 shows that the identified set of α^{IV} is the union of the 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 weighted average of local average treatment effects in situations where at least two LATEs, $\alpha_{k,k-1}$ and $\alpha_{k',k'-1}$, have opposite signs. Fortunately, however, we are still able to recover the sign of α^{IV} as long as all the LATEs stand on the same side of zero. Furthermore, we can recover the sign of all the LATEs from the observed data and Lemma 3.2.[16] 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 secures (on average) higher wages. Hence, in this case, the "direction consistency" of LATEs is positive.[17]

Corollary 3.2. *Let Assumption 3.1, 3.2, and 3.3 hold for T.*

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

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

[16] As discussed in Appendix A.1, we actually allow the sign of $\alpha_{k,k-1}$ for the "ineffectual subgroups", representing compliers subgroups making no contributions to the objective of interest, to be unknown.

[17] Although Corollary 3.2 does not require any sign restriction to ensure direction consistency of the LATEs, imposing them might be useful for inference. It is useful especially when some compliers with small, or close to zero, probability (by Lemma 3.1) with estimated LATEs have opposite sign as compared to other complier groups. This inconsistency of direction may be caused by small samples, and, removing those small probability events via sign restrictions, will improve the identified set substantially. Such sign restriction, or direction consistency as we call it, is commonly assumed in the treatment effects partial identification literature, usually referred to as "monotonicity" assumptions. 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 3.2. The proof follows from Lemma A.2 in Appendix and Theorem 3.2. □

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 some $\Delta_k \mathbb{E}(Y|Z) = 0$, the results above still hold with possibility $\alpha^{IV} = 0$, as long as their corresponding $TV_{(Y,T),k} > 0$.[18]

Second strategy. Our second strategy is built upon the relation between α^{IV} and α^{Mis}, and the identified sets of $\{\Delta p_k\}_{k=1}^{K}$. Recall from Corollary 3.1 that $\alpha^{IV} = \xi \alpha^{Mis}$, where $\xi = \gamma_k^{IV}/\gamma_k^{Mis} = \sum_{k=1}^{K} \gamma_k^{IV} \Delta p_k$. Based on the definition of ξ, we have:

$$\min_{k=1,2,...,K} \{\Delta p_k\} \leq \xi = \sum_{k=1}^{K} \gamma_k^{IV} \Delta p_k \leq \max_{k=1,2,...,K} \{\Delta p_k\}. \tag{14}$$

Denote our second identified set of α^{IV} as $\Theta^{p}(\mathbf{P})$, where the superscript p represents its key components $\{\Theta_k^{p}(\mathbf{P})\}_{k=1}^{K}$. $\Theta^{p}(\mathbf{P})$ can be characterized by the Theorem below.

Theorem 3.3. *Let Assumption 3.1, 3.2, and 3.3 hold for T. Then,*

$$\Theta^{p}(\mathbf{P}) = \left\{ \alpha^{Mis} \times \Delta p : \ \Delta p \in \bigcup_{k=1,2,...,K} \Theta_k^{p}(\mathbf{P}) \right\}, \tag{15}$$

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

Proof of Theorem 3.3. See Appendix A.1.8. □

Theorem 3.3 gives the general form of the identified set $\Theta^{p}(\mathbf{P})$, based on both the identifiable estimand α^{Mis} and the identified sets of $\{\Delta p_k\}_{k=1}^{K}$. If $\alpha^{Mis} = 0$, α^{IV} is point identified as zero. Again, there are situations where $\Theta^{p}(\mathbf{P})$ may fail to recover the sign of α^{IV}, when at least one set $\Theta_k^{p}(\mathbf{P})$ includes both positive and negative elements. However, we argue that, in most empirical applications, Δp_k should be at least nonnegative. This is quite intuitive because it indicates that, for compliers, the probability of having their treatment status correctly observed, if they were treated, is larger than the probability of being wrongly observed as treated, if they were untreated (false positive). For example, in the study of the effects of SNAP in the US, up to 50% of true participants does not report to be treated and only less than 5% of true non-participants reports to be treated (Meyer et al., 2018). Hence, in this case, the “direction consistency” of the difference in misclassification probabilities is positive ($0.50 - 0.05 = 0.45 > 0$). Such a feature is often mild to assume

[18] As long as $TV_{(Y,T),k} > 0$, the identified set of $\alpha_{k,k-1}$ has an explicit expression as in Lemma 3.2(ii), and it is not Θ. If $TV_{(Y,T),k} = 0$ for at least one k, then $\Theta^{\alpha}(\mathbf{P}) = \Theta$.

and identifiable by Lemma 3.3.[19]

Corollary 3.3. *Let Assumption 3.1, 3.2, and 3.3 hold for T. 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

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

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

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

Proof of Corollary 3.3. It follows directly from Lemma A.2 in Appendix and Theorem 3.3. □

The Corollary above gives the sign of α^{IV}, and the explicit expression of $\Theta^p(\mathbf{P})$, when the direction consistency of $\{\Delta p_k\}_{k=1}^K$ holds.

First vs Second strategy. The two strategies introduced thus far are both compatible with the observable data under Assumptions 3.1, 3.2 and 3.3. Moreover, they make distinct contributions to the partial identification of α^{IV} because they are based on different sources of information. Since the two identified sets 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 $\Theta_k^p(\mathbf{P})$.

Corollary 3.4. *Let Assumption 3.1, 3.2 and 3.3 hold for T. $\Theta^\alpha(\mathbf{P})$ and $\Theta^p(\mathbf{P})$ can be rewritten as follows:*

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

where $\alpha^{IV} = \sum_{k=1}^K \gamma_k^{IV} \alpha_{k,k-1}$ and $\xi = \sum_{k=1}^K \gamma_k^{IV} \Delta p_k$ are weighted average of LATEs and weighted average of the LATMs, respectively, and Δp is any generic value in $\Theta_k^p(\mathbf{P})$.

Proof of Corollary 3.4. See Appendix A.1.9. □

Corollary 3.4 delivers four crucial messages. First, in general, unless more information are available, it is not a-priori obvious which identified set outperforms the other, since $\alpha_{k,k-1}/\Delta p_k$ may

[19] Our second strategy does not require any sign restrictions on $\{\Delta p_k\}_{k=1}^K$. However, for finite sample inference, such restrictions may be helpful to rule out the possibility that Δp_k have the opposite sign for some compliers, with small or close to zero proportion, with respect to other compliers.

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 $\Delta p \in \Theta_k^p(\mathbf{P})$ and all k, the closer to 1 is the ratio $\Delta p/\xi$, the narrower is the identified set delivered by strategy 2 (that is, the narrower is the set $\Theta^p(\mathbf{P})$ of α^{IV}). Fourth, at the limit, if, for all k, $\Delta p_k = \xi$, that is, the data satisfy homogeneous misclassification, then $\Theta_k^p(\mathbf{P}) = \xi$. In this last case, point identification is achieved by $\Theta^p(\mathbf{P})$ as follows:

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

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 $\xi \Delta_k \mathbb{E}(Y|Z)/\Delta_k \mathbb{E}(T|Z)$,[20] from Corollary 3.4 we have:

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

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 are in order. First, following the method of intersecting the bounds, which is commonly applied in the treatment effect partial identification literature,[21] there is no issue preventing us from intersecting $\Theta^{\alpha}(\mathbf{P})$ and $\Theta^p(\mathbf{P})$ to achieve even tighter bound. However, when considering inference, adopting only one identification set (instead of both) 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, Theorems 3.2 and 3.3 will be identical. In addition, our first two identified sets will also coincide with that in Ura (2018), because $K = 1$ and $\sum_{k' \neq k} TV_{(Y,T),k'}$ will degenerate to zero.

3.4 Partial Identification of α^{IV} using external information

The increase availability of administrative records of program receipt inspires our third main partial identification strategy. Accurate information about the extent of misreporting in survey data can offer a potential strategy to reduce the bounds of program benefits. These information can be obtained from external sources of information, such as: administrative data, validation studies, repeated measurements of the same individual, or from economic theory. Suppose the practitioner has some prior or external information about the possible range of ξ, and this range is narrower

[20] 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.

[21] See, for example, Manski (1990), Heckman and Vytlacil (1999), Manski and Pepper (2000), Chesher (2010).

than that in Equation (14). Then we can replace $\Theta_k^p(\mathbf{P})$ in $\Theta^p(\mathbf{P})$ by this known range to further tighten the bounds.

Denote our third identified set of α^{IV} as $\Theta^{\xi}(\mathbf{P})$, where the superscript ξ indicates extra information about measurement accuracy. At the risk of repetition, recall from Corollary 3.1 that $\alpha^{IV} = \xi\alpha^{Mis}$, where $\xi = \gamma_k^{IV}/\gamma_k^{Mis} = \sum_{k=1}^{K}\gamma_k^{IV}\Delta p_k$.

Theorem 3.4. *Let Assumption 3.1 and 3.2 hold for T. 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 $\alpha^{Mis} \geq 0$, then $\alpha^{IV} \geq 0$ and $\Theta^{\xi}(\mathbf{P}) = \left[\underline{\xi}\alpha^{Mis}, \overline{\xi}\alpha^{Mis}\right]$.

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

Proof of Theorem 3.4. See Appendix A.1.10. □

Intuitively, the constants $\underline{\xi}$ and $\overline{\xi}$ are two bounds of the weighted average of LATMs. A similar approach of using external information is adopted, in various contexts, by Lewbel (2007), Kreider and Pepper (2007), Molinari (2008), Battistin and Sianesi (2011), Kreider et al. (2012) and Battistin et al. (2014). By using these extra information, the identified set $\Theta^{\xi}(\mathbf{P})$ will be at least as good as that in Corollary 3.3 (second strategy). If no extra information about the measurement accuracy is available, one could (in principle) simply set $\underline{\xi}$ and $\overline{\xi}$ as the lower and upper bounds 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.

From Theorem 3.4, 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 ($p_{1,k} = 1$ and $p_{0,k} = 0$), 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 P-LATE, combined with external information, can be used as a bias reduction method with respect to a naïve IV estimator.

4 Inference

In this section, we construct the confidence intervals of the partially identified α^{IV}. Recall that, given the data generating process $\mathbf{P}$ of (Y,T,Z), $\Theta_k^{\alpha}(\mathbf{P})$ denotes the identified set of $\alpha_{k,k-1}$, and $\Theta_k^p(\mathbf{P})$ denotes the identified set of Δp_k. Since, in practice, the partial identification of α^{IV} is based on some 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 of the identified sets $\alpha_{k,k-1}$ and Δp_k. Second, we construct the confidence intervals for both $\alpha_{k,k-1}$ 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.

4.1 Moment inequalities of the identified sets

One feasible estimation of the identified sets is via the bootstrap-based testing of moment inequalities method proposed by Chernozhukov et al. (2019). We follow Ura (2018) and extends his results to accommodate multiple or multi-valued instrumental variables. 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 4.1. *Let Assumption 3.1-(ii)-(iv), 3.2-(i) and 3.3 hold for T. 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:

$$\mathbb{E}\left[-\varphi_k sign(\alpha_{k,k-1})Y\right] \leq 0, \tag{16}$$

$$\mathbb{E}\left\{\varphi_k\left[|\alpha_{k,k-1}|h(Y,T) - sign(\alpha_{k,k-1})Y\right]\right\} \leq 0, \ \forall h \in \mathbf{H} \tag{17}$$

$$\mathbb{E}\left[\varphi_k sign(\alpha_{k,k-1})Y - |\alpha_{k,k-1}|\left(1 - \sum_{k' \neq k}\varphi_{k'}h_{k'}(Y,T)\right)\right] \leq 0, \ \forall h_{k'} \in \mathbf{H}. \tag{18}$$

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

$$\mathbb{E}\left[-\varphi_k sign(\Delta p_k)T\right] \leq 0, \tag{19}$$

$$\mathbb{E}\left\{\varphi_k\left[|\Delta p_k|h(Y,T) - sign(\Delta p_k)T\right]\right\} \leq 0, \ \forall h \in \mathbf{H} \tag{20}$$

$$\mathbb{E}\left[\varphi_k sign(\Delta p_k)T - |\Delta p_k|\left(1 - \sum_{k' \neq k}\varphi_{k'}h_{k'}(Y,T)\right)\right] \leq 0, \ \forall h_{k'} \in \mathbf{H}, \tag{21}$$

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 4.1. See Appendix A.1.11.[22] □

$\Delta_k\mathbb{E}[h(Y,T)|Z]$ with $h \in \mathbf{H}$ can bound the total variation distance:

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

[22] 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, in this paper we may not distinguish h and $h_{k'}$ elsewhere if it is not necessary, and we use h to denote any generic function in $\mathbf{H}$.

and φ_k helps rewrite the conditional moments to unconditional ones:

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

for $Q \in \{Y, T, h(Y, T)\}$.

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 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 4.1. *The parameter space $\Theta \times \mathbf{\Pi} \times \mathscr{P}_0$ satisfies the following conditions:*[23]

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

(ii) All random variables inside $\mathbb{E}[\cdot]$ in Lemma 4.1 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 4.1 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 4.1. 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 4.2.[24] 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 4.1 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. We refer to $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$ as the "approximated identified sets", and their convergence will be formally defined in Assumption 4.2 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 that describe the approximated identified sets. Then, $p_n = 1 + \kappa_n + \kappa_n^{K-1}$. Assumption 4.2 below outlines the sufficient assumptions on the DGP and the partition $\{I_{n,l}\}_{l=1}^{L_n}$ that ensure the convergence of the approximated identified sets $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$ to $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^p(\mathbf{P})$.

[23] Confidence intervals of π_k and α^{Mis} are needed since they are regarded as nuisance parameters in the inference process and will be estimated in advance.

[24] When $\mathbf{H}$ has finite dimension, we can simply let $\mathbf{H}_n = \mathbf{H}$.

Assumption 4.2. *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 $I_{n+1,1}, I_{n+1,2}, \dots, I_{n+1,L_{n+1}}$ is a refinement of the partition $I_{n,1}, I_{n,2}, \dots, I_{n,L_n}$.

(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 4.2-(i) restricts the smoothness of the density function of observable (Y,T) and 4.2-(ii) implies the sequence $\{\mathbf{H}_n\}$ satisfying $\mathbf{H}_n \subset \mathbf{H}_{n+1} \subset \cdots \subset \mathbf{H}$. Assumption 4.2-(iii) is used to make sure that the partition becomes finer as sample size increases. Assumption 4.2-(iv) is borrowed from Chernozhukov et al. (2019) for the asymptotic performance of the confidence interval.

If $\mathbf{H}_n = \{h_{n,1}, h_{n,2}, \dots, h_{n,\kappa_n}\}$ based on partition $\{I_{n,l}\}_{l=1}^{L_n}$ satisfies Assumption 4.2, we have the convergence of $\widetilde{\Theta}_k^\alpha(\mathbf{P})$ and $\widetilde{\Theta}_k^p(\mathbf{P})$ to $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^p(\mathbf{P})$, as formally stated by the following Lemma.

Lemma 4.2. *Let Assumption 3.1, 3.2, 3.3, 4.1 and 4.2 hold for T. 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 $(\pi, \mathbf{P}) \in \mathbf{\Pi} \times \mathscr{P}_0$.*

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

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

Proof of Lemma 4.2. See Appendix A.1.12. □

Given the convergence result in Lemma 4.2, we can now proceed to the inference stage of the approximated identified sets.

4.2 Confidence intervals of the approximated identified sets

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 $\eta_\pi \in (0, \eta/2)$, suppose $(1-\eta_\pi)$-confidence intervals of π_{k-1} and π_k, $\mathscr{C}_{\pi_{k-1}}(\eta_\pi)$ and $\mathscr{C}_{\pi_k}(\eta_\pi)$, are available to the practitioner. We can then construct a $(1-\eta-2\eta_\pi)$-

confidence interval of θ_k:

$$\mathscr{C}_{\theta_k}(\eta + 2\eta_\pi) = \bigcup_{\pi_k \in \mathscr{C}_{\pi_k}(\eta_\pi),\ \pi_{k-1} \in \mathscr{C}_{\pi_{k-1}}(\eta_\pi)} \Big\{\theta_k \in \Theta:\ \tau(\theta_k, \pi_k, \pi_{k-1}) \le c_k(\eta)\Big\}, \tag{22}$$

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.2. The testing procedure is for the p_n moment inequalities which characterize the approximated identified sets.[25] To obtain Equation (22), π_{k-1} and π_k are regarded as nuisance parameters and estimated in advance. The following Theorem holds for both $\theta_k = \alpha_{k,k-1}$ and $\theta_k = \Delta p_k$.

Theorem 4.1. *Let Assumption 3.1, 3.2, 3.3, 4.1 and 4.2 hold for T. 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.2.*

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

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

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

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

Proof of Theorem 4.1. See Appendix A.1.13. □

Theorem 4.1-(i) shows that the confidence interval $\mathscr{C}_{\theta_k}(\eta+2\eta_\pi)$ defined in Equation (22) covers any point in the identified set with probability at least $(1-\eta-2\eta_\pi)$ uniformly over $\mathscr{P}_0$. In addition, Theorem 4.1-(ii) tells us that the confidence interval, evaluated at the true π_{k-1}, π_k, will exclude any fixed point outside the identified set 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 the nuisance parameter π and its associated confidence interval are used to construct Equation (22).

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:

$$\hat{\mathscr{C}}_{\theta_k}(\eta) = \Big\{\theta_k \in \Theta:\ \tau(\theta_k, \hat{\pi}_k, \hat{\pi}_{k-1}) \le \hat{c}_k(\eta)\Big\}, \tag{23}$$

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

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

in Appendix B.2 of Chernozhukov et al. (2019).[26] Moreover, the simpler version confidence interval $\hat{\mathscr{C}}_{\theta_k}(\eta)$ in Equation (23) can also be applied to construct $\mathscr{C}^{\alpha}(\beta^{\alpha})$ or $\mathscr{C}^{p}(\beta^{p})$ for practical purpose.

4.3 Confidence intervals of α^{IV}

Given the results in Theorems 3.2-3.4 and given the confidence intervals of $\alpha_{k,k-1}$ and Δp_k, we can now move on to construct a confidence interval of α^{IV}. Firstly, we propose a $(1-\beta^{\alpha})$-confidence interval $\mathscr{C}^{\alpha}(\beta^{\alpha})$, according to the first partial identification strategy in Theorem 3.2:

$$\mathscr{C}^{\alpha}(\beta^{\alpha}) = \bigcup_{k=1,2,..,K} \mathscr{C}_{\alpha_{k,k-1}}(\eta + 2\eta_{\pi}), \tag{24}$$

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

In addition, another $(1-\beta^{p})$-confidence interval of α^{IV}, denoted by $\mathscr{C}^{p}(\beta^{p})$, is based on the second partial identification strategy in Theorem 3.3:

$$\mathscr{C}^{p}(\beta^{p}) = \bigcup_{\alpha \in \mathscr{C}_{\alpha Mis}\left(\eta_{\alpha Mis}\right)} \left\{ \alpha \times \Delta p: \ \Delta p \in \bigcup_{k=1,2,..,K} \mathscr{C}_{\Delta p_k}(\eta + 2\eta_{\pi}) \right\}, \tag{25}$$

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

The last confidence interval, denoted by $\mathscr{C}^{\xi}(\beta^{\xi})$, comes from our partial identification strategy with external sources of information in Theorem 3.4:

$$\mathscr{C}^{\xi}(\beta^{\xi}) = \bigcup_{\alpha \in \mathscr{C}_{\alpha Mis}\left(\eta_{\alpha Mis}\right)} \left\{ \alpha \times \Delta p: \ \Delta p \in \left[\underline{\xi}, \overline{\xi}\right] \right\}, \tag{26}$$

where $\beta^{\xi} = \eta_{\alpha^{Mis}}$ and $\underline{\xi}, \overline{\xi}$ are given 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 4.1. *Let the assumptions in Theorem 4.1 hold for T. 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 (24)-(26) all control their sizes asymptotically and uniformly over $\mathscr{P}_0$, i.e.*

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

Proof of Corollary 4.1. See Appendix A.1.14. □

Corollary 4.1 proves that, for all the three confidence intervals of α^{IV}, their asymptotic coverage

[26] Although our moment inequalities in Lemma 4.1 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 Appendix A.7 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.

rate, at any point inside the associated identified set, achieves the desired level.[27] 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 $\eta_{\alpha^{Mis}} \leq \beta^{\alpha}$.

The results thus far can be summarized as follows. The partial identification of α^{IV} can be achieved by following one of three strategies. Strategy 1 depends only on the bounds of LATEs. Strategy 2 uses the estimand α^{Mis} together with the bounds of LATMs $\{\Delta p_k\}_{k=1}^{K}$. Strategy 3 exploits external sources of information to restrict the possible range of $\{\Delta p_k\}_{k=1}^{K}$ and further improve the bounds of the second strategy.

5 Extensions and Applications

This section is organized in four parts. First, we sketch two extensions of P-LATE which are fully developed in the Appendix. Second, we show how to implement P-LATE in practice. Third, we sketch the main ideas and results behind our Monte Carlo simulations which are fully presented in the Appendix. Finally, we apply our method to measure the benefits of participating to the 401(k) pension plan on savings.

5.1 Extensions

In Appendix A.4 and A.5, we present two main extensions that are briefly discussed here.

Multiple treatments or repeated measurements. The results in Section 3.3 and 3.4 require only one binary treatment indicator T. Nevertheless, if there are multiple treatment proxies (or repeated measurements), we can (potentially) further tighten the bounds of α^{IV}, since each proxy may carry different and relevant information about the actual (and unobservable) treatment D.[28] Based on the results presented in Appendix A.4, when multiple treatment proxies (or repeated measurements) are available, all three confidence intervals can be obtained in the same manner as in Equations (24)-(26). Moreover, there are two main differences between our approach and that commonly used in the literature when multiple treatments (or repeated measurements) are available to the practitioner.[29] 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 (and binary) treatment dummies from the same discrete treatment variable and capturing various pieces of useful information in the same measurement.

[27] More details about how we construct the confidence intervals of the there identified sets $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are given in Appendix A.3.

[28] 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.

[29] 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).

Including covariates. In many applications, the instruments may be confounded without conditioning on some covariates. In addition, treatment effects may be heterogeneous across the population characterized by different attributes. Hence, in the identification of causal effects, particular attention has been paid to accounting for covariates (see e.g. Abadie (2003), Frölich (2007) and Angrist and Fernandez-Val (2010) among many others). Following this literature, in Appendix A.5 we define our target parameter the conditional IV estimand $\alpha^{IV}(x)$, which can be expressed as a weighted average of the conditional LATEs $\{\alpha_{k,k-1}(x)\}_{K=1}^{k}$:

$$\alpha^{IV}(x) = \frac{\text{Cov}(Y, g(Z)|X=x)}{\text{Cov}(D, g(Z)|X=x)} = \sum_{k=1}^{K} \gamma_k^{IV}(x)\alpha_{k,k-1}(x), \tag{27}$$

where X is a vector of observables with support Ω_X, and $\gamma_k^{IV}(x)$ is the weight such that $\sum_{k=1}^{K}\gamma_k^{IV}(x) = 1$ for $\forall x \in \Omega_X$. In this case, we show that the extension of the three main partial identification strategies to accommodate for covariates is straightforward. This is because their respective identified sets for $\alpha^{IV}(x)$ are the conditional-on-covariates version of the identified sets obtained in Section 3.3, once Assumptions 3.1, 3.2 and 3.3 hold conditional on X. In Appendix A.6, we provide guidance on the inference procedure which, following Dehejia and Wahba (1999) and Battistin and Sianesi (2011), is based on the idea of stratification matching.

5.2 How to use P-LATE in practice

We provide some guidance about: (i) How to choose among the three partial identification strategies; and (ii) How to incorporate external information about the misclassification error and calculate the bounds using strategy 3.

First, the choice of which strategy to adopt depends on the information available. If no prior or external information about measurement accuracy is available, then both strategy 1 and 2 can be applied with the available dataset. Moreover, based on the discussion of Corollary 3.4, 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 pointed out at the end of Section 3.3, if the available instrument is binary, there is no choice to make between strategy 1 and 2 because they are exactly the same and they coincide with the strategy by Ura (2018). 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 is likely to help improving the bounds of α^{IV}. 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% confidence interval is $[0.52, 2.48]$. The objective of strategy 3 is to combine these information with information about the misclassification error in order to obtain the tightest bounds of α^{IV}. 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, $\xi = 1 - w^n - w^p$. In order to show how to implement at best the third strategy, we consider four examples and illustrate how one should set the range $[\underline{\xi}, \overline{\xi}]$ in each scenario. We also calculate the corresponding confidence intervals of α^{IV} using Equation (26).

Example 1. The first example mimics a context where, for each group of compliers, C_k, the number of false positive is lower than the number of false negative: that is, $0 \leq p_{0,k} \leq 1 - p_{1,k}$ (implying $0 \leq w^p \leq w^n$). This is a common situation under poor recalling of treatment status. Moreover, suppose only w^n is known. 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 $\xi \in [1 - 2w^n, 1 - w^n]$. Assume $w^n = 0.50$. Given the aforementioned estimate of α^{Mis}, the 95% confidence interval for α^{IV} in this case would be $[0, 1.49]$.[30]

Example 2. In the second example, consider again a context where, for each group of compliers, C_k, $0 \leq p_{0,k} \leq 1 - p_{1,k}$. However, differently from before, suppose only w^p is known. Then, $\xi \leq 1 - 2w^p$ because $w^p \leq w^n$, and $\xi \geq -w^p$ because $1 - w^n$ is nonnegative. In this case, a practitioner can set $\xi \in [-w^p, 1 - 2w^p]$. Suppose $w^p = 0.05$. Given the estimate of α^{Mis}, the 95% confidence interval for α^{IV} would be $[-0.12, 2.23]$.[31]

Four remarks follow directly from these first two examples. Firstly, if the value of false negative w^n is larger than 0.5 (that is, more than 50% of individuals who are truly treated report to be untreated), the range of ξ, in Example 1, would lie in an interval including both negative and positive values and would make Θ^{ξ} fail to identify the sign of α^{IV}. This would occur even if the confidence interval of α^{Mis} stood on one side of zero. This is intuitive because it suggests that the data collected are heavily contaminated by misclassification error. In such a situation, we warn the practitioner to be cautious about the interpretation of the results obtained. Secondly, if the only available information is the value of false positive, in general such information is likely to be too weak to recover, confidently, the sign of α^{IV}. This is because, in Example 2, when the false positive $w^p \leq 0.5$, ξ would also lie in an interval including both negative and positive values and would make, again, Θ^{ξ} fail to identify the sign of α^{IV}. In such a situation, we recommend to impose further restrictions, such as the probability of false negative to be at most 0.5, leading to a narrower bound $\xi \in [0.5 - w^p, 1 - 2w^p]$. The latter choice should be motivated by the specific context. Thirdly,

[30] The confidence interval (CI) is calculated using Equation (26). The ending points correspond to the smallest and largest points in 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 $\overline{\xi}$, 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 $\alpha^{IV} \in [0.52 \times 0, 2.48 \times 0.5] = [0, 1.49]$. The rule is slightly more complicated to apply if the confidence interval of the 2SLS estimate 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% confidence interval of this 2SLS estimate is $[-0.08, 3.08]$, which is less precise than before because the range is wider and contains both positive and negative values of the effect. In this case, if we apply the same rule, the CI would be calculate as follows: $\alpha^{IV} \in [-0.08 \times 0.5, 3.08 \times 0.5] = [-0.04, 1.54]$.

[31] The complication in this case is caused by the negative $\underline{\xi}$. Given the estimate of α^{Mis}, the CI of α^{IV} is $[2.48 \times (-0.05), 2.48 \times 0.9] = [-0.12, 2.23]$. Whereas, if $\widetilde{\alpha}^{Mis} = 1.5$, but the 95% confidence interval of this 2SLS estimate is $[-0.08, 3.08]$, the CI would be $\alpha^{IV} \in [(-0.08) \times 0.9, 3.08 \times 0.9] = [-0.07, 2.8]$.

both remarks imply that the information about false negative is in general more powerful than the information about false positive. This is also intuitive because, knowing the probability of false negative is equivalent to knowing the probability of true treated correctly reporting their treatment status (which is equal to $1 - w^n$). It can be thought as a measure of the quality of the sample collected. Finally, even if $\underline{\xi}$ and $\overline{\xi}$ are both positive, it is possible that $\Theta^{\xi}(\mathbf{P})$ may fail to identify the sign of α^{IV} if the confidence interval of α^{Mis} is on both side of zero.

Example 3. The third example mimics a context where the practitioner does not know the approximate number of false positive or false negative, but only an upper bound of w^n and w^p: For example, $w^n \leq 0.5$ and $w^p \leq 0.05$. In this case, the range of ξ can be set as $[0.45, 1]$, where the lower bound is computed by 1 minus the summation of the two maximum values of w^n and w^p, and the upper bound by 1 minus the summation of the two minimum values, which are 0. One special case is when the practitioner knows that the false positive is approximately 0; in this case $\xi \in [0.5, 1]$. Let us take the range $\xi \in [0.45, 1]$ as prior information. Given the estimate of α^{Mis}, the 95% confidence interval for α^{IV} would be $[0.23, 2.48]$.[32]

Example 4. In the last example, 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, the identified set $\Theta^{\xi}(\mathbf{P})$ 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% confidence interval of $[0.23, 1.12]$.[33] However, it is worth pointing out that, in practice, exactly because the value of w^n and w^p are likely to be only approximations of their true values, the point estimate obtained in this case will be biased with respect to the true α^{IV}. Nevertheless, our P-LATE estimator can still be used in place of a conventional IV estimator as a bias reduction method. Moreover, our simulation results in Appendix A.7 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}.

5.3 Monte Carlo Simulations

In Appendix A.7, 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 4. 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 Equation (23). Based on this, the confidence intervals of α^{IV} are constructed in the same manners as in Equations (24), (25) and (26). We explore extensively 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 P-LATE represents a reliable alternative estimator when practitioners can only use a mismeasured

[32] Whereas, if $\tilde{\alpha}^{Mis} = 1.5$, but the 95% confidence interval of this 2SLS estimate is $[-0.08, 3.08]$, the CI of α^{IV} would be again $[-0.08, 3.08]$.

[33] Whereas, if $\tilde{\alpha}^{Mis} = 1.5$, but the 95% confidence interval of this 2SLS estimate is $[-0.08, 3.08]$, the CI of α^{IV} would be $[-0.036, 1.39]$.

binary treatment T in place of D to estimate the benefits of a program. Moreover, P-LATE becomes very powerful, and works at best, when external information about the accuracy of the measurement error can be taken into account.

5.4 Application to the 401(k) pension plan

In this Section, we use our method to measure the benefits of participating to the 401(k) pension plan on savings and compare the results with the existing literature. This is one of the most popular defined contribution retirement plan, which is aimed at increasing financial savings through tax deducibility of the contributions to retirement accounts. The effects of this plan have been studied by, among others, Poterba et al. (1995), Abadie (2003) and Ura (2018). One of the main characteristics of the plan is that it is provided by employers, hence only workers in firms offering such program are eligible.

There are two main difficulties in measuring its benefits: endogenous participation to the plan and misreporting of participation. The first problem may arise due to unobserved differences in saving behaviors. That is, participants to this plan may save more than those who do not participate, even in the absence of the 401(k). Hence, a comparison of accumulated financial assets between participants and non-participants is likely to yield a positive bias of the true effect of the program. Whereas, the second problem may arise because individuals find it difficult to remember or understand their pension plan, leading to the issue of reporting error. Indeed, Gustman et al. (2007) document that about one-fourth of respondents to the Health and Retirement Study (HRS) survey misreport their pension plan. Furthermore, Dushi and Iams (2010) document that, in the Survey of Income and Program Participation (SIPP) survey, over 17% of participants to the plan self-report as non-participants (false negative) and almost 10% of non-participants self-report as participants (false positive). Understanding the benefits of such programs is relevant for the economic well-being of future retirees because these plans are an important part of retirement income security.

We use data from the SIPP survey round of 1991. The construction of the dataset follows Abadie (2003). Hence, our sample includes only households where at least one person is employed and has no income from self-employment. Moreover, the sample is restricted to individuals with annual family income between \$10,000 to \$200,000, because eligibility to the plan is rare outside of this interval. Table 1 reports the summary statistics of the main variables used in the analysis. The average family net financial assets (which is the outcome Y) is around \$19,000, roughly 27% of the observations report to participate in the 401(k) pension plan (which is the misreported treatment T), whereas 39% are eligible to the plan (which is the 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.

Table 2 reports the empirical results. Column (1) reports the conventional 2SLS estimates as shown in column (3) of Table 2 by Abadie (2003). This represents our benchmark result of a biased

Table 1: 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.

Table 2: Empirical results

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Existing results		P-LATE				
Abadie	Ura	Strategy	Strategy 3			
(2003)	(2018)	1 & 2	w^n only	w^p only	appr. w^n and w^p	exact w^n and w^p
9.4 (5.3, 13.5)	(4.4, 28.3)	(4.4, 28.3)	(3.3, 11.2)	(-1.4, 10.9)	(1.8, 13.6)	6.8 (3.8, 9.8)

Notes: Results in this Table are in 1,000$ units. Column (1) reports the conventional 2SLS estimates as shown in column (3) of Table 2 by Abadie (2003). Column (2) reports the best 95% confidence interval of the local average treatment effect as shown in Table 2 by Ura (2018). Column (3)-(6) report the 95% confidence interval of P-LATE under different assumptions regarding the misclassification probabilities. Finally, column (7) delivers a point estimate of the effect.

point estimate. The effect is statistically significant and says that participating to the 401(k) plan increases the total financial assets by roughly $9,400, with a 95% confidence interval lying between $5-13,000. Column (2) reports the best 95% confidence interval of the local average treatment effect as shown in Table 2 by Ura (2018). This is our benchmark result for P-LATE. As one can notice, although the interval accounts for the misclassification error of the treatment variable, the upper bound is more than twice as large as the upper bound of the previous interval. Column (3)-(6) report the 95% confidence interval of P-LATE under different assumptions regarding the misclassification probabilities. Column (3) assumes no information about the misclassification probabilities. Since the instrumental variable is binary, strategy 1 and 2 coincide with one another and are equivalent to the method developed by Ura (2018). Furthermore, column (4) assumes that only the probability of false negative w^n is known; column (5) assumes that only the probability of false positive w^p is known; column (6) assumes that only an upper bound of both these probabilities is known. As one can see, the confidence intervals delivered by P-LATE are similar, and in some cases even better, than those in column (1). Finally, column (7) assumes that both probabilities are known, hence in this case P-LATE delivers a point estimate of the effect.

6 Conclusion

In the evaluation of treatment effects, the 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 weighted average of local average treatment effects (LATE), which is a parameter that can be estimated to measure the effects of a treatment in case of non-compliance. 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 causal and the identifiable parameter that can be recovered from the observed data. Then, we provide strategies to set identify the weighted average of LATEs and to further tighten the bounds using external information on the misclassification probabilities.

Overall, our article shows that researchers who aim to measure treatment effects with a misclassified binary treatment can obtain bounds of the weighted average of LATEs. These bounds can potentially be tight, provided accurate information about the extent of misreporting in survey data can be found. These information can come from the increase availability of administrative records of treated individuals. There are applications where such information are readily available. In other applications, one could also rely on small validation studies, repeated measurements of the same individual, as well as economic theory. Our main conclusion is that our 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 instrument(s) are available.

References

ABADIE, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of econometrics*, 113, 231–263. [24], [27], [28], [46], [71]

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117. [9]

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

ANGRIST, J. AND I. FERNANDEZ-VAL (2010): "Extrapolate-ing: External validity and overidentification in the late framework," Tech. rep., National Bureau of Economic Research. [24], [69]

ANGRIST, J. D. (2001): "Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice," *Journal of business & economic statistics*, 19, 2–28. [70]

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 Experimentsa," in *Handbook of Field Experiments*, ed. by A. V. Banerjee and E. Duflo, North-Holland, vol. 1 of *Handbook of Economic Field Experiments*, 73 – 140. [2]

BATTISTIN, E., M. D. NADAI, AND B. SIANESI (2014): "Misreported schooling, multiple measures and returns to educational qualifications," *Journal of Econometrics*, 181, 136 – 150. [5], [17]

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. [5], [8], [9], [17], [24], [55], [69]

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. [13]

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]

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]

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

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]

CALVI, R., A. LEWBEL, AND D. TOMMASI (2018): "Women's Empowerment and Family Health: Estimating LATE with Mismeasured Treatment," . [5], [8]

CARD, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. [4], [9] CHALAK, K. (2017): "Instrumental variables methods with heterogeneity and mismeasured instruments," *Econometric Theory*, 33, 69–104. [5], [9]

CHEN, X., C. A. FLORES, AND A. FLORES-LAGUNES (2018): "Going beyond LATE Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*, 53, 1050–1099. [13]

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2019): "Inference on Causal and Structural Parameters using Many Moment Inequalities," *The Review of Economic Studies, forthcoming*. [3], [18], [20], [21], [22], [34], [49], [53]

CHESHER, A. (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78, 575–601. [10], [16]

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. [24], [69]

DITRAGLIA, F. J. AND C. GARCÍA-JIMENO (2018): "Mis-classified, Binary, Endogenous Regressors: Identification and Inference," Tech. rep., National Bureau of Economic Research. [5]

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]

FRAZIS, H. AND M. A. LOEWENSTEIN (2003): "Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables," *Journal of Econometrics*, 117, 151 – 178. [4], [5], [8], [9]

FRÖLICH, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35–75. [24], [67]

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. [9]

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. [23]

HECKMAN, J. J. AND R. ROBB (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, 30, 239–267. [70]

HECKMAN, J. J. AND E. J. VYTLACIL (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the national Academy of Sciences*, 96, 4730–4734. [16]

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]

HOAGLAND, A. (2019): "Who Do Innovations Reach? The Influence of Trainings on Mental Health Treatments," Tech. rep., Working Paper. [5]

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]

Huber, M. and K. Wuthrich (2018): "Local average and quantile treatment effects under endogeneity: a review," Tech. rep., Working paper. [5]

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. [4]

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. [4]

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [2], [5], [6], [7], [10], [35], [36]

Jiang, Z. and P. Ding (2019): "Measurement errors in the binary instrumental variable model," Tech. rep., Working paper. [5]

Kane, T. J., C. E. Rouse, and D. Staiger (1999): "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper 7235. [4]

Kasahara, H. and K. Shimotsu (2019): "Identification of Regression Models with a Misclassified and Endogenous Binary Regressor," Tech. rep., Working paper. [5]

Kedagni, D. (2019): "Identifying treatment effects in the presence of confounded types," Tech. rep., Working paper. [5]

Kitagawa, T. (2009): "Identification region of the potential outcome distributions under instrument independence," Tech. rep., CEMMAP working paper. [10]

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

Kreider, B. (2010): "Regression Coefficient Identification Decay in The Presence of Infrequent Classification Errors," *The Review of Economics and Statistics*, 92, 1017–1023. [4]

Kreider, B., C. Gundersen, and D. Jolliffe (2012): "Identifying the effects of food stamps on child health outcomes when participation is endogenous and misreported," . [5], [17]

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. [4], [5], [17]

Lewbel, A. (2007): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75, 537–551. [4], [5], [8], [9], [17]

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

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. [16]

——— (1997): "Monotone treatment response," *Econometrica: Journal of the Econometric Society*, 1311–1334. [13]

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. [13], [16]

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12, S200–S216. [13]

Meyer, B. D. and N. Mittag (2019a): "Misreporting of Government Transfers: How Important are Survey Design and Geography?" *Southern Economic Journal*. [3], [4]

——— (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. [3], [4]

Meyer, B. D., N. Mittag, and R. M. Goerge (2018): "Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation," Working Paper 25143, National Bureau of Economic Research. [3], [4], [14]

Meyer, B. D., W. K. C. Mok, and J. X. Sullivan (2015): "Household Surveys in Crisis," *Journal of Economic Perspectives*, 29, 199–226. [2], [4]

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], [4]

Molinari, F. (2008): "Partial identification of probability distributions with misclassified data," *Journal of Econometrics*, 144, 81 – 117. [4], [5], [17]

Nguimkeu, P., A. Denteh, and R. Tchernis (2018): "On the estimation of treatment effects with endogenous misreporting," *Journal of Econometrics*. [4]

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. [27]

Stephens Jr., M. and T. Unayama (2020): "Estimating the Impacts of Program Bene
ts: Using Instrumental Variables with Underreported and Imputed Data," *The Review of Economics and Statistics*, n/a. [5], [9]

Swanson, S. A., M. A. Hernán, M. Miller, J. M. Robins, and T. S. Richardson (2018): "Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes," *Journal of the American Statistical Association*, 113, 933–947. [7]

Tommasi, D. (2019): "Control of resources, bargaining power and the demand of food: Evidence from PROGRESA," *Journal of Economic Behavior Organization*, 161, 265 – 286. [5]

Ura, T. (2018): "Heterogeneous treatment effects with mismeasured endogenous treatment," *Quantitative Economics*, 9, 1335–1370. [5], [6], [9], [12], [16], [18], [24], [27], [28], [37], [38], [49], [55], [69]

Yanagi, T. (2019): "Inference on local average treatment effects for misclassified treatment," *Econometric Reviews*, 38, 938–960. [5]

A Appendix

This Appendix contains seven sections with proofs, additional material and analysis. The information are organized as follows. Appendix A.1 contains the proofs of Section 3. Appendix A.2 provides the details about the two-step multiplier bootstrap method introduced by Chernozhukov et al. (2019) and about how to use it to construct the confidence intervals for P-LATE. Appendix A.3 provides the details about how, in practice, confidence intervals are constructed in our paper. Appendix A.4 provides the details about the partial identification results using multiple treatment proxies. Appendix A.5 and A.6 provide the details about how to use P-LATE in empirical applications with covariates. Appendix A.7 contains the Monte Carlo simulations, tables and figures.

A.1 Proofs of Section 3

A.1.1 Proof of Theorem 3.1

Proof of Theorem 3.1. Assumption 3.2-(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),

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

For any $z_l, z_w \in \Omega_Z$, by the definition of T and Assumption 3.2-(i) (extended unconfoundedness),

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

Due to Assumption 3.1-(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

$D_{k-1} \leq D_k$[34]. Therefore,

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

Plug (A2) into (A1), we get

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

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

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

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

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

□

A.1.2 Proof of Corollary 3.1

Proof of Corollary 3.1. For $\forall k$, by definitions of γ_k^{IV} and γ_k^{Mis} we have:

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

□

A.1.3 Effectual and ineffectual subgroups

Let us first introduce some notation and definitions. For any generic random variable A and B, denote f_A the distribution function of A, and denote $f_{A|B}$ the conditional distribution function of A given B. If A is a discrete random variable, then $f_A(a)$ represents the probability of taking value a. Furthermore, we define an *effectual subgroup* the subgroup of compliers (C_k) with nonzero γ_k^{IV} and nonzero $\alpha_{k,k-1}$ such that this subgroup has nonzero contribution to α^{IV}. On the contrary, we define an *ineffectual subgroup* the subgroup of compliers (C_k) with either zero weight γ_k^{IV}, or zero LATE $\alpha_{k,k-1}$. That is, an ineffectual subgroup is a group of compliers that makes no contribution to α^{IV}.

From equation (4), it is easy to see that if $\mathbb{E}(Y|Z=z_k) - \mathbb{E}(Y|Z=z_{k-1}) = 0$, then the subgroup k is one of the ineffectual subgroups. In addition, a subgroup k such that $\sum_{l=k}^{K}(g(z_l) - \mathbb{E}[g(Z)])\pi_l = 0$

[34] In discrete IV setting, $P(z_{k-1}) \leq P(z_k)$ implies $D_{k-1} \leq D_k$ can be simply proved by contradiction under Assumption 3.1-(ii) that $Z \perp (Y_1, Y_0, D_k)$ for $k = 0, 1, ..., K$.

is also an ineffectual subgroup. An effectual subgroup k', instead, is such that $\mathbb{E}(Y|Z=z_{k'})-\mathbb{E}(Y|Z=z_{k'-1})\neq 0$ and $\sum_{l=k'}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l\neq 0$.

Lemma A.1. *Suppose Assumptions 3.1-(ii)-(iv), 3.2-(i) and 3.3 hold. Denote S^e as the collection of indicators for effectual subgroups (C_k). Then the effectual and the ineffectual subgroups are identifiable. For any effectual subgroup $k\in S^e$, we have $TV_{(Y,T),k}>0$, and the sign of $\alpha_{k,k-1}$ is identified by the sign of $\Delta_k\mathbb{E}(Y|Z)$.*

Proof of Lemma A.1. This proof is based on the proof of Theorem 1 in Imbens and Angrist (1994). Since the order of the support Ω_Z is given, we have that $Pr(z_k)\geq Pr(z_{k-1})$ for $k=1,2,...,K$. By virtue of (4), we know that

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

where $\Delta_k\mathbb{E}(Y|Z)$ is identifiable given the joint distribution of (Y,Z). If $\Delta_k\mathbb{E}(Y|Z)=0$, then it is either $\alpha_{k,k-1}=0$, or $P(z_k)-P(z_{k-1})=0$ implying $\gamma_k=0$. In addition, because $g(\cdot)$ is a known function, we can also identify whether $\sum_{l=k}^{K}(g(z_l)-\mathbb{E}[g(Z)])\pi_l$ is zero or not based on the distribution of Z. Thus, those ineffectual subgroups are identified, as well as those effectual subgroups.

For those effectual subgroups, we have $\Delta_k\mathbb{E}(Y|Z)\neq 0$ and $Pr(z_k)-Pr(z_{k-1})\neq 0$. Therefore, $TV_{(Y,T),k}>0$ follows directly from Lemma A.2-(ii). In addition,

$$\text{sign}(\alpha_{k,k-1})=\text{sign}\left(\frac{\Delta_k\mathbb{E}(Y|Z)}{\Pr(C_k)}\right)=\text{sign}(\Delta_k\mathbb{E}(Y|Z)),$$

where $\text{sign}(x)=1[x\geq 0]-1[x<0]$, and the last equality is due that $Pr(z_k)>Pr(z_{k-1})$. Hence, the sign of $\alpha_{k,k-1}$ of the effectual subgroup can be identified by the sign of $\Delta_k\mathbb{E}(Y|Z)$. □

Denote the collection of the effectual subgroups $S^e\subset\{1,2,...,K\}$. If $k\in S^e$, then subgroup (C_k) has nonzero contribution to the weighted average treatment effect α^{IV}. By definition of the effectual subgroup, α^{IV} and α^{Mis}, we have $\gamma_k\alpha_{k,k-1}=\gamma_k^{Mis}\alpha_{k,k-1}=0$ for $k\notin S^e$. Therefore,

$$\alpha^{IV}=\sum_{k\in S^e}\gamma_k\alpha_{k,k-1},\ \text{and}\ \alpha^{Mis}=\sum_{k\in S^e}\gamma_k^{Mis}\alpha_{k,k-1}.$$

Then, the sufficient conditions needed for the proofs about the identified sets α^{IV}, based on the expression of either α^{IV} or α^{Mis}, can be relaxed in the sense that they can be only imposed on S^e.

A.1.4 Proof of Lemma 3.1

Proof of Lemma 3.1. 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.1.5 Proof of Lemma 3.2

The proofs of Lemma 3.2 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.2 below. In what follows, we first prove Lemma A.2 and then proceed to the proof of Lemma 3.2.

Lemma A.2. *Under Assumptions 3.1-(ii)-(iv), 3.2-(i) and 3.3, 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.2. (i) This is a multi-valued instrument 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

$$Pr\left[\left\{(y,t) \in \Omega_Y \times \{0,1\} : \left|f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)\right| \neq 0\right\}\right] > 0. \tag{A7}$$

It can be verified by proof by contradiction as below. Suppose $\Delta_k \mathbb{E}(Y|Z) \neq 0$ but the probability in (A7) is zero. It means with probability one that

$$\begin{aligned}
& \left|f_{(Y,T)|Z=z_k}(y,t) - f_{(Y,T)|Z=z_{k-1}}(y,t)\right| = 0\\
\Leftrightarrow\;& f_{(Y,T)|Z=z_k}(y,t) = f_{(Y,T)|Z=z_{k-1}}(y,t), \text{ for both } t = 0,1\\
\Leftrightarrow\;& \sum_{t=0,1} f_{(Y,T)|Z=z_k}(y,t) = \sum_{t=0,1} f_{(Y,T)|Z=z_{k-1}}(y,t)\\
\Leftrightarrow\;& f_{Y|Z=z_k}(y) = f_{Y|Z=z_{k-1}}(y)\\
\Leftrightarrow\;& \Delta_k \mathbb{E}(Y|Z) = \int y[f_{Y|Z=z_k}(y) - f_{Y|Z=z_{k-1}}(y)] d\mu_Y(y) = 0,
\end{aligned} \tag{A8}$$

which contradicts $\Delta_k \mathbb{E}(Y|Z) \neq 0$. Therefore, $|\Delta_k \mathbb{E}(Y|Z)| > 0$ implies (A7), and we have that $TV_{(Y,T),k} > 0$ by definition. □

Now we can proceed to the proof of Lemma 3.2.

Proof of Lemma 3.2. (i) If $TV_{(Y,T),k} = 0$, then by Lemma A.2 $\Delta_k\mathbb{E}(Y|Z) = 0$, and any $\alpha_{k,k-1} \in \Theta$ satisfies the inequalities (5), (6) and (7). 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 3.3(ii), via replacing $\Delta_k\mathbb{E}(T|Z)$ and Δp_k in the proof of Lemma 3.3(ii) by $\Delta_k\mathbb{E}(Y|Z)$ and $\alpha_{k,k-1}$ respectively. Due to space limitation, we skip it. □

A.1.6 Proof of Lemma 3.3

Proof of Lemma 3.3. (i) If $TV_{(Y,T),k} = 0$, $\Delta_k\mathbb{E}(T|Z) = 0$ by Lemma A.2, and any $\Delta p_k \in [-1,1]$ satisfies the inequalities (8), (9) and (10). 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

$$f_{(Y,T)|Z=z_0}(y,t) = f_{(Y,T)|Z=z_1}(y,t) = ... = f_{(Y,T)|Z=z_K}(y,t) = f_{(Y,T)}(y,t) \tag{A9}$$

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:

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

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 3.1-(ii)-(iv) and 3.2-(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 (A9). 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 3.1-(ii)-(iv) and 3.2-(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 3.1-(ii)-(iv) and 3.2-(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 3.1-(ii)-(iv) and 3.2-(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:

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

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

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

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 $m \le k-1$. Furthermore, for $\forall m \ge k$,

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

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 $m \ge k$. Hence, we have shown that the DGP P_L^* generates (Y, T, Z).

Next, consider Δp_k under P_L^*:

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

which fulfills the proof of (a).

(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

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

First of all, noticing that

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

It's easy to check that DGP P_U^* satisfies Assumptions 3.1-(ii)-(iv) and 3.2-(i). Denote f^{*U} as the distribution functions under P_U^*. We first show that f^{*U} is well-defined in the sense that (b.1) the summation of the probabilities of all possible combinations for $\{D_k\}_{k=0}^K$ is one, and (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.

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

(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.

$$\begin{aligned}\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) &= f_{(Y,T,H)|Z=z_k}(y,t,0.5) - f_{(Y,T,H)|Z=z_{k-1}}(y,t,0.5)\\ &= f_{(Y,T)|Z=z_k}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) \geq 0] - f_{(Y,T)|Z=z_{k-1}}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) \geq 0]\\ &= \Delta_k f_{(Y,T)|Z}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) \geq 0] \geq 0. \end{aligned} \tag{A10}$$

Moreover,

$$\begin{aligned}\Delta_k f_{(Y,T,H)|Z}(y,t,-0.5) &= f_{(Y,T,H)|Z=z_k}(y,t,-0.5) - f_{(Y,T,H)|Z=z_{k-1}}(y,t,-0.5)\\ &= f_{(Y,T)|Z=z_k}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) < 0] - f_{(Y,T)|Z=z_{k-1}}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) < 0]\\ &= \Delta_k f_{(Y,T)|Z}(y,t)1[\Delta_k f_{(Y,T)|Z}(y,t) < 0] \leq 0. \end{aligned} \tag{A11}$$

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

(b.3) From (A10) and (A11) we have that

$$\begin{aligned}&\sum_{t=0,1}\int \left[\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) + \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)\right] d\mu_Y(y)\\ =&\sum_{t=0,1}\int \Delta_k f_{(Y,T)|Z}(y,t) d\mu_Y(y) = 0, \end{aligned} \tag{A12}$$

and

$$\begin{aligned}&\sum_{t=0,1}\int \left[\Delta_k f_{(Y,T,H)|Z}(y,t,0.5) - \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5)\right] d\mu_Y(y)\\ =&\sum_{t=0,1}\int \Delta_k f_{(Y,T)|Z}(y,t)\text{sign}(\Delta_k f_{(Y,T)|Z}(y,t)) d\mu_Y(y)\\ =&\sum_{t=0,1}\int |\Delta_k f_{(Y,T)|Z}(y,t)| d\mu_Y(y)\\ =&2TV_{(Y,T)|Z}. \end{aligned} \tag{A13}$$

Based on (A12) and (A13), we get that

$$\sum_{t=0,1}\int \Delta_k f_{(Y,T,H)|Z}(y,t,0.5) d\mu_Y(y) = TV_{(Y,T)|Z}, \tag{A14}$$

$$\sum_{t=0,1}\int \Delta_k f_{(Y,T,H)|Z}(y,t,-0.5) d\mu_Y(y) = -TV_{(Y,T)|Z}. \tag{A15}$$

Given (A14) and (A15), 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 $\forall m \leq k-1$,

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

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 3.1-(ii)-(iv) and 3.2-(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.1.7 Proof of Theorem 3.2

Proof of Theorem 3.2. We have $\min_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\} \leq \alpha_{k,k-1} \leq \max_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\}$ holds for $\forall k$. Since the weight $\gamma_k^{IV} \geq 0$ for all $k = 1,2,\dots,K$, then

$$\sum_{k=1}^{K}\gamma_k^{IV}\alpha_{k,k-1} \leq \max_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\}\sum_{k=1}^{K}\gamma_k^{IV} = \max_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\},$$
$$\sum_{k=1}^{K}\gamma_k^{IV}\alpha_{k,k-1} \geq \min_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\}\sum_{k=1}^{K}\gamma_k^{IV} = \min_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\}.$$

Thus, $\min_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\} \leq \alpha^{IV} \leq \max_{k\in\{1,2,\dots,K\}}\{\alpha_{k,k-1}\}$. Because each LATE is partially identified by $\Theta_k^{\alpha}(\mathbf{P})$, based on which we have $\alpha^{IV} \in \bigcup_{k=1,2,\dots,K}\Theta_k^{\alpha}(\mathbf{P})$. □

A.1.8 Proof of Theorem 3.3

Proof of Theorem 3.3. Since $\xi = \sum_{k=1}^{K}\gamma_k^{IV}\Delta p_k$, we know that

$$\min_{k=1,2,\dots,K}\{\Delta p_k\} \leq \xi \leq \max_{k=1,2,\dots,K}\{\Delta p_k\}.$$

Thus, the identified set for ξ is $\bigcup_{k=1,2,\dots,K}\Theta_k^{p}(\mathbf{P})$. It results from Equation (3), $\alpha^{IV} = \xi\alpha^{Mis}$, that

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

where α^{Mis} is identifiable by observable data. □

A.1.9 Proof of Corollary 3.4

Proof of Corollary 3.4. It yields from the definitions of $\Delta_k\mathbb{E}(Y|Z)$ and $\Delta_k\mathbb{E}(T|Z)$ that

$$\Delta_k\mathbb{E}(Y|Z)/\Delta_k\mathbb{E}(T|Z) = \alpha_{k,k-1}/\Delta p_k.$$

From Theorem 3.2, we have that

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

where the last equality is due to the definition of $\Theta^{p}_{k}(\mathbf{P})$. Similarly, from Theorem 3.3 and (3)

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

□

A.1.10 Proof of Theorem 3.4

Proof of Theorem 3.4. 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 $\bar{\xi} \alpha^{Mis}$, and its sign is determined by the sign of α^{Mis}. □

A.1.11 Proof of Lemma 4.1

Proof of Lemma 4.1. Recall $\pi_k = Pr(Z = z_k)$. Denote $\varphi_k = \frac{1[Z=z_k]\pi_{k-1} - 1[Z=z_{k-1}]\pi_k}{\pi_k \pi_{k-1}}$. Similar to Abadie (2003)'s binary instrument case, we can get for any generic random variable Q and $\forall k$,

$$\begin{aligned}
\Delta_k \mathbb{E}[Q|Z] = & \mathbb{E}[Q|Z = z_k] - \mathbb{E}[Q|Z = z_{k-1}] \\
= & \frac{1}{\pi_k} \mathbb{E}[\pi_k Q|Z = z_k] - \frac{1}{\pi_{k-1}} \mathbb{E}[\pi_{k-1} Q|Z = z_{k-1}] \\
= & \mathbb{E}\left[\frac{Q \times 1[Z = z_k]}{\pi_k} \right] - \mathbb{E}\left[\frac{Q \times 1[Z = z_{k-1}]}{\pi_{k-1}} \right] \\
= & \mathbb{E}\left[\frac{1[Z = z_k]\pi_{k-1} - 1[Z = z_{k-1}]\pi_k}{\pi_k \pi_{k-1}} Q \right] \\
= & \mathbb{E}[\varphi_k Q], \qquad \text{(A16)}
\end{aligned}$$

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

$$\begin{aligned}\Delta_k \mathbb{E}[h(Y,T)|Z] =& \sum_{t=0,1} \int h(y,t)\Delta_k f_{(Y,T)|Z}(y,t) d\mu_Y(y) \\ \leq& \frac{1}{2}\sum_{t=0,1}\int \left|\Delta_k f_{(Y,T)|Z}(y,t)\right| d\mu_Y(y) = TV_{(Y,T),k}, \end{aligned} \tag{A17}$$

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, (A17) also implies that for $\forall h \in \mathbf{H}$,

$$1 - \sum_{k' \neq k} TV_{(Y,T),k'} \leq 1 - \sum_{k' \neq k} \Delta_{k'}\mathbb{E}[h_{k'}(Y,T)|Z], \tag{A18}$$

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 (A17) and (A18) 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 (A16) to the above inequalities gives us the desired results. Same arguments can be applied to prove the results for Θ_k^p. □

A.1.12 Proof of Lemma 4.2

Proof of Lemma 4.2. Assumption 4.2(ii) implies $\mathbf{H}_n \subset \mathbf{H}_{n+1} \subset \cdots \subset \mathbf{H}$. Thus, it is straightforward that the approximated identified set $\widetilde{\Theta}_k^{\alpha}(\mathbf{P})$ and $\widetilde{\Theta}_k^{p}(\mathbf{P})$ cover the identified sets $\Theta_k^{\alpha}(\mathbf{P})$ 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 (A17) 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

$$TV_{(Y,T),k} = \sup_{h\in\mathbf{H}} \Delta_k\mathbb{E}[h(Y,T)|Z]. \tag{A19}$$

In addition, because $H = h_k^*(Y,T)$, where the random variable H is defined in the proof of Lemma 3.3 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 $l=1,2,...,L_n$

$$\begin{aligned}
&\max_{(y,t)\in I_{n,l}} \Delta_k f_{(Y,T)|Z}(y,t) - \min_{(y',t')\in I_{n,l}} \Delta_k f_{(Y,T)|Z}(y',t') \\
\leq &\max_{(y,t)\in I_{n,l}} f_{(Y,T)|Z=z_k}(y,t) - \min_{(y,t)\in I_{n,l}} f_{(Y,T)|Z=z_{k-1}}(y,t) - \min_{(y',t')\in I_{n,l}} f_{(Y,T)|Z=z_k}(y',t') \\
&\qquad + \max_{(y',t')\in I_{n,l}} f_{(Y,T)|Z=z_{k-1}}(y',t') \\
\leq &2M_0(\frac{2M_1}{L_n})^m. \qquad \text{(A20)}
\end{aligned}$$

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 (A20) 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,

$$h_k^*(y,t)\Delta_k f_{(Y,T)|Z}(y,t) - h_{k,n}^*(y,t)\Delta_k f_{(Y,T)|Z}(y,t) \leq M_n, \qquad \text{(A21)}$$

since the left hand side of (A21) is zero and $M_n \geq 0$. Secondly, for (y,t) such that $h_k^*(y,t) = h_{k,n}^*(y,t)$ and $|\Delta_k f_{(Y,T)|Z}(y,t)| \leq M_n$, (A21) 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, (A21) holds for $\forall (y,t) \in \Omega_Y \times \{0,1\}$.

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

which gives the first convergence in Lemma 4.2, since $M_n \to 0$ uniformly over $\mathbf{\Pi} \times \mathscr{P}_0$.

Moreover, recall that (A18) 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,

$$1-\sum_{k'\neq k} TV_{(Y,T),k'} = \inf_{\{h_{k'}\}\in\mathbf{H}^{K-1}} \left[1-\sum_{k'\neq k}\Delta_{k'}\mathbb{E}[h_{k'}(Y,T)|Z]\right] = 1-\sum_{k'\neq k}\sup_{h_{k'}\in\mathbf{H}}\Delta_{k'}\mathbb{E}[h_{k'}(Y,T)|Z]. \quad \text{(A22)}$$

Hence, it follows from (A21) and (A22) 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.1.13 Proof of Theorem 4.1

By abuse of notation, denote by Θ the parameter space of θ_k, and denote by $\widetilde{\Theta}_k^{\theta}(\mathbf{P})$ the approximated identified set of θ_k. Before we proceed to the proof of Theorem 4.1, let us introduce a theorem used in Theorem 21 by Ura (2018), which is taken from Corollary 5.1 and Theorem 6.1 in Chernozhukov et al. (2019) of the asymptotic size and power of the test statistic $\tau(\theta_k, \pi_k, \pi_{k-1})$. Due that the test is for the p_n inequalities which characterize the approximated identified set, results in Theorem A.1 below are also derived for the the approximated identified set $\widetilde{\Theta}_k^{\theta}(\mathbf{P})$.

Theorem A.1. *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

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

Under assumptions in Theorem 4.1,

(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 4.1 can be shown as below.

Proof of Theorem 4.1. 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 4.1, for any $\mathbf{P}\in\mathscr{P}_0$ such that $\theta_k\in\widetilde{\Theta}_k^\theta(\mathbf{P})$, we can get

$$\begin{aligned}
&Pr\left[\theta_k\notin\mathscr{C}_{\theta_k}(\eta+2\eta_\pi)\right]\\
=&Pr\left[\theta_k\notin\mathscr{C}_{\theta_k}(\eta+2\eta_\pi),A\right]+Pr\left[\theta_k\notin\mathscr{C}_{\theta_k}(\eta+2\eta_\pi),A^C\right]\\
\leq&Pr\left[\theta_k\notin\mathscr{C}_{\theta_k}(\eta+2\eta_\pi),A\right]+Pr\left[A^C\right]\\
\leq&Pr\left[\theta_k\notin\mathscr{C}_{\theta_k}(\eta+2\eta_\pi),A\right]+Pr\left[\pi_k^0\notin\mathscr{C}_{\pi_k}(\eta_\pi)\right]+Pr\left[\pi_{k-1}^0\notin\mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right]\\
\leq&Pr\left[\tau(\theta_k,\pi_k^0,\pi_{k-1}^0)>c_k(\eta),A\right]+Pr\left[\pi_k^0\notin\mathscr{C}_{\pi_k}(\eta_\pi)\right]+Pr\left[\pi_{k-1}^0\notin\mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right]\\
\leq&Pr\left[\tau(\theta_k,\pi_k^0,\pi_{k-1}^0)>c_k(\eta)\right]+Pr\left[\pi_k^0\notin\mathscr{C}_{\pi_k}(\eta_\pi)\right]+Pr\left[\pi_{k-1}^0\notin\mathscr{C}_{\pi_{k-1}}(\eta_\pi)\right],
\end{aligned}\tag{A23}$$

where the second last inequality is by definition of $\mathscr{C}_{\theta_k}$. Therefore, it follows from Theorem A.1-(i), Assumption 4.1 and the convergence of $\widetilde{\Theta}_k^\theta(\mathbf{P})$ to $\Theta_k^\theta(\mathbf{P})$ in Lemma 4.2, 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.1-(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.1-(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 4.1-(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 (16)-(18) is violated. If (16) 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 (17) does not hold, it implies that

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

Based on the first convergence result in Lemma 4.2 and the fact that $\varphi_k|\alpha_{k,k-1}|$ is bounded by

Assumption 4.1, there exists some $h_k \in \mathbf{H}_n$ such that when n is large enough,

$$\mathbb{E}\left\{\varphi_k\left[|\alpha_{k,k-1}|h_k(Y,T)-\text{sign}(\alpha_{k,k-1})Y\right]\right\}>0. \tag{A24}$$

Let $c>0$ be the value of the left hand side of (A24). 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 (18) 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 (A24) can be applied to arrive the same conclusion, based on the second convergence in Lemma 4.2 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.1-(ii).

Case 2. $\theta_k=\Delta p_k$. Since $\Delta p_k\notin\Theta_k^p(\mathbf{P})$, at least one of Equations (19)-(21) is violated. The same arguments for **Case 1** can be applied to achieve the desired results. □

A.1.14 Proof of Corollary 4.1

Proof of Corollary 4.1. (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

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

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 4.1-(i) that

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

(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 $\alpha^{Mis}=\frac{\text{Cov}(Y,g(Z))}{\text{Cov}(T,g(Z))}$. 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

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

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 4.1 and Assumption 4.1,

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

(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 $\theta \in \Theta^\xi(\mathbf{P})$. □

A.2 Details about the Two-Step Multiplier Bootstrap

This section provides the 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 4.1 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.

A.2.1 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,

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

where $\hat{m}_j^B(\theta_k, \pi_k, \pi_{k-1}) = \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left[g_j(V_i, \theta_k, \pi_k, \pi_{k-1}) - \hat{m}_j(\theta_k, \pi_k, \pi_{k-1}) \right]$. 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 $\{V_i\}_{i=1}^n$.

A.3 How confidence intervals are constructed in practice

This section provides some details about the procedure to construct the confidence intervals of the there identified sets $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$. For illustration simplicity, we focus on the case of a single treatment proxy T case. By abuse of notation, we use θ_k to denote either $\alpha_{k,k-1}$ or Δp_k, and use $\hat{\mathscr{C}}_{\theta_k}(\eta)$ to represent their identified sets respectively.

Step 1. Construct identified sets of θ_k for $k = 1, 2, ..., K$.

(a) Partition the support of (Y, T) into L_n subsets as $\Omega_Y \times \{0, 1\} = \bigcup_{l=1,2,...,L_n} I_{n,l}$, where the subscript n means that the number of partitions L_n may increase with sample size n. For example, if Y takes two values {0,1}, then $L_n = 4$ with $I_{n,1} = (1, 1)$, $I_{n,2} = (1, 0)$, $I_{n,3} = (0, 1)$ and $I_{n,4} = (0, 0)$.

(b) Generate a function $h : \Omega_Y \times \{0, 1\} \mapsto \{0.5, -0.5\}$ such that $h(y, t)$ is a constant (either 0.5 or -0.5) over each partition $(y, t) \in I_{n,l}$. Since there are L_n partitions of $\Omega_Y \times \{0, 1\}$, then the total number of all possible functions h is $\kappa_n := 2^{L_n}$. From now on, we eliminate the subscript n in function h, and denote the series of functions h as $\{h_j\}_{j=1,2,...,\kappa_n}$. Using the example from step 1(a), there will be $\kappa_n = 2^4 = 16$ different functions h, as

		h_1	h_2	$h_{n,3}$	$\cdots$	h_6	h_7	$\cdots$	h_{12}	h_{13}	$\cdots$	h_{16}
Partitions	$I_{n,1} = (1, 1)$	0.5	-0.5	-0.5	...	0.5	-0.5	$\cdots$	0.5	-0.5	$\cdots$	0.5
	$I_{n,2} = (1, 0)$	0.5	-0.5	0.5	...	0.5	-0.5	$\cdots$	0.5	-0.5	$\cdots$	-0.5
	$I_{n,3} = (0, 1)$	0.5	-0.5	0.5	...	0.5	0.5	$\cdots$	-0.5	-0.5	$\cdots$	-0.5
	$I_{n,4} = (0, 0)$	0.5	-0.5	0.5	...	-0.5	0.5	$\cdots$	-0.5	0.5	$\cdots$	-0.5

where h_1 is constant 0.5 on all 4 partitions; h_2 is constant -0.5 on all 4 partitions; h_3 to h_6 map 1 partition to -0.5 and rest 3 partitions to 0.5; h_7 to h_{12} map 2 partitions to -0.5 and rest 2 partitions to 0.5; and h_{13} to h_{16} map 3 partitions to -0.5 and 1 partition to 0.5.

(b.1) Remark: The simulation results in Section A.7 suggests that the number of partitions κ_n impacts the width of $\mathscr{C}^{\alpha}(\beta^{\alpha})$ differently, based on the misclassification probability, the number of available treatment proxies, and the instrument strength. However, since its effects on $\mathscr{C}^{p}(\beta^{p})$ are negligible, the choice of κ_n is not critical when the second strategy is applied. The same reasoning applies to the third strategy. As for the first strategy, a data-driven process of choosing κ_n is left for future research.[35]

(c) Design the grid of candidate values $(\delta_1, \delta_2, ..., \delta_M)$ for θ_k based on its parameter space Θ, for instance, an evenly distributed grid between the two boundaries of Θ. If $\theta_k = \alpha_k$ and $\Omega_Y = \{0, 1\}$, then the grid is over the interval $[-1, 1]$. If $\theta_k = \Delta p_k$, then the grid is also over the interval $[-1, 1]$.[36]

[35] For binary instrument cases, Ura (2018) suggests a possible rule-of-thumb choice for κ_n such that at least a certain number of observations, e.g. 30, are contained in each partition $I_{n,l}$.

[36] In practice, we may restrict the grid for Δp_k within $[0, 1]$. It is reasonable because negative values of Δp_k indicates the correct classification probability of treated individuals as treated is less than the incorrect classification probability of untreated individuals as treated. It contradicts the assumption $p_{1,k} \geq p_{0,k}$ which is typically invoked in the literature, see e.g. Battistin and Sianesi (2011), and implies that the quality of the collected data is too poor to draw any reliable conclusions.

(d) Obtain $\sqrt{n}$-consistent estimator $\hat{\pi}_k$ of $\pi_k = Pr(Z = z_k)$ and its associated η_π-confidence interval $\mathscr{C}_{\pi_k}(\eta_\pi)$.

(e) In what follows, we take $\theta_k = \alpha_{k,k-1}$ as an example to illustrate the rest of this step. Recall that $V_i = (Y_i, T_i, S_i, Z_i)$. Given the κ_n functions $\{h_j\}$ in (b), for each candidate value δ_m with $m = 1, 2, ..., M$ in the grid of θ_k in (c), following Lemma 4.1 to calculate the sample analogue of the moment inequalities $\hat{m}_j(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})$, and their sample variance $\hat{\sigma}_j^2(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})$ as:

$$\begin{aligned}\hat{m}_j(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) &= \frac{1}{n}\sum_{i=1}^{n} g_j(V_i, \delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}),\\ \hat{\sigma}_j^2(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) &= \frac{1}{n}\sum_{i=1}^{n}\left[g_j(V_i, \delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) - \hat{m}_j(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})\right]^2,\end{aligned}$$

where

$$\begin{aligned}&g_1(V_i, \delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) = -\hat{\varphi}_{i,k}\text{sign}(\delta_m)Y_i,\\ &g_j(V_i, \delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) = \hat{\varphi}_{i,k}\left[|\delta_m|h_k(Y_i, T_i) - \text{sign}(\delta_m)Y_i\right], \text{ for } j = 2, ..., \kappa_n + 1,\\ &g_j(V_i, \delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) = \hat{\varphi}_{i,k}\text{sign}(\delta_m)Y_i - |\delta_m|\left(1 - \sum_{k'\neq k}\hat{\varphi}_{i,k'}h_{k'}(Y_i, T_i)\right), \text{ for } j = \kappa_n + 2, ..., p_n,\end{aligned}$$

with $p_n = 1 + \kappa_n + \kappa_n^{K-1}$ and $\hat{\varphi}_{i,k} = \frac{1[Z_i = z_k]\hat{\pi}_{k-1} - 1[Z_i = z_{k-1}]\hat{\pi}_k}{\hat{\pi}_k\hat{\pi}_{k-1}}$ as defined in Lemma 4.1.

(f) For each δ_m in the grid of θ_k, calculate the test statistic for $H_0:\ \mathbb{E}[g_j(\delta_m, \pi_k, \pi_{k-1})] \leq 0$ for all $j = 1, 2, ..., p_n$, defined as

$$\tau(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) = \max_{1\leq j\leq p_n}\frac{\sqrt{n}\hat{m}_j(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})}{\hat{\sigma}_j(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})}.$$

(g) For each $(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})$, calculate $c_k(\eta)$, the critical value of the test statistic $\tau(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1})$ follows the detailed algorithm in Appendix A.2.

(h) Repeat (e)-(g) for all candidate values in the grid of θ_k and all $(\tilde{\pi}_k, \tilde{\pi}_{k-1}) \in \mathscr{C}_{\pi_k}(\eta_\pi) \times \mathscr{C}_{\pi_{k-1}}(\eta_\pi)$. Then, construct the $(1 - \eta - 2\eta_\pi)$-confidence interval of θ_k as $\left[\underline{\delta}_k, \overline{\delta}_k\right]$, where

$$\begin{aligned}\underline{\delta}_k &= \min_{m=1,...,M,\ \tilde{\pi}_k\in\mathscr{C}_{\pi_k}(\eta_\pi),\ \tilde{\pi}_{k-1}\in\mathscr{C}_{\pi_{k-1}}(\eta_\pi)}\left\{\delta_m:\ \tau(\delta_m, \tilde{\pi}_k, \tilde{\pi}_{k-1}) \leq c_k(\eta)\right\},\\ \overline{\delta}_k &= \max_{m=1,...,M,\ \tilde{\pi}_k\in\mathscr{C}_{\pi_k}(\eta_\pi),\ \tilde{\pi}_{k-1}\in\mathscr{C}_{\pi_{k-1}}(\eta_\pi)}\left\{\delta_m:\ \tau(\delta_m, \tilde{\pi}_k, \tilde{\pi}_{k-1}) \leq c_k(\eta)\right\}.\end{aligned} \tag{A25}$$

(h') In practice, the confidence interval of θ_k in step (h) can be replaced by a simplified version as discussed in Section 4 and (23), denoted by $\hat{\mathscr{C}}_{\theta_k}(\eta) := \left[\hat{\underline{\delta}}_k, \hat{\overline{\delta}}_k\right]$:

$$\begin{aligned} \hat{\underline{\delta}}_k &= \min_{m=1,\dots,M} \left\{\delta_m : \ \tau(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) \leq c_k(\eta)\right\}, \\ \hat{\overline{\delta}}_k &= \max_{m=1,\dots,M} \left\{\delta_m : \ \tau(\delta_m, \hat{\pi}_k, \hat{\pi}_{k-1}) \leq c_k(\eta)\right\}, \end{aligned} \tag{A26}$$

where we only repeat the process in (e)-(g) for all candidates δ_m while fixing $(\hat{\pi}_k, \hat{\pi}_{k-1})$ to be the $\sqrt{n}$-consistent estimators of (π_k, π_{k-1}).

Step 2. Construct the confidence interval of the identified sets of α^{IV}. Denote $(\underline{\delta}_k^{\alpha}, \overline{\delta}_k^{\alpha})$ and $(\underline{\delta}_k^{p}, \overline{\delta}_k^{p})$ as the corresponding lower and upper bounds of $\alpha_{k,k-1}$ and Δp_k in (A25), respectively.

(a) For the confidence interval of $\Theta^{\alpha}(\mathbf{P})$ is given by

$$\mathscr{C}^{\alpha}(\beta^{\alpha}) = \left[\min_{k=1,2,\dots,K} \left\{\underline{\delta}_k^{\alpha}\right\}, \max_{k=1,2,\dots,K} \left\{\overline{\delta}_k^{\alpha}\right\}\right],$$

where $\underline{\delta}_k^{\alpha}$ and $\overline{\delta}_k^{\alpha}$ can be replaced by the simplified version in (A26).

(b) For the confidence interval of $\Theta^{p}(\mathbf{P})$, firstly, calculate the $\eta_{\alpha^{Mis}}$-confidence interval $\mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})$ of the estimand α^{Mis}. Then, we can construct the confidence interval as

$$\mathscr{C}^{p}(\beta^{p}) = \left[\min_{k=1,2,\dots,K,\ \alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{\alpha \times \underline{\delta}_k^{p}\right\}, \max_{k=1,2,\dots,K,\ \alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{\alpha \times \overline{\delta}_k^{p}\right\}\right],$$

where $\underline{\delta}_k^{p}$ and $\overline{\delta}_k^{p}$ can be replaced by the simplified version in (A26).

(c) For the confidence interval of $\Theta^{\xi}(\mathbf{P})$, still calculate the $\eta_{\alpha^{Mis}}$-confidence interval $\mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})$ of the Wald estimand α^{Mis} first. Then, create the possible range of ξ, i.e. $[\underline{\xi}, \overline{\xi}]$, according to the external information (e.g. following the suggestions discussed in Section 4). At last, construct the confidence interval as

$$\mathscr{C}^{\xi}(\beta^{\xi}) = \left[\min_{\alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{\alpha \times \underline{\xi}\right\}, \max_{\alpha \in \mathscr{C}_{\alpha^{Mis}}(\eta_{\alpha^{Mis}})} \left\{\alpha \times \overline{\xi}\right\}\right].$$

A.4 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 3.2 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. Let us first introduce a lemma analogue to Lemmas 3.2 and 3.3, but under two treatment measures T and S. Denote $\Theta_k^{p^W}(\mathbf{P})$ as $\Theta_k^{p}(\mathbf{P})$ associated with $W \in \{T, S\}$.

Lemma A.3. *Let Assumptions 3.1-(ii)-(iv) and 3.3 hold, and suppose Assumption 3.2-(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*

$$\Theta_k^{\alpha}(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z)}{TV_{(Y,T,S),k}}, \frac{\Delta_k \mathbb{E}(Y|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z) < 0. \end{cases} \tag{A27}$$

(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 (A27) 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*

$$\Theta_k^{p^W}(\mathbf{P}) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(W|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}, \frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}}\right], & \text{if } \Delta_k \mathbb{E}(W|Z) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(W|Z) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(W|Z)}{TV_{(Y,T,S),k}}, \frac{\Delta_k \mathbb{E}(W|Z)}{1-\sum_{k'\neq k} TV_{(Y,T,S),k'}}\right], & \text{if } \Delta_k \mathbb{E}(W|Z) < 0. \end{cases} \tag{A28}$$

(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 (A28) 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 3.3, 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

$$
\begin{aligned}
Z \sim f_Z,\ \ (D_{k-1},D_k)|_Z = (0,1),\ \ D_l \leq D_w \text{ if } l < w\\
(Y_1,T_1,S_1)|_{(\{D_k\}_{k=0}^K,Z)} &\sim f_{(Y,T,S)|Z=z_k}\\
(Y_0,T_0,S_0)|_{(\{D_k\}_{k=0}^K,Z)} &\sim f_{(Y,T,S)|Z=z_{k-1}}.
\end{aligned}
$$

P^*_U should be changed to

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

□

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 3.1-(ii)-(iv) and 3.3 hold for T, and suppose Assumption 3.2-(i) is satisfied by both T and S. For $k = 1,2,...,K$,*

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

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 $k = 1, 2, ..., K$

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

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, $TV_{(Y,T,S),k} \leq Pr(C_k) \leq 1 - \sum_{k'\neq k} TV_{(Y,T,S),k'}$ follows from same proof of Lemma 3.1. □

Lemma A.4 says that the bounds of $Pr(C_k)$ shrink from $[TV_{(Y,T),k}, 1 - \sum_{k'\neq k} TV_{(Y,T),k'}]$, when only a single proxy T is used, to $[TV_{(Y,T,S),k}, 1 - \sum_{k'\neq k} TV_{(Y,T,S),k'}]$, when both T and S are used. The improved identified set of $Pr(C_k)$ also leads to narrower bounds of (i) LATEs, (ii) LATMs, and (iii) the IV estimand α^{IV}. The identified sets of $\alpha_{k,k-1}$ and Δp_k, and their sharpness results, are summarized by Lemma A.3. In what follows, we focus on refining our partial identification strategies for α^{IV} when multiple treatment proxies are available.

First strategy. Recall that, in our first partial identification strategy, we denote the identified sets of α^{IV} using only the sets of $\{\alpha_{k,k-1}\}_{k=1}^{K}$ as $\Theta^{\alpha}(\mathbf{P})$. 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 3.1, 3.3 hold hold for T. Suppose T and S satisfy Assumption 3.2.*

(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 3.2, Lemma A.4, and Lemmas A.2 and A.3 in Appendix A.1. □

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 $\alpha^{IV} \geq 0$. Similarly for A.1-(ii), $\Delta_k\mathbb{E}(Y|Z) \leq 0$, but $TV_{(Y,T,S),k} > 0$

for all k, leads to the same expression of $\Theta^{\alpha}(\mathbf{P})$ and $\alpha^{IV} \leq 0$. If the direction consistency of LATEs does not hold, the general form of $\Theta^{\alpha}(\mathbf{P})$ will simply be the union of the $\{\Theta^{\alpha}_{k}(\mathbf{P})\}_{k=1}^{K}$ 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). It is important to point out that, at this stage, no information on the misclassification error, nor on α^{Mis}, have been used yet.

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 proxies available.

Denote the estimand α^{Mis} in Equation (2), associated with T and S, as $\alpha^{Mis,T}$ and $\alpha^{Mis,S}$, respectively. Furthermore, denote the LATMs, for T and S, as Δp^{T}_{k} and Δp^{S}_{k}, respectively. Recall that we use $\Theta^{p}(\mathbf{P})$ to denote the identified set of α^{IV}, using α^{Mis} and the identified sets of $\{\Delta p_k\}_{k=1}^{K}$. 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 3.1, 3.3 hold for T and S. T and S are both binary and satisfy Assumption 3.2.*

(i) For $W \in \{T,S\}$, if $\Delta_k \mathbb{E}(W|Z) > 0$ for $\forall k = 1,2,...,K$, then

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

(ii) For $W \in \{T,S\}$, if $\Delta_k \mathbb{E}(W|Z) < 0$ for $\forall k = 1,2,...,K$, then

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

Proof of Corollary A.2. The proof follows from Theorem 3.3, Lemma A.4, and Lemmas A.2 and A.3 in Appendix A.1. □

Corollary A.2 states that, by employing multiple treatment measures, there are two sources of gains compared to Corollary 3.3. 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 $(\underline{\xi}^T, \overline{\xi}^T)$ and $(\underline{\xi}^S, \overline{\xi}^S)$ the lower and upper bounds of the LATMs Δp_k^T and Δp_k^S, respectively. Like before, these bounds may come from external sources of information.

Corollary A.3. *Let Assumption 3.1, 3.3 hold for T and S. T and S are both binary and satisfy Assumption 3.2. 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 3.4 and the fact that α^{IV} lies in both of the identified sets derived using T and S. □

We can summarize the results of this subsection as follows. When multiple treatment proxies are available, we have three further strategies to partially identify α^{IV}. The improvements of the identified sets 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 identified sets we can obtain, potentially, even tighter bounds of α^{IV}.

A.5 Extension: P-LATE with covariates

This section shows how to use P-LATE in the presence of covariates. We proceed in two steps. First, we introduce the extended assumptions required for P-LATE, as well as the identification target. Second, we present the main partial identification results.

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 $P(z, x) = \mathbb{E}(D|Z = z, X = x)$.

Assumption A.1. *(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 $\forall x \in \Omega_X$, $P(z, x)$ 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 $P(z_l, x) \leq P(z_w, x)$ implies $g(z_l) \leq g(z_w)$, or $P(z_l, x) \leq P(z_w, x)$ implies $g(z_l) \geq g(z_w)$;

Assumption A.2. *(Conditional ascending order) 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 $P(z_l, x) \leq P(z_w, x)$, and this order is known.*

Assumptions A.1 and A.2 extend Assumptions 3.1, 3.2 and 3.3 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:

$$\alpha^{IV}(x) = \frac{\text{Cov}(Y, g(Z)|X = x)}{\text{Cov}(D, g(Z)|X = x)} = \sum_{k=1}^{K} \gamma_k^{IV}(x)\alpha_{k,k-1}(x), \tag{A29}$$

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, which could be a proxy for D, or could correspond to reported values of D that are misclassified for some observations. In this case, we can obtain the biased conditional IV estimand $\alpha^{Mis}(x)$:

$$\alpha^{Mis}(x) = \frac{\text{Cov}(Y, g(Z)|X = x)}{\text{Cov}(T, g(Z)|X = x)} = \sum_{k=1}^{K} \gamma_k^{Mis}(x)\alpha_{k,k-1}(x), \tag{A30}$$

with weights

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

Derivations of (A29) and (A30) can be obtained under Assumption A.1 by applying similar arguments used in the proof of Theorem 3.1 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.1 hold for T. 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 the conditional ascending order in Assumption A.2, let us 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.5. *Under Assumptions A.1 and A.2, 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.5. This proof is a direct extension of the proof of Lemma 3.1 when 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 identified sets 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.1, we have that the conditional LATE satisfies

$$\Delta_k\mathbb{E}(Y|Z, X = x) = \alpha_{k,k-1}(x)P(C_k|X = x), \tag{A31}$$

based on which we can derive a identified set of $\alpha_{k,k-1}(x)$, denoted by $\Theta_k^{\alpha}(\mathbf{P},x) \subset \Theta$. Similarly, the following equation holds for each $\Delta p_k(x)$:

$$\Delta_k \mathbb{E}(T|Z,X=x) = \Delta p_k(x) P(C_k|X=x). \tag{A32}$$

which can be used to construct the identified set of $\Delta p_k(x)$, denoted by $\Theta_k^{p}(\mathbf{P},x) \subset [-1,1]$. Given (A31), (A32) and Lemma A.5, we can obtain the following Lemmas that establish the identified sets of $\alpha_{k,k-1}(x)$ and $\Delta p_k(x)$.

Lemma A.6. *Let Assumption A.1 and A.2 hold for T. 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:

$$\Theta_k^{\alpha}(\mathbf{P},x) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}, \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k}(x)}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z,X=x) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(Y|Z,X=x) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{TV_{(Y,T),k(x)}}, \frac{\Delta_k \mathbb{E}(Y|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}\right], & \text{if } \Delta_k \mathbb{E}(Y|Z,X=x) < 0; \end{cases} \tag{A33}$$

(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 (A33) is the sharp identified set of $\alpha_{k,k-1}(x)$.

Proof of Lemma A.6. The proof is a direct extension of the proof of Lemma 3.2 when conditional on X. □

Lemma A.7. *Let Assumption A.1 and A.2 hold for T. 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 $\Theta_k^{p}(\mathbf{P},x) = [-1,1]$. Whereas, if $TV_{(Y,T),k}(x) > 0$, then:

$$\Theta_k^{p}(\mathbf{P},x) = \begin{cases} \left[\frac{\Delta_k \mathbb{E}(T|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}, \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{TV_{(Y,T),k}(x)}\right], & \text{if } \Delta_k \mathbb{E}(T|Z,X=x) > 0, \\ \{0\}, & \text{if } \Delta_k \mathbb{E}(T|Z,X=x) = 0, \\ \left[\frac{\Delta_k \mathbb{E}(T|Z,X=x)}{TV_{(Y,T),k}(x)}, \frac{\Delta_k \mathbb{E}(T|Z,X=x)}{1-\sum_{k' \neq k} TV_{(Y,T),k'}(x)}\right], & \text{if } \Delta_k \mathbb{E}(T|Z,X=x) < 0; \end{cases} \tag{A34}$$

(ii) If $\max_{0 \leq m \leq K} TV_{(Y,T),m}(x) = 0$, then $\Theta_k^{p}(\mathbf{P},x) = [-1,1]$ is the sharp identified set of $\Delta p_k(x)$. Whereas, if $TV_{(Y,T),k}(x) > 0$ and $TV_{(Y,T),k'}(x) = 0$ for all $k' \neq k$, then $\Theta_k^{p}(\mathbf{P},x)$ in (A34) is the sharp identified set of $\Delta p_k(x)$.

Proof of Lemma A.7. The proof is a direct extension of the proof of Lemma 3.3 when conditional on X. □

For $x \in \Omega_X$, the identified set of $\alpha^{IV}(x)$ constructed using either the identified sets of $\{\alpha_{k,k-1}(x)\}_{k=1}^K$ or $\{\Delta p_k(x)\}_{k=1}^K$ or external information are denoted as $\Theta^\alpha(\mathbf{P},x)$, $\Theta^p(\mathbf{P},x)$ and $\Theta^\xi(\mathbf{P},x)$, respectively. Given Lemmas A.6 and A.7, 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.1 and A.2 hold for T. Then, for $\forall x \in \Omega_X$:

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

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

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

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

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

Strategy 2 with covariates. Let Assumption A.1 and A.2 hold for T. Then, for $\forall x \in \Omega_X$:

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

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

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

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

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

Strategy 3 with covariates. Let Assumption A.1 hold for T. 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 $\alpha^{Mis}(x) \geq 0$, then $\alpha^{IV}(x) \geq 0$ and $\Theta^\xi(\mathbf{P},x) = \left[\underline{\xi}(x)\alpha^{Mis}(x), \overline{\xi}(x)\alpha^{Mis}(x)\right]$.

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

where values of $\underline{\xi}(x)$ and $\overline{\xi}(x)$ may be obtained from external sources of information, such as: validation studies, administrative data, repeated measurements of the same individual, or from economic theory.

Further extension. We conclude this section by showing the technical challenge one would face if the identification target was the unconditional IV estimand:

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

Unlike Frölich (2007), since Z is discrete, the expression of $\mathrm{Cov}(Y,g(Z))/\mathrm{Cov}(D,g(Z))$ in terms of the conditional LATEs $\{\alpha_{k,k-1}(X)\}_{k=1}^{K}$ is not straightforward and might requires further restrictions.

Consider first the mean independence of Z to X, $\mathbb{E}[Z|X]=\mathbb{E}[Z]$. Although it weakens the necessity of assuming the unconfoundedness of the instrument(s), it may be infeasible in some empirical studies. Without restricting the mean independence of Z, $\mathbb{E}[Y(g(Z)-\mathbb{E}[g(Z)])|X]$ can no longer be expressed as some function of the conditional LATEs $\{\alpha_{k,k-1}(X)\}_{k=1}^{K}$.[37] Suppose that $\mathbb{E}[Z|X]=\mathbb{E}[Z]$. Then:

$$\begin{aligned}\mathbb{E}\Big[Y\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]&=\sum_{k=1}^{K}\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X)\alpha_{k,k-1}(X),\\ \mathbb{E}\Big[D\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]&=\sum_{k=1}^{K}\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X).\end{aligned}\tag{A35}$$

However, even if (A35) is satisfied,

$$\begin{aligned}\frac{\mathrm{Cov}(Y,g(Z))}{\mathrm{Cov}(D,g(Z))}&=\sum_{k=1}^{K}\frac{\mathbb{E}_X\left[\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X)\alpha_{k,k-1}(X)\right]}{\sum_{k=1}^{K}\mathbb{E}_X\left[\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X)\right]}\\ &\neq\sum_{k=1}^{K}\frac{\mathbb{E}_X\left[\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X)\right]}{\sum_{k=1}^{K}\mathbb{E}_X\left[\Pr(C_k|X)\sum_{l=k}^{K}\left(g(z_l)-\mathbb{E}[g(Z)]\right)\pi_l(X)\right]}\mathbb{E}_X[\alpha_{k,k-1}(X)]\end{aligned}$$

where the inequality holds in general. This is because, for two random variables A and B, $\mathbb{E}[AB]\neq\mathbb{E}[A]\mathbb{E}[B]$ unless A is uncorrelated with B. Therefore, we cannot conduct the partial identification using a method analogue to Strategies 1, 2 and 3, because all of them rely on the target estimand being rewritten as weighted average of some form, either conditional or unconditional, of LATEs.

Secondly, we wonder if $\mathrm{Cov}(Y,g(Z))/\mathrm{Cov}(D,g(Z))$ and $\mathbb{E}_X\left[\alpha^{IV}(X)\right]$ are two different estimands. Because if they are the same, then the suggested estimation method for $\mathbb{E}_X\left[\alpha^{IV}(X)\right]$ can be applied

[37] Because if $\mathbb{E}[Z|X]\neq\mathbb{E}[Z]$, then the term

$$\sum_{l=0}^{K}\mathbb{E}(Y|Z=z_0,X)(g(z_l)-\mathbb{E}[g(Z)])\pi_l(X)=\mathbb{E}[Y|Z=z_0,X](\mathbb{E}[g(Z)|X]-\mathbb{E}[g(Z)])=0$$

does not hold in general (unless $\mathbb{E}[Y|Z=z_0,X]=0$), which is required for deriving (A35).

to $\mathrm{Cov}(Y, g(Z))/\mathrm{Cov}(D, g(Z))$. However,

$$\begin{aligned}\frac{\mathrm{Cov}(Y, g(Z))}{\mathrm{Cov}(D, g(Z))} =& \frac{\mathbb{E}_X\Big\{\mathbb{E}\Big[Y\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]\Big\}}{\mathbb{E}_X\Big\{\mathbb{E}\Big[D\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]\Big\}} \\ \neq& \mathbb{E}_X\left\{\frac{\mathbb{E}\Big[Y\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]}{\mathbb{E}\Big[D\Big(g(Z)-\mathbb{E}[g(Z)]\Big)\Big|X\Big]}\right\} = \mathbb{E}_X\left[\alpha^{IV}(X)\right],\end{aligned}$$

where the inequality holds, unless the denominator $\mathbb{E}\left[D\left(g(Z)-\mathbb{E}[g(Z)]\right)|X\right]$ is degenerate, i.e. is not a function of X. One sufficient condition for degenerate $\mathbb{E}\left[D\left(g(Z)-\mathbb{E}[g(Z)]\right)|X\right]$ is that $(D, Z) \perp X$, which, however, is quite restrictive and infeasible in many scenarios such as observational studies.

A.6 Details: How to incorporate covariates in practice

There are two main scenarios. First, the simplest case is when all covariates in X take on a finite number of values, that is, Ω_X is a finite set. Angrist and Fernandez-Val (2010) (Section 4) also study the estimation and inference for conditional treatment effects when assuming covariates are discrete. Assuming covariates are discrete is not required for partial identification of $\alpha^{IV}(x)$, but it maintains the inference process very similar to before. Indeed, in this case, a practitioner can simply implement the same inference process outlined in Section 4 (and further detailed in Appendix A.3) for each covariate-defined 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.[38] In this case, we suggest to follow a method adopted by Dehejia and Wahba (1999) and Battistin and Sianesi (2011) which is based on the idea of stratification matching. More specifically, each of our main partial identification strategies can be implemented following three steps:

Step 1. For the sake of dimension reduction, denote $e(x) = \Pr(T = 1|X = x)$ as the observable propensity score of the treatment proxy T, which is an index summarizing the information contained in covariates. Estimate $e(x)$ from a logit or probit regression, 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 the sample 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. This step is equivalent to converting the continuous variable $e(x)$ into a discrete one.

Step 3. Within each stratum, proceed with the chosen partial identification strategy and conduct inference following the detailed procedure outlined in Section 4 (and further detailed in Appendix A.3). Specifically for Strategy 3, this means to obtain first an estimate of $\alpha^{Mis}(x)$ and its confidence interval, via a conventional 2SLS, using a sample for each stratum. Then, following Equation (26), construct the confidence interval of $\alpha^{IV}(x)$ using information on the misclassification error (see Section 5.2 for guidance).

Three final remarks are in order. Firstly, we focus on the conditional IV estimand, $\alpha^{IV}(x)$, because this parameter has a clear relationship with the conditional LATEs, which are the foundation of our partial identification strategies in the presence of covariates. Alternatively, one could target the unconditional IV estimand. For example, in case of a binary instrument, the practitioner has already the tools to target the unconditional LATE (Ura, 2018). This would be the same if the practitioner applied our strategies in the same context. Whereas, in the more general case, when

[38] 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 methods.

instrument(s) are discrete, targeting the unconditional IV estimand is not as straightforward. One way to construct a bound for the overall treatment effect $\mathbb{E}_X[\alpha^{IV}(x)]$ is to take expectations of the lower and upper bounds of the identified set of $\alpha^{IV}(x)$. However, a rigorous exploration along this line, particularly the development of methods able to deal with continuous covariates without relying on stratification, is beyond the scope of this paper and left for future research.[39]

Secondly, in the context of an endogenous treatment response model with valid instrument(s) and exogenous covariates, the 2SLS estimator is commonly adopted and it is often justified by imposing linear model restrictions with constant treatment response (see e.g. Heckman and Robb (1985) and Angrist (2001) among many others). Hence, it is interesting to see what is the potential advantage for us if we use these further restrictions. Suppose, first, that the binary treatment D is observed and the practitioner assumes (i) $Y = \alpha D + \beta X + V$ (model linearity and constant effect), with V being the unobservable error term, and (ii) $E[V|g(Z), X] = 0$ (instrument validity). Then, the conditional IV estimand defined in Equation (27), $\alpha^{IV}(x)$, becomes invariant to covariates:

$$\alpha^{IV}(x) = \frac{Cov(Y, g(Z)|X = x)}{Cov(D, g(Z)|X = x)} = \frac{\alpha Cov(D, g(Z)|X = x) + \beta Cov(V, g(Z)|X = x)}{Cov(D, g(Z)|X = x)} = \alpha,$$

and, under mild regularity conditions, the probability limit of the linear 2SLS estimator converges to the true value:

$$\begin{pmatrix} \hat{\alpha}^{2SLS} \\ \hat{\beta}^{2SLS} \end{pmatrix} = \left[\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} g(Z_i) \\ X_i \end{pmatrix} \begin{pmatrix} D_i \\ X_i \end{pmatrix}' \right]^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} g(Z_i) \\ X_i \end{pmatrix} Y_i \right] \xrightarrow{p} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{A36}$$

However, when D is unobserved and T is used as a proxy, if we use the same assumptions (i) and (ii) as above, the mismeasured conditional IV estimand $\alpha^{Mis}(x)$ becomes:[40]

$$\alpha^{Mis}(x) = \frac{Cov(Y, g(Z)|X = x)}{Cov(T, g(Z)|X = x)} = \frac{Cov(Y, g(Z)|X = x)}{Cov(D, g(Z)|X = x)} \frac{Cov(D, g(Z)|X = x)}{Cov(T, g(Z)|X = x)} = \frac{\alpha}{\xi(x)},$$

where $\xi(x)$ is the weighted average of the conditional LATMs. This expression makes clear that a feasible linear 2SLS estimator $\widetilde{\alpha}^{2SLS}$, obtained by replacing D with T in Equation (A36), does not converge in probability to $\alpha^{Mis}(x)$ (nor, of course, to α). This is because $\alpha^{Mis}(x)$ varies with x through $\xi(x)$, whereas $\widetilde{\alpha}^{2SLS}$ is constant for all individuals. Thus, even when model linearity and constant effect are assumed, without making further (strong) restrictions on the misclassification error, our strategy 3 cannot be implemented by using the conventional 2SLS estimator on the entire

[39] The challenge is the following. Suppose the covariates are continuous, then the inference procedure should be conducted for identified set characterized by unconditional moment inequalities, rather than conditional ones. This is because the probability of observing samples conditional on covariates at any fixed value $x \in \Omega_X$ is zero. However, consider Strategy 1 as an example; one can show that:

$$\mathbb{E}_X\left[\min_k\{\alpha_{k,k-1}(X)\}\right] \leq \mathbb{E}_X[\alpha^{IV}(X)] \leq \mathbb{E}_X\left[\max_k\{\alpha_{k,k-1}(X)\}\right].$$

Since the random variables $\min_k\{\alpha_{k,k-1}(X)\}$ and $\max_k\{\alpha_{k,k-1}(X)\}$ are order statistics whose distributions are complicated and unknown, we fail to obtain explicit identified set of $\mathbb{E}_X[\alpha^{IV}(X)]$ characterized by unconditional moment inequalities. In Appendix A.5 we provide further details and insights.

[40] See Appendix A.5 for the definition of the estimand $\alpha^{Mis}(x)$ and further details.

sample in place of $\alpha^{Mis}(x)$. Hence, also with a linear model and constant effect, the way to proceed to obtain bounds of the conditional IV estimand $\alpha^{IV}(x)$ is to follow the procedure outlined in Step 1 to 3.

Finally, the discussion above raises the question of whether and under what conditions it is possible to link the feasible linear 2SLS estimators with covariates $\widetilde{\alpha}^{2SLS}$ to the conditional IV estimand $\alpha^{IV}(\text{x})$. As discussed in Abadie (2003), Section 5, when the instrument is binary and $Pr(Z=1|X)$ is linear in X, the infeasible linear 2SLS estimator $\hat{\alpha}^{2SLS}$ converges to the best linear least squares approximation of LATE. The investigation of whether $\widetilde{\alpha}^{2SLS}$ has a similar interpretation, and how to use it for bias correction, is beyond the scope of this paper and left for future research.

A.7 Details of Monte Carlo Simulations

Consider the following data generating process (DGP):

$$\begin{aligned} Y_0 &= 0.5 + 0.2X + O + V_0, \\ Y_1 &= 1.5 + 0.2X + O + V_1, \\ Y &= DY_1 + (1-D)Y_0, \end{aligned}$$

where Y_0 and Y_1 are potential outcomes, X is an independently generated covariate taking values in $\Omega_X = \{0,1,2,3\}$ with equal probabilities, O is unobservable (omitted variable), and V_0 and V_1 are standard normal random errors. The unobserved true treatment D is generated by

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

where $\gamma_0 = -2$, $\gamma_1 \in \{1, 1.5\}$ (instrument strength) and $\gamma_2 = 0.5$. The randomly generated discrete instrument Z takes values in a finite set $\Omega_Z = \{0,1,2\}$ with probabilities $\pi_0 = 0.4, \pi_1 = 0.4, \pi_2 = 0.2$. Assume we can observe (Y, Z, T, S, X), where T and S are misclassified endogenous binary treatment proxies generated by:

$$\begin{aligned} T &= DT_1 + (1-D)T_0, \text{ where } T_0 = 1[\Phi(U_T) < \underline{\mu}_T], \;\; T_1 = 1[\Phi(U_T) \geq \overline{\mu}_T], \\ S &= DS_1 + (1-D)S_0, \text{ where } S_0 = 1[\Phi(U_S) < \underline{\mu}_S], \;\; S_1 = 1[\Phi(U_S) \geq \overline{\mu}_S]. \end{aligned} \tag{A37}$$

where U_T and U_S are the key components of the misclassification error of the treatment proxies T and S, respectively. For $W = \{T, S\}$, $\underline{\mu}_W$ is the probability of opposite reporting, and $1 - \overline{\mu}_W$ is the probability of correctly reporting. In our simulations, we set $\underline{\mu}_T = 0.05$ and $\overline{\mu}_T \in \{0.1, 0.3\}$; $\underline{\mu}_S = 0.05$ and $\overline{\mu}_S = 0.25$. Furthermore, the unobservable terms $(O, V_D, U_T, U_S)'$ follow a joint normal distribution:

$$\begin{pmatrix} O \\ V_D \\ U_T \\ U_S \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 & 0.2 & 0.2 \\ -0.5 & 1 & -0.2 & -0.1 \\ 0.2 & -0.2 & 1 & 0 \\ 0.2 & -0.1 & 0 & 1 \end{pmatrix}\right],$$

Since U_T, U_S are endogenous, in the sense that they are correlated with $(O, V_D)'$, the misclassification errors (T_1, T_0) and (S_1, S_0) are also endogenous.

Table A1 reports the true values of the LATEs $\alpha_{k,k-1}$ in panel (a), and of the LATMs of proxy T ($\Delta p_k^T = p_{1,k}^T - p_{0,k}^T$) in panel (b), conditional on $X = x$ and under different values of $(\underline{\mu}_T, \overline{\mu}_T)$, as well as their identified sets $\Theta_k^\alpha(\mathbf{P})$ and $\Theta_k^{p^T}(\mathbf{P})$. As we can see, the identified set of LATEs becomes wider as the probability of correct recall $1 - \overline{\mu}_T$ decreases (more misclassification error), while the true value of $\alpha_{k,k-1}$ remains the same. In addition, as $1 - \overline{\mu}_T$ decreases, the true value of Δp_k^T decreases, with its identified set moving downward but not necessarily expanding. Finally, it is clear that the information provided by the multiple proxies helps to narrow down the identified sets.

Figure A1 and A2 plot the true values of α^{IV}, $\alpha^{Mis,T}$ and $\alpha^{Mis,S}$, as well as the identified sets of α^{IV}, using our three strategies. In both figures the instrument strength is $\gamma_1 = 1$, whereas the misclassification error is $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$, respectively. The full set of numerical results, for each figure and different values of instrument strength, are reported in Table A2. Specifically for strategy 3, we consider only two examples discussed in Section 5.2. First, Example 1, with $\xi^T \in [\underline{\xi}^T, \overline{\xi}^T]$, where $\underline{\xi}^T = 1 - 2\overline{\mu}_T$ and $\overline{\xi}^T = 1 - \overline{\mu}_T$, the resulting identified set is denoted by $\Theta^{\xi}(\mathbf{P})$. Second, Example 4, with $\underline{\xi}^T = \overline{\xi}^T = 1 - \overline{\mu}_T - \underline{\mu}_T$, with the resulting identified set denoted by $\Theta^{\xi}_{*}(\mathbf{P})$.

Looking at the figures, one can see that, as the probability of correctly reporting $1-\overline{\mu}_T$ decreases (more misclassification error), the bias of $\alpha^{Mis,T}$ gets worse. In addition, the identified sets $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ all become wider. Moreover, the upper bounds of $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are smaller than $\alpha^{Mis,T}$ and $\alpha^{Mis,S}$ in all different settings. The upper bound of $\Theta^{\alpha}(\mathbf{P})$ becomes smaller than $\alpha^{Mis,T}$ when, for example, the correct reported probability is relatively small ($\gamma_1 = 1, \overline{\mu}_T = 0.3$), or when multiple proxies are available. This implies that, by implementing our method, the results would outperform those obtained following a naïve IV approach. Besides, consistent with the theoretical results, using multiple proxies T and S reduces the width of the identified sets of α^{IV}. Finally, in Appendix one can see that, as expected, using a stronger instrument significantly narrows down the width of the identified sets.

Table A1: Identified Set of LATEs and LATMs under Endogenous Misclassification

(a) LATE

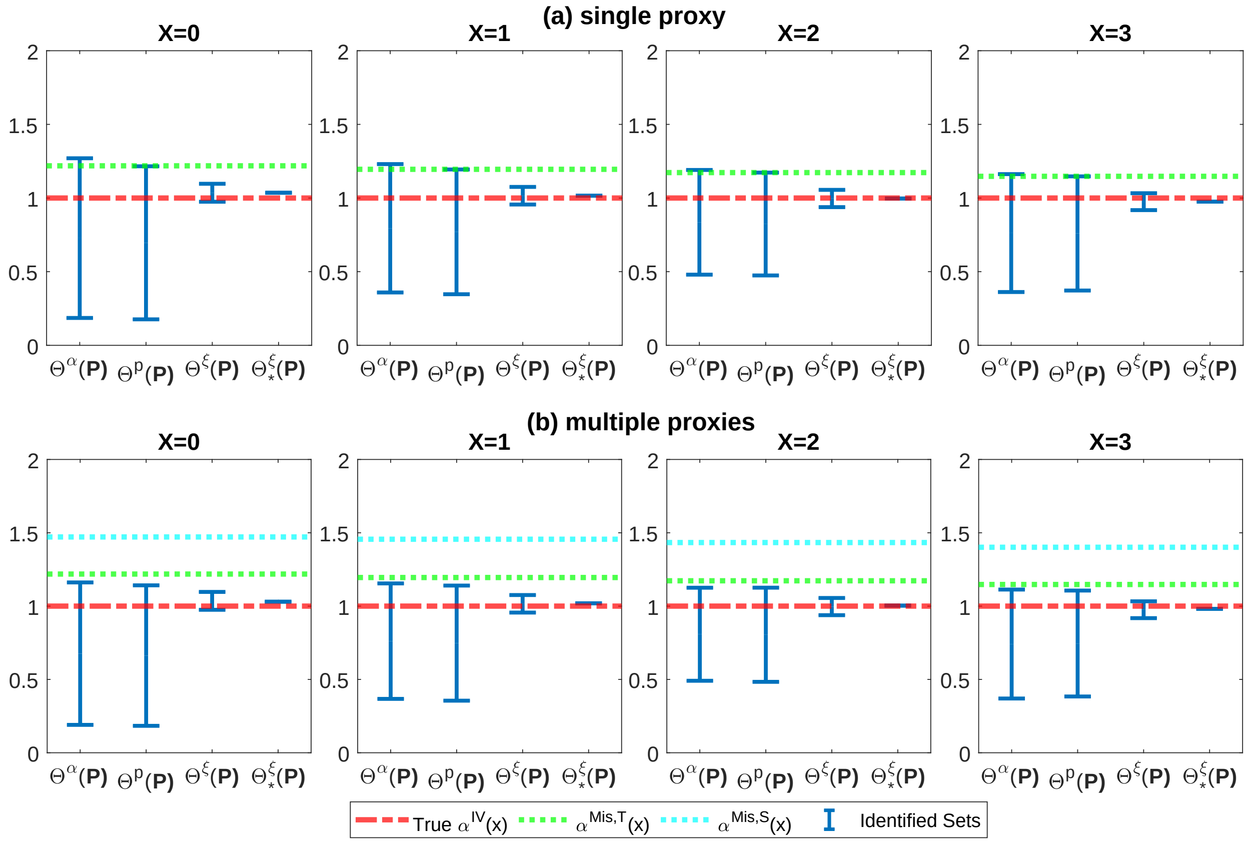
			$\Theta_1^\alpha(\mathbf{P})$			$\Theta_2^\alpha(\mathbf{P})$		
		$(\underline{\mu}_T,\overline{\mu}_T)$	$\alpha_{1,0}$	singe	multiple	$\alpha_{2,1}$	single	multiple
$\gamma_1 = 1$	$X = 0$	(0.05,0.1)	1	[0.194,1.320]	[0.199,1.217]	1	[0.382,1.201]	[0.386,1.136]
		(0.05,0.3)		[0.177,1.717]	[0.190,1.327]		[0.374,1.540]	[0.384,1.234]
	$X = 1$	(0.05,0.1)	1	[0.220,1.276]	[0.226,1.175]	1	[0.407,1.208]	[0.412,1.141]
		(0.05,0.3)		[0.201,1.697]	[0.217,1.301]		[0.395,1.534]	[0.407,1.232]
	$X = 2$	(0.05,0.1)	1	[0.255,1.272]	[0.262,1.184]	1	[0.426,1.183]	[0.431,1.121]
		(0.05,0.3)		[0.235,1.672]	[0.254,1.294]		[0.405,1.496]	[0.420,1.202]
	$X = 3$	(0.05,0.1)	1	[0.289,1.259]	[0.296,1.168]	1	[0.443,1.171]	[0.450,1.115]
		(0.05,0.3)		[0.263,1.633]	[0.286,1.271]		[0.425,1.483]	[0.442,1.196]
$\gamma_1 = 1.5$	$X = 0$	(0.05,0.1)	1	[0.528,1.241]	[0.547,1.160]	1	[0.690,1.152]	[0.705,1.105]
		(0.05,0.3)		[0.455,1.630]	[0.522,1.275]		[0.644,1.426]	[0.685,1.173]
	$X = 1$	(0.05,0.1)	1	[0.577,1.245]	[0.598,1.164]	1	[0.702,1.154]	[0.719,1.107]
		(0.05,0.3)		[0.497,1.622]	[0.567,1.271]		[0.648,1.422]	[0.694,1.173]
	$X = 2$	(0.05,0.1)	1	[0.614,1.232]	[0.632,1.151]	1	[0.698,1.142]	[0.718,1.102]
		(0.05,0.3)		[0.537,1.598]	[0.608,1.268]		[0.642,1.408]	[0.692,1.165]
	$X = 3$	(0.05,0.1)	1	[0.653,1.232]	[0.671,1.155]	1	[0.698,1.142]	[0.719,1.102]
		(0.05,0.3)		[0.582,1.596]	[0.652,1.263]		[0.632,1.399]	[0.689,1.161]

(b) LATM

			$\Theta_1^{p^T}(\mathbf{P})$			$\Theta_2^{p^T}(\mathbf{P})$		
		$(\underline{\mu}_T,\overline{\mu}_T)$	Δp_1^T	single	multiple	Δp_2^T	single	multiple
$\gamma_1 = 1$	$X = 0$	(0.05,0.1)	0.765	[0.146,0.991]	[0.149,0.914]	0.832	[0.317,0.997]	[0.320,0.943]
		(0.05,0.3)	0.520	[0.090,0.879]	[0.097,0.680]	0.611	[0.227,0.933]	[0.233,0.748]
	$X = 1$	(0.05,0.1)	0.774	[0.171,0.994]	[0.176,0.916]	0.836	[0.337,0.999]	[0.341,0.944]
		(0.05,0.3)	0.526	[0.105,0.884]	[0.113,0.678]	0.622	[0.242,0.941]	[0.250,0.756]
	$X = 2$	(0.05,0.1)	0.781	[0.199,0.992]	[0.204,0.924]	0.841	[0.360,0.999]	[0.364,0.946]
		(0.05,0.3)	0.539	[0.125,0.892]	[0.136,0.690]	0.631	[0.257,0.951]	[0.267,0.764]
	$X = 3$	(0.05,0.1)	0.785	[0.229,0.996]	[0.234,0.924]	0.847	[0.378,0.998]	[0.384,0.951]
		(0.05,0.3)	0.548	[0.146,0.905]	[0.158,0.705]	0.640	[0.272,0.949]	[0.283,0.765]
$\gamma_1 = 1.5$	$X = 0$	(0.05,0.1)	0.791	[0.423,0.996]	[0.439,0.931]	0.868	[0.598,0.999]	[0.611,0.959]
		(0.05,0.3)	0.554	[0.253,0.906]	[0.290,0.709]	0.676	[0.436,0.964]	[0.463,0.793]
	$X = 1$	(0.05,0.1)	0.797	[0.461,0.996]	[0.478,0.930]	0.872	[0.608,0.999]	[0.622,0.958]
		(0.05,0.3)	0.562	[0.280,0.915]	[0.320,0.717]	0.682	[0.439,0.964]	[0.470,0.795]
	$X = 2$	(0.05,0.1)	0.802	[0.497,0.996]	[0.511,0.931]	0.877	[0.611,0.999]	[0.628,0.964]
		(0.05,0.3)	0.573	[0.308,0.917]	[0.349,0.727]	0.688	[0.441,0.966]	[0.475,0.799]
	$X = 3$	(0.05,0.1)	0.807	[0.528,0.996]	[0.543,0.934]	0.881	[0.611,0.999]	[0.629,0.964]
		(0.05,0.3)	0.580	[0.334,0.916]	[0.374,0.724]	0.697	[0.438,0.969]	[0.477,0.805]

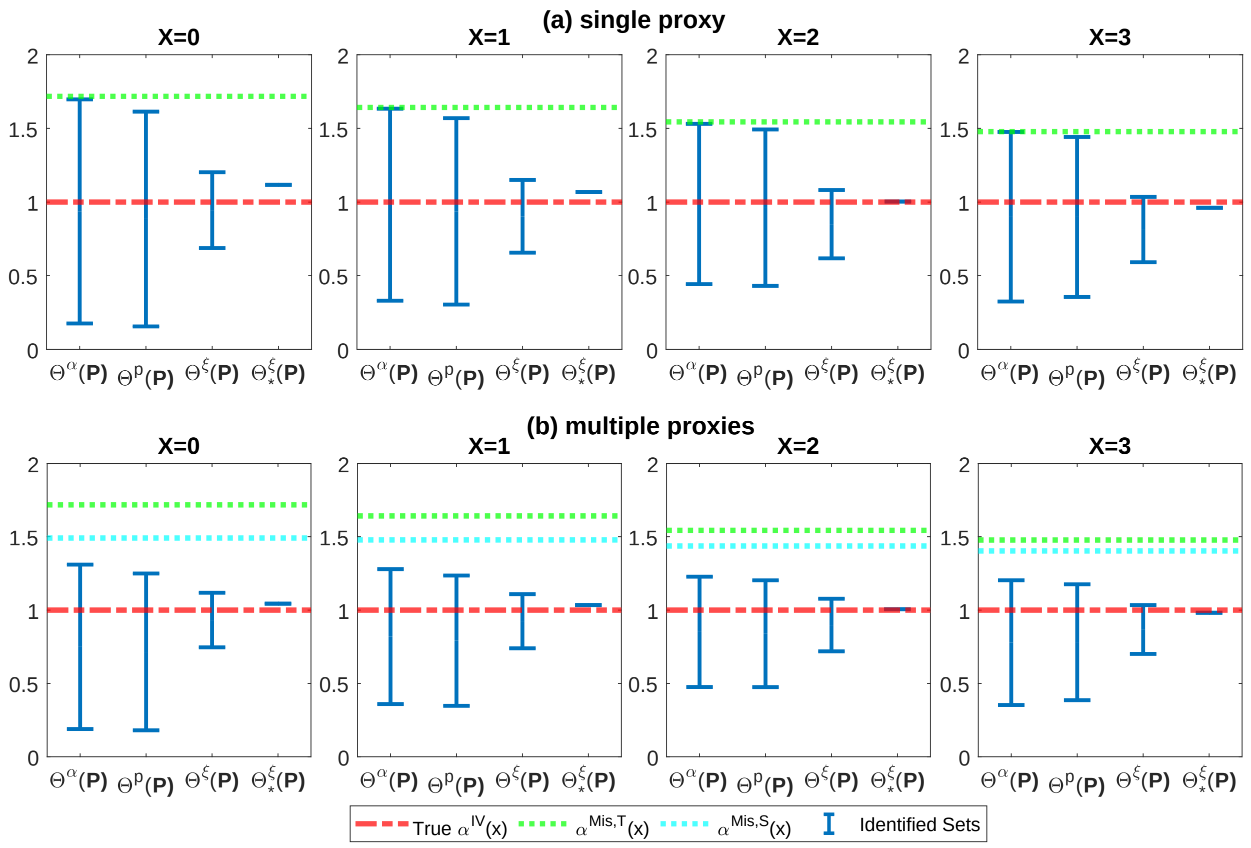
Notes: In panel (a), columns $\alpha_{k,k-1}$ $(k = 1,2)$ display the true values of LATEs. Columns $\Theta_k^\alpha(\mathbf{P})$ is the true identified set for LATE $\alpha_{k,k-1}$, "single" is when only T is available, and "multiple" is the identified set using both T and S. In panel (b), columns Δp_k^T $(k = 1,2)$ are the true values of the LATMs of proxy T. Columns $\Theta_k^{p^T}(\mathbf{P})$ $(k = 1,2)$ are the true identified set for Δp_k^T. Columns "single" display the identified sets using only T, and columns "multiple" report the identified sets under multiple proxies T, S, given in Lemma A.3 in Appendix.

Figure A1: True Identified Sets ($\gamma_1 = 1, \bar{\mu}_T = 0.1$)



Notes: The red, green and cyan lines are the true values of $\alpha^{IV}(x)$, $\alpha^{Mis,T}(x)$ and $\alpha^{Mis,S}(x)$ using T and S as proxy, respectively. The upper and lower bars of blue intervals are the ending points of the identified sets. In panel (a), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$, $\Theta^{\xi}(\mathbf{P})$ and $\Theta^{\xi}_{*}(\mathbf{P})$ report the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T as proxy. In panel (b), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are the intersection of the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T, S as proxies. In both panels, $\Theta^{\xi}(\mathbf{P})$ is constructed given $[\underline{\xi}^{W}(x), \bar{\xi}^{W}(x)]$, where $\underline{\xi}^{W}(x) = 1 - 2\overline{\mu}_W$, $\overline{\xi}^{W}(x) = 1 - \overline{\mu}_W$ for all $x \in \Omega_X$ and $W = \{T, S\}$. $\Theta^{\xi}_{*}(\mathbf{P})$ is a special case of the third identification strategy, where the possible range of $\xi^{W}(x)$ with $W = \{T, S\}$ is set to be a point $\underline{\xi}^{W}(x) = \bar{\xi}^{W}(x) = 1 - \overline{\mu}_W - \underline{\mu}_W$ for all $x \in \Omega_X$.

Figure A2: True Identified Sets ($\gamma_1 = 1, \bar{\mu}_T = 0.3$)



Notes: The red, green and cyan lines are the true values of $\alpha^{IV}(x)$, $\alpha^{Mis,T}(x)$ and $\alpha^{Mis,S}(x)$ using T and S as proxy, respectively. The upper and lower bars of blue intervals are the ending points of the identified sets. In panel (a), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$, $\Theta^{\xi}(\mathbf{P})$ and $\Theta^{\xi}_{*}(\mathbf{P})$ report the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T as proxy. In panel (b), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are the intersection of the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T, S as proxies. In both panels, $\Theta^{\xi}(\mathbf{P})$ is constructed given $[\underline{\xi}^{W}(x), \bar{\xi}^{W}(x)]$, where $\underline{\xi}^{W}(x) = 1 - 2\overline{\mu}_W$, $\overline{\xi}^{W}(x) = 1 - \overline{\mu}_W$ for all $x \in \Omega_X$ and $W = \{T, S\}$. $\Theta^{\xi}_{*}(\mathbf{P})$ is a special case of the third identification strategy, where the possible range of $\xi^{W}(x)$ with $W = \{T, S\}$ is set to be a point $\underline{\xi}^{W}(x) = \bar{\xi}^{W}(x) = 1 - \overline{\mu}_W - \underline{\mu}_W$ for all $x \in \Omega_X$.

Next, we study the finite sample properties of the confidence intervals $C^j(\beta^j)$ with $j = \alpha, p, \xi$ proposed in Section 4. To focus on their performance in practical applications, we compute the simplified version of the confidence intervals of $\alpha_{k,k-1}$ and Δp_k with $k = 1, 2$ as in Equation (23). Based on this, the confidence intervals of α^{IV} are constructed in the same manners as in Equations (24), (25) and (26). Monte Carlo simulations are implemented with sample size $n = 8,000$ and 1,000 replications. We choose the number of partition $\kappa_n = 4$, size $\eta = 0.05$ and $\eta_{\alpha^{Mis}} = 0.01$ (for $C^p(\beta^p)$) and $\eta_{\alpha^{Mis}} = 0.05$ (for $C^\xi(\beta^\xi)$). We calculate the coverage rates as how often the confidence interval includes a given parameter value out of 1,000 simulations. The critical values used to build the confidence intervals are obtained by two-step multiplier bootstrap with size $\beta = 0.1\%$ for the moment selection and 500 bootstrap repetitions.[41]

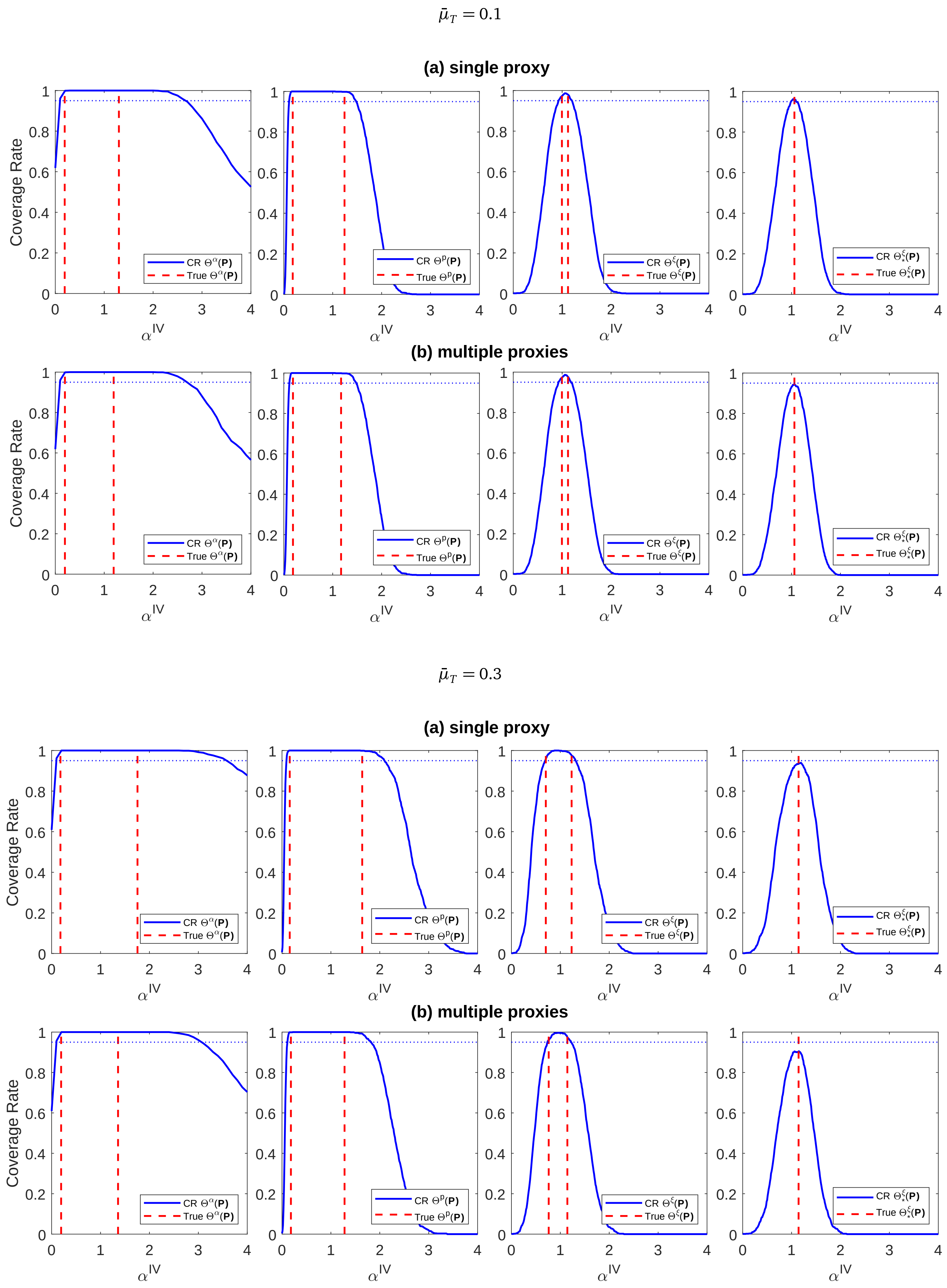
Figure A3 plots the coverage rates of the confidence intervals associated with $\overline{\mu}_T = 0.1$ and $\gamma_1 = 1$, for $X = 0$. The complete set of results is reported below.[42] One can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%, while $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. This is intuitive because the second strategy utilizes the information from both α^{Mis} and the identified sets of $\{\Delta p_k\}_{k-1}^K$. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available and/or when instrument strength becomes stronger.

Our conclusion is that P-LATE represents a reliable alternative estimator when practitioners can only use a mismeasured binary treatment T in place of D to estimate the benefits of a program. Moreover, P-LATE becomes very powerful, and works at best, when external information about the accuracy of the measurement error can be taken into account.

[41] See the two-step multiplier bootstrap in Appendix A.2 for the details about the moment selection and its size β.

[42] Here we provide further robustness checks associated to $\overline{\mu}_T \in \{0.1, 0.3\}$ and $\gamma_1 \in \{1, 1.5\}$, and for all stratification of the covariate X. The qualitative results do not change.

Figure A3: Coverage Rate of Confidence Intervals ($\gamma_1 = 1, \underline{\mu}_T = 0.05, X = 0$)



$\bar{\mu}_T = 0.1$

$\bar{\mu}_T = 0.3$

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1$, $\underline{\mu}_T = 0.05$ and stratification $X = 0$. We vary $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Table A2: Identified Set of α^{IV} under Endogenous Misclassification ($\alpha^{IV} = 1$)

(a) Single proxy

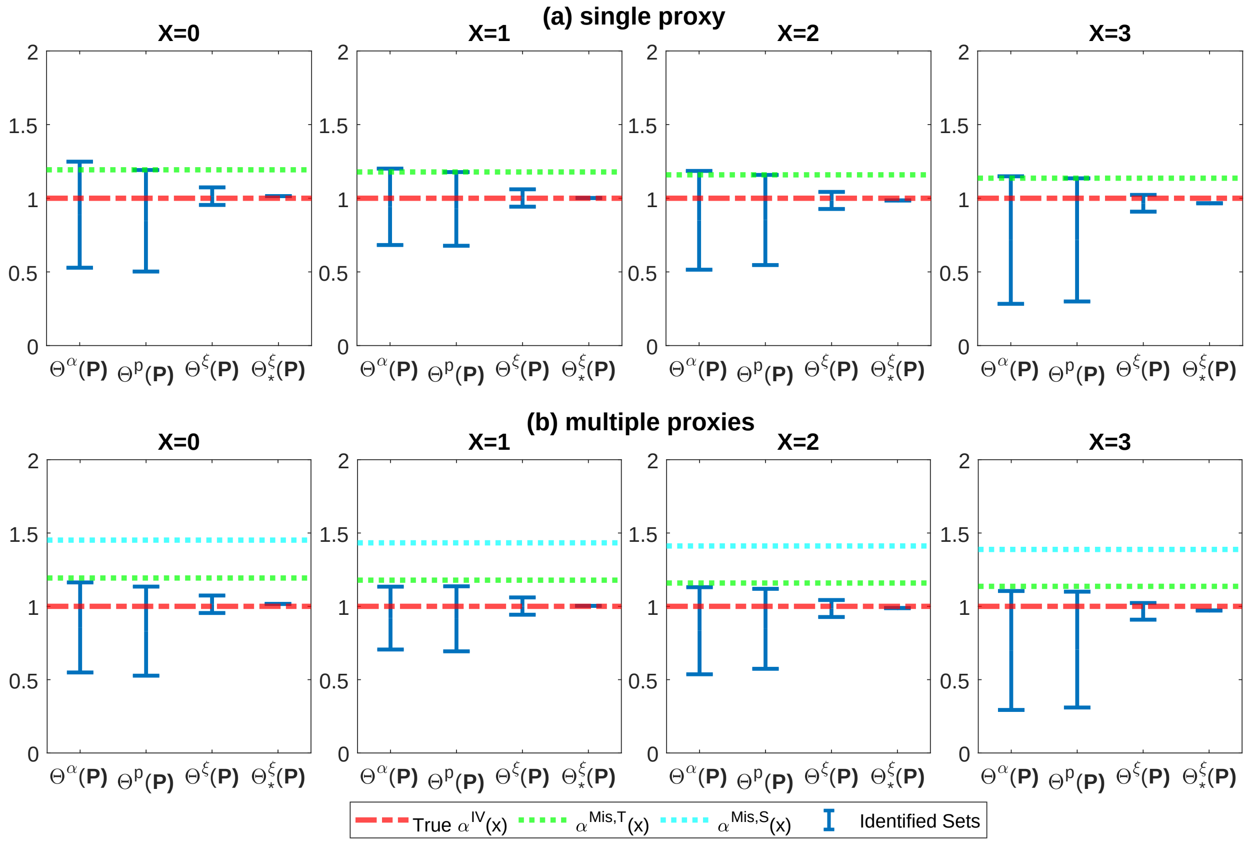
		$(\underline{\mu}_T, \overline{\mu}_T)$	$\alpha^{Mis,T}$	$\Theta^{\alpha}(\mathbf{P})$	$\Theta^{p}(\mathbf{P})$	$\Theta^{\xi}(\mathbf{P})$	$\Theta^{\xi}_{*}(\mathbf{P})$
$\gamma_1 = 1$	$X = 0$	(0.05,0.1)	1.247	[0.194,1.320]	[0.182,1.243]	[0.997,1.122]	1.060
		(0.05,0.3)	1.759	[0.180,1.755]	[0.159,1.640]	[0.703,1.231]	1.143
	$X = 1$	(0.05,0.1)	1.235	[0.220,1.276]	[0.212,1.234]	[0.988,1.112]	1.050
		(0.05,0.3)	1.724	[0.201,1.697]	[0.180,1.623]	[0.690,1.207]	1.121
	$X = 2$	(0.05,0.1)	1.221	[0.255,1.272]	[0.243,1.220]	[0.977,1.099]	1.038
		(0.05,0.3)	1.681	[0.235,1.672]	[0.210,1.598]	[0.672,1.176]	1.092
	$X = 3$	(0.05,0.1)	1.209	[0.289,1.259]	[0.277,1.207]	[0.967,1.088]	1.028
		(0.05,0.3)	1.654	[0.263,1.633]	[0.241,1.569]	[0.662,1.158]	1.075
$\gamma_1 = 1.5$	$X = 0$	(0.05,0.1)	1.190	[0.528,1.241]	[0.504,1.189]	[0.952,1.071]	1.011
		(0.05,0.3)	1.597	[0.455,1.630]	[0.404,1.540]	[0.639,1.118]	1.038
	$X = 1$	(0.05,0.1)	1.195	[0.577,1.245]	[0.551,1.195]	[0.956,1.076]	1.016
		(0.05,0.3)	1.595	[0.497,1.622]	[0.447,1.538]	[0.638,1.116]	1.036
	$X = 2$	(0.05,0.1)	1.186	[0.614,1.232]	[0.589,1.184]	[0.949,1.067]	1.008
		(0.05,0.3)	1.581	[0.537,1.598]	[0.488,1.528]	[0.633,1.107]	1.028
	$X = 3$	(0.05,0.1)	1.189	[0.653,1.232]	[0.628,1.188]	[0.951,1.070]	1.010
		(0.05,0.3)	1.583	[0.582,1.596]	[0.529,1.535]	[0.633,1.108]	1.029

(b) Multiple proxies

		$(\underline{\mu}_T, \overline{\mu}_T)$	$\alpha^{Mis,T}$	$\alpha^{Mis,S}$	$\Theta^{\alpha}(\mathbf{P})$	$\Theta^{p}(\mathbf{P})$	$\Theta^{\xi}(\mathbf{P})$	$\Theta^{\xi,T}_{*}(\mathbf{P})$	$\Theta^{\xi,S}_{*}(\mathbf{P})$
$\gamma_1 = 1$	$X = 0$	(0.05,0.1)	1.247	1.510	[0.199,1.217]	[0.187,1.173]	[0.997,1.122]	1.060	1.057
		(0.05,0.3)	1.759		[0.187,1.299]	[0.180,1.255]	[0.755,1.133]	1.143	
	$X = 1$	(0.05,0.1)	1.235	1.500	[0.226,1.175]	[0.218,1.163]	[0.988,1.112]	1.050	1.050
		(0.05,0.3)	1.724		[0.217,1.301]	[0.211,1.263]	[0.750,1.125]	1.121	
	$X = 2$	(0.05,0.1)	1.221	1.479	[0.262,1.184]	[0.249,1.154]	[0.977,1.099]	1.038	1.035
		(0.05,0.3)	1.681		[0.254,1.294]	[0.243,1.237]	[0.740,1.111]	1.092	
	$X = 3$	(0.05,0.1)	1.209	1.462	[0.296,1.168]	[0.287,1.140]	[0.967,1.088]	1.028	1.024
		(0.05,0.3)	1.654		[0.286,1.271]	[0.275,1.226]	[0.731,1.097]	1.075	
$\gamma_1 = 1.5$	$X = 0$	(0.05,0.1)	1.190	1.450	[0.547,1.160]	[0.523,1.141]	[0.952,1.071]	1.011	1.015
		(0.05,0.3)	1.597		[0.522,1.275]	[0.498,1.216]	[0.725,1.088]	1.038	
	$X = 1$	(0.05,0.1)	1.195	1.453	[0.598,1.164]	[0.573,1.143]	[0.956,1.076]	1.016	1.017
		(0.05,0.3)	1.595		[0.567,1.271]	[0.544,1.219]	[0.728,1.091]	1.036	
	$X = 2$	(0.05,0.1)	1.186	1.442	[0.632,1.151]	[0.613,1.132]	[0.949,1.067]	1.008	1.010
		(0.05,0.3)	1.581		[0.608,1.268]	[0.584,1.218]	[0.726,1.089]	1.028	
	$X = 3$	(0.05,0.1)	1.189	1.449	[0.671,1.155]	[0.650,1.139]	[0.951,1.070]	1.010	1.015
		(0.05,0.3)	1.583		[0.652,1.263]	[0.627,1.214]	[0.724,1.086]	1.029	

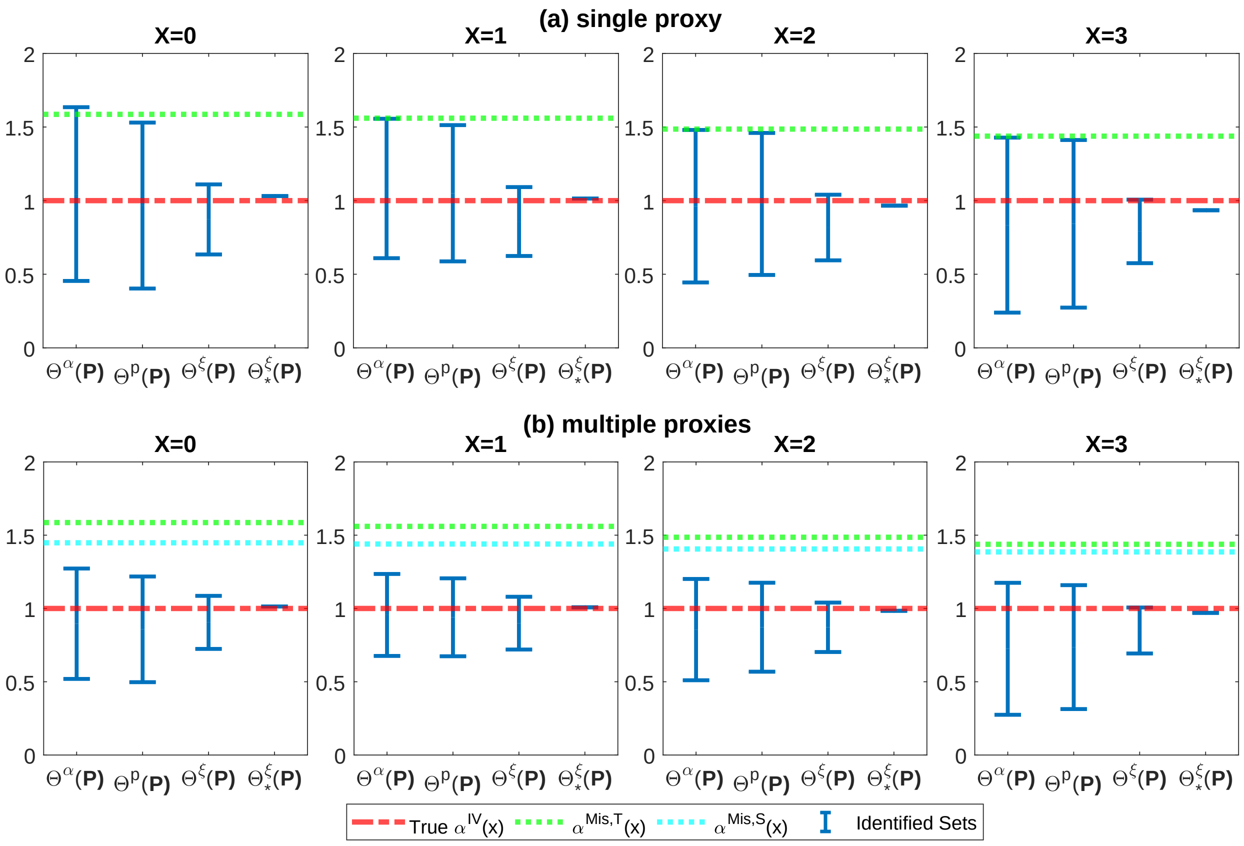
Notes: Columns $\alpha^{Mis,T}$ and $\alpha^{Mis,S}$ are true values of $\alpha^{Mis}(x)$ using T and S as proxy, respectively. In panel (a), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$, $\Theta^{\xi}(\mathbf{P})$ and $\Theta^{\xi}_{*}(\mathbf{P})$ report the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T as proxy. In panel (b), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are the intersection of the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T, S as proxies. In both panels, $\Theta^{\xi}(\mathbf{P})$ is constructed given $[\underline{\xi}^W(x), \overline{\xi}^W(x)]$, where $\underline{\xi}^W(x) = 1 - 2\overline{\mu}_W$, $\overline{\xi}^W(x) = 1 - \overline{\mu}_W$ for all $x \in \Omega_X$ and $W = \{T, S\}$. $\Theta^{\xi}_{*}(\mathbf{P})$ is a special case of the third identification strategy, where the possible range of $\xi^W(x)$ with $W = \{T, S\}$ is set to be a point $\underline{\xi}^W(x) = \bar{\xi}^W(x) = 1 - \overline{\mu}_W - \underline{\mu}_W$ for all $x \in \Omega_X$.

Figure A4: True Identified Sets ($\gamma_1 = 1.5, \bar{\mu}_T = 0.1$)



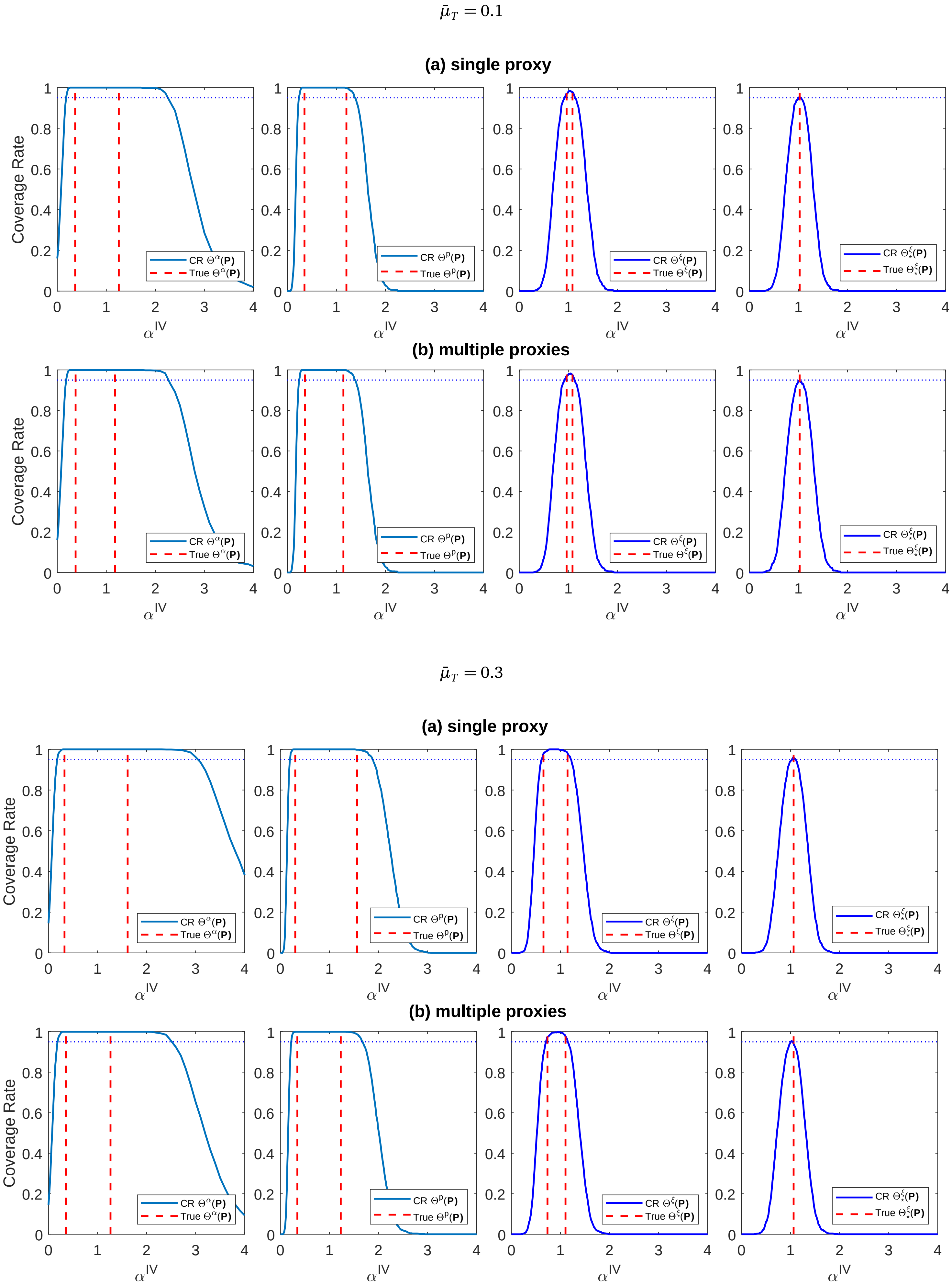
Notes: The red, green and cyan lines are the true values of $\alpha^{IV}(x)$, $\alpha^{Mis,T}(x)$ and $\alpha^{Mis,S}(x)$ using T and S as proxy, respectively. The upper and lower bars of blue intervals are the ending points of the identified sets. In panel (a), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$, $\Theta^{\xi}(\mathbf{P})$ and $\Theta^{\xi}_{*}(\mathbf{P})$ report the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T as proxy. In panel (b), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are the intersection of the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T, S as proxies. In both panels, $\Theta^{\xi}(\mathbf{P})$ is constructed given $[\underline{\xi}^{W}(x), \bar{\xi}^{W}(x)]$, where $\underline{\xi}^{W}(x) = 1 - 2\overline{\mu}_W$, $\overline{\xi}^{W}(x) = 1 - \overline{\mu}_W$ for all $x \in \Omega_X$ and $W = \{T, S\}$. $\Theta^{\xi}_{*}(\mathbf{P})$ is a special case of the third identification strategy, where the possible range of $\xi^{W}(x)$ with $W = \{T, S\}$ is set to be a point $\underline{\xi}^{W}(x) = \bar{\xi}^{W}(x) = 1 - \overline{\mu}_W - \underline{\mu}_W$ for all $x \in \Omega_X$.

Figure A5: True Identified Sets ($\gamma_1 = 1.5, \bar{\mu}_T = 0.3$)



Notes: The red, green and cyan lines are the true values of $\alpha^{IV}(x)$, $\alpha^{Mis,T}(x)$ and $\alpha^{Mis,S}(x)$ using T and S as proxy, respectively. The upper and lower bars of blue intervals are the ending points of the identified sets. In panel (a), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$, $\Theta^{\xi}(\mathbf{P})$ and $\Theta^{\xi}_{*}(\mathbf{P})$ report the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T as proxy. In panel (b), $\Theta^{\alpha}(\mathbf{P})$, $\Theta^{p}(\mathbf{P})$ and $\Theta^{\xi}(\mathbf{P})$ are the intersection of the true identified sets of $\alpha^{IV}(x)$, conditional on $X = x$, using T, S as proxies. In both panels, $\Theta^{\xi}(\mathbf{P})$ is constructed given $[\underline{\xi}^{W}(x), \bar{\xi}^{W}(x)]$, where $\underline{\xi}^{W}(x) = 1 - 2\overline{\mu}_{W}$, $\overline{\xi}^{W}(x) = 1 - \overline{\mu}_{W}$ for all $x \in \Omega_X$ and $W = \{T, S\}$. $\Theta^{\xi}_{*}(\mathbf{P})$ is a special case of the third identification strategy, where the possible range of $\xi^{W}(x)$ with $W = \{T, S\}$ is set to be a point $\underline{\xi}^{W}(x) = \bar{\xi}^{W}(x) = 1 - \overline{\mu}_{W} - \underline{\mu}_{W}$ for all $x \in \Omega_X$.

Figure A6: Coverage Rate of Confidence Intervals ($\gamma_1 = 1, \underline{\mu}_T = 0.05, X = 1$)



$\bar{\mu}_T = 0.1$

(a) single proxy

(b) multiple proxies

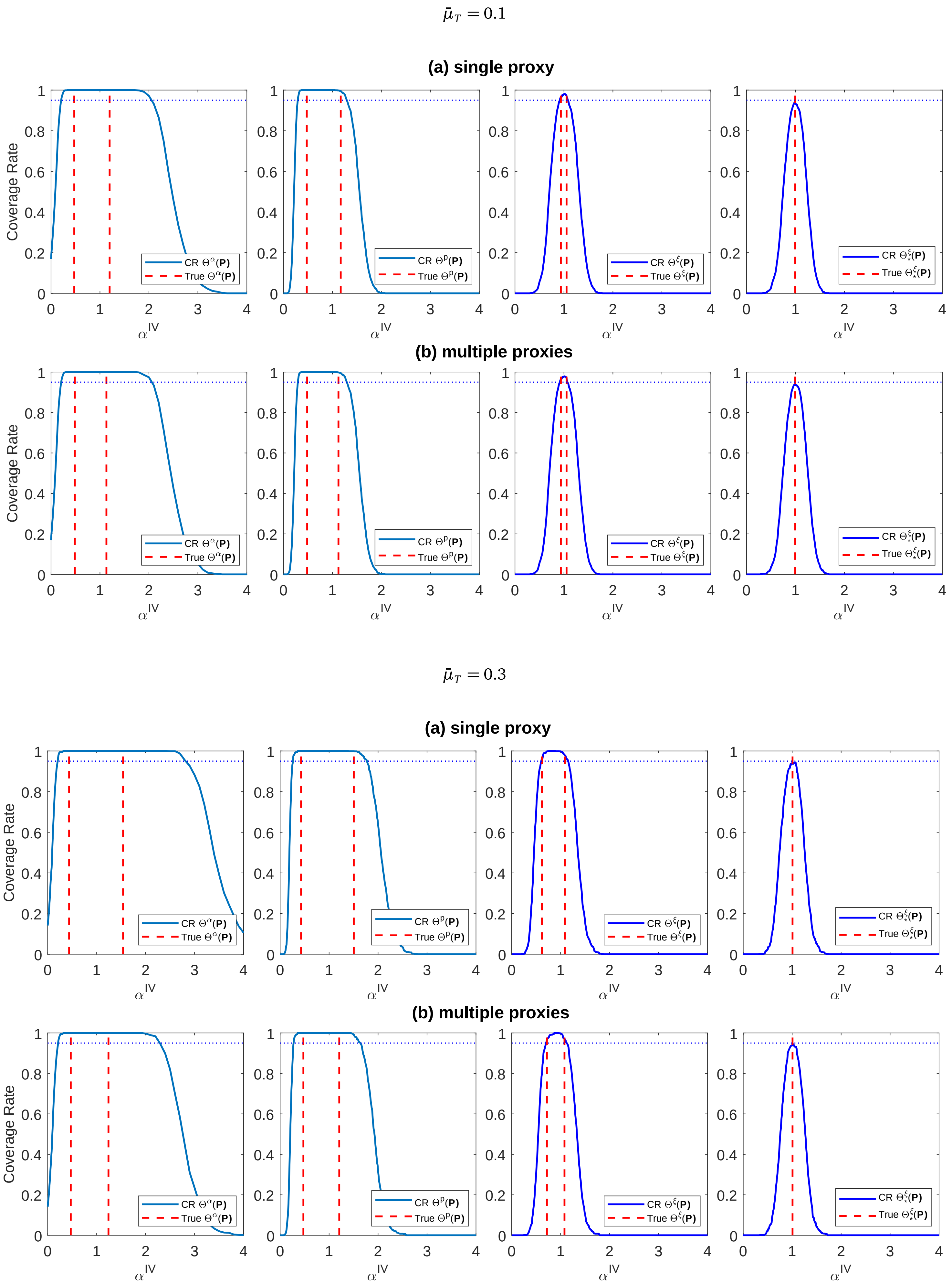
$\bar{\mu}_T = 0.3$

(a) single proxy

(b) multiple proxies

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1$, $\underline{\mu}_T = 0.05$ and stratification $X = 1$. We vary $\bar{\mu}_T = 0.1$ and $\bar{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^{\alpha}(\beta^{\alpha})$ in the sense that, compared to $\mathscr{C}^{\alpha}(\beta^{\alpha})$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^{\xi}(\beta^{\xi})$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A7: Coverage Rate of Confidence Intervals ($\gamma_1 = 1, \underline{\mu}_T = 0.05, X = 2$)



$\bar{\mu}_T = 0.1$

(a) single proxy

(b) multiple proxies

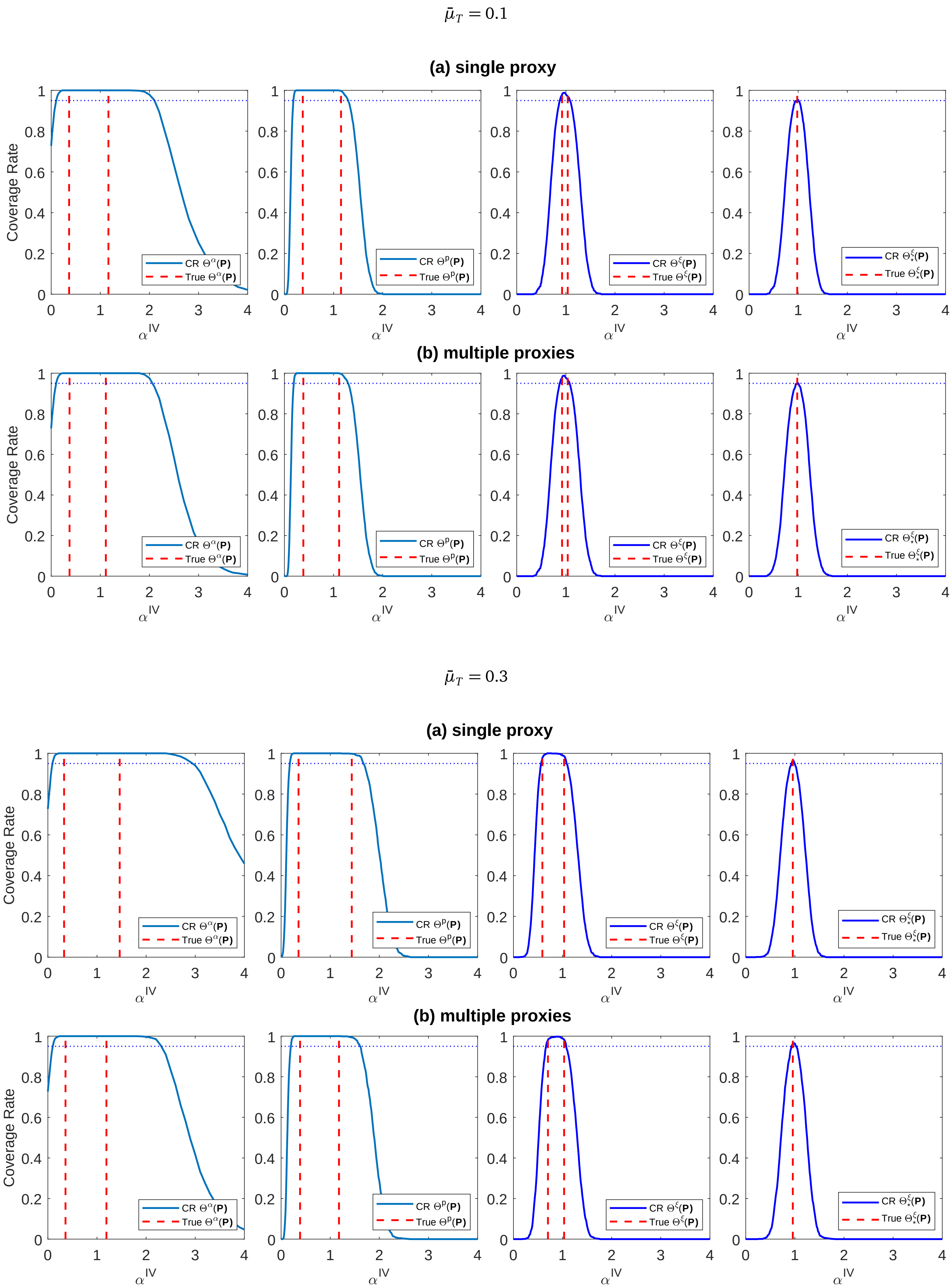
$\bar{\mu}_T = 0.3$

(a) single proxy

(b) multiple proxies

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1$, $\underline{\mu}_T = 0.05$ and stratification $X = 2$. We vary $\bar{\mu}_T = 0.1$ and $\bar{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \bar{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A8: Coverage Rate of Confidence Intervals ($\gamma_1 = 1, \underline{\mu}_T = 0.05, X = 3$)


$\bar{\mu}_T = 0.1$

(a) single proxy

(b) multiple proxies

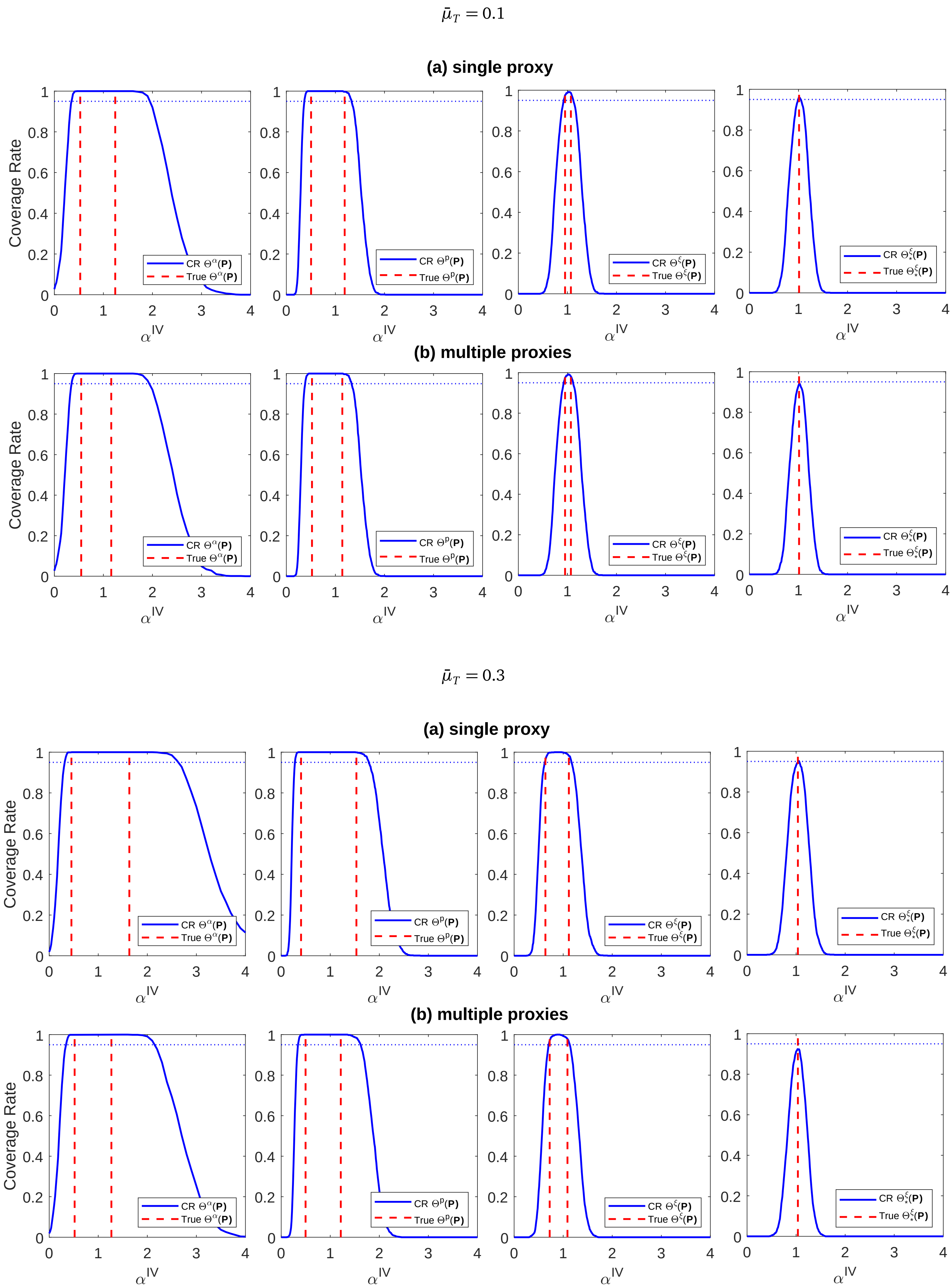
$\bar{\mu}_T = 0.3$

(a) single proxy

(b) multiple proxies

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1$, $\underline{\mu}_T = 0.05$ and stratification $X = 3$. We vary $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A9: Coverage Rate of Confidence Intervals ($\gamma_1 = 1.5, \underline{\mu}_T = 0.05, X = 0$)



$\bar{\mu}_T = 0.1$

(a) single proxy

(b) multiple proxies

$\bar{\mu}_T = 0.3$

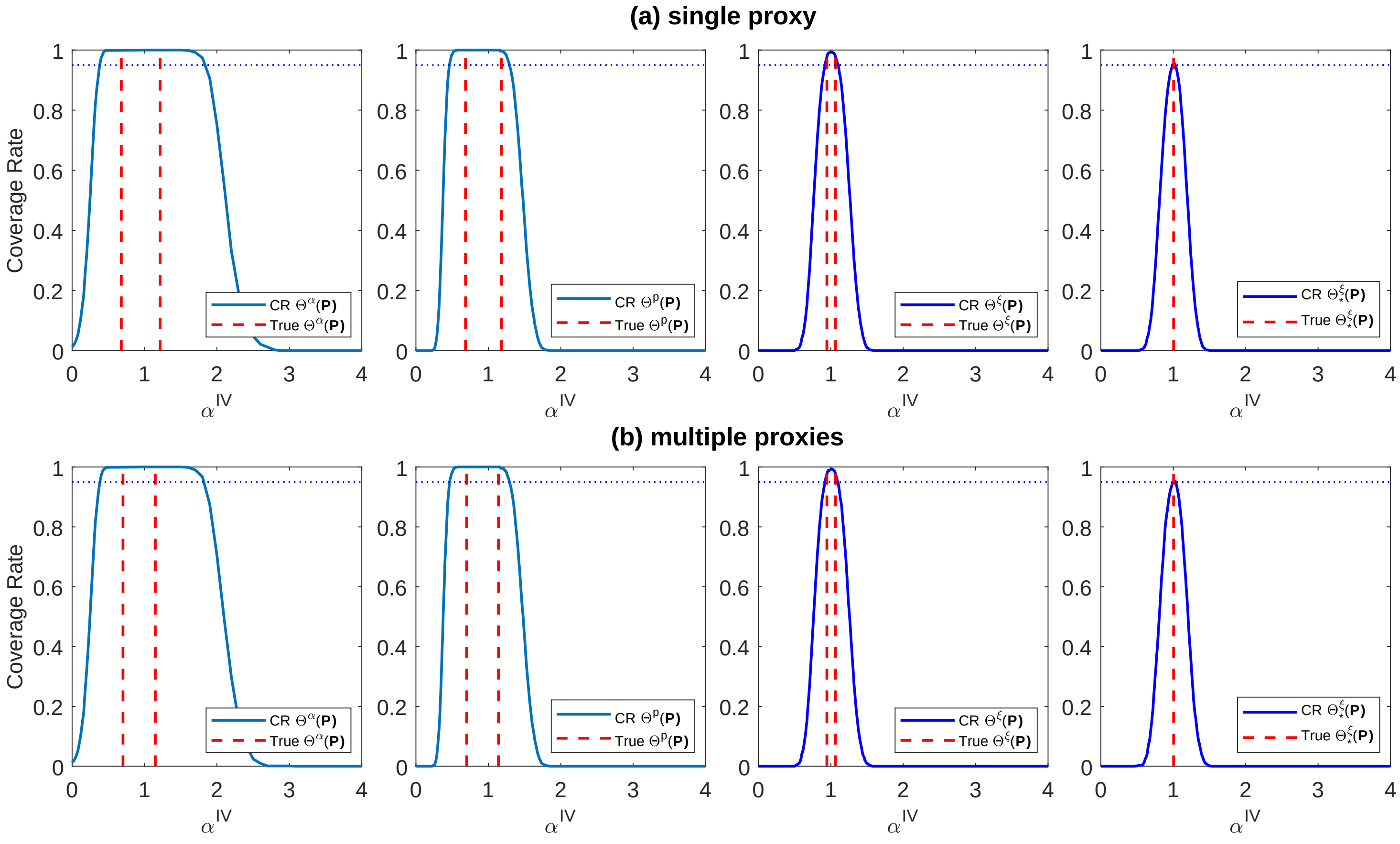
(a) single proxy

(b) multiple proxies

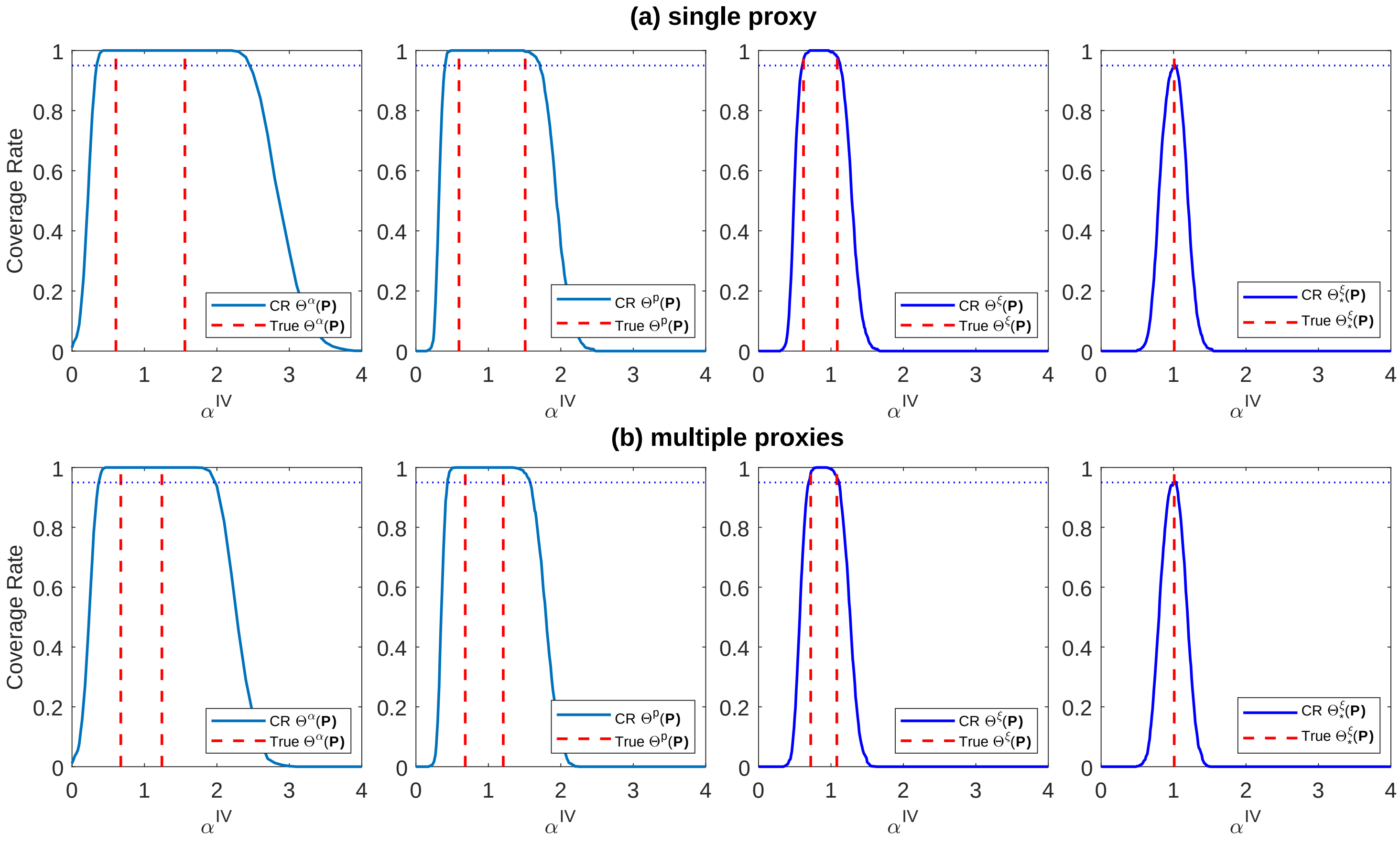
Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1.5$, $\underline{\mu}_T = 0.05$ and stratification $X = 0$. We vary $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A10: Coverage Rate of Confidence Intervals ($\gamma_1 = 1.5, \underline{\mu}_T = 0.05, X = 1$)

$\bar{\mu}_T = 0.1$



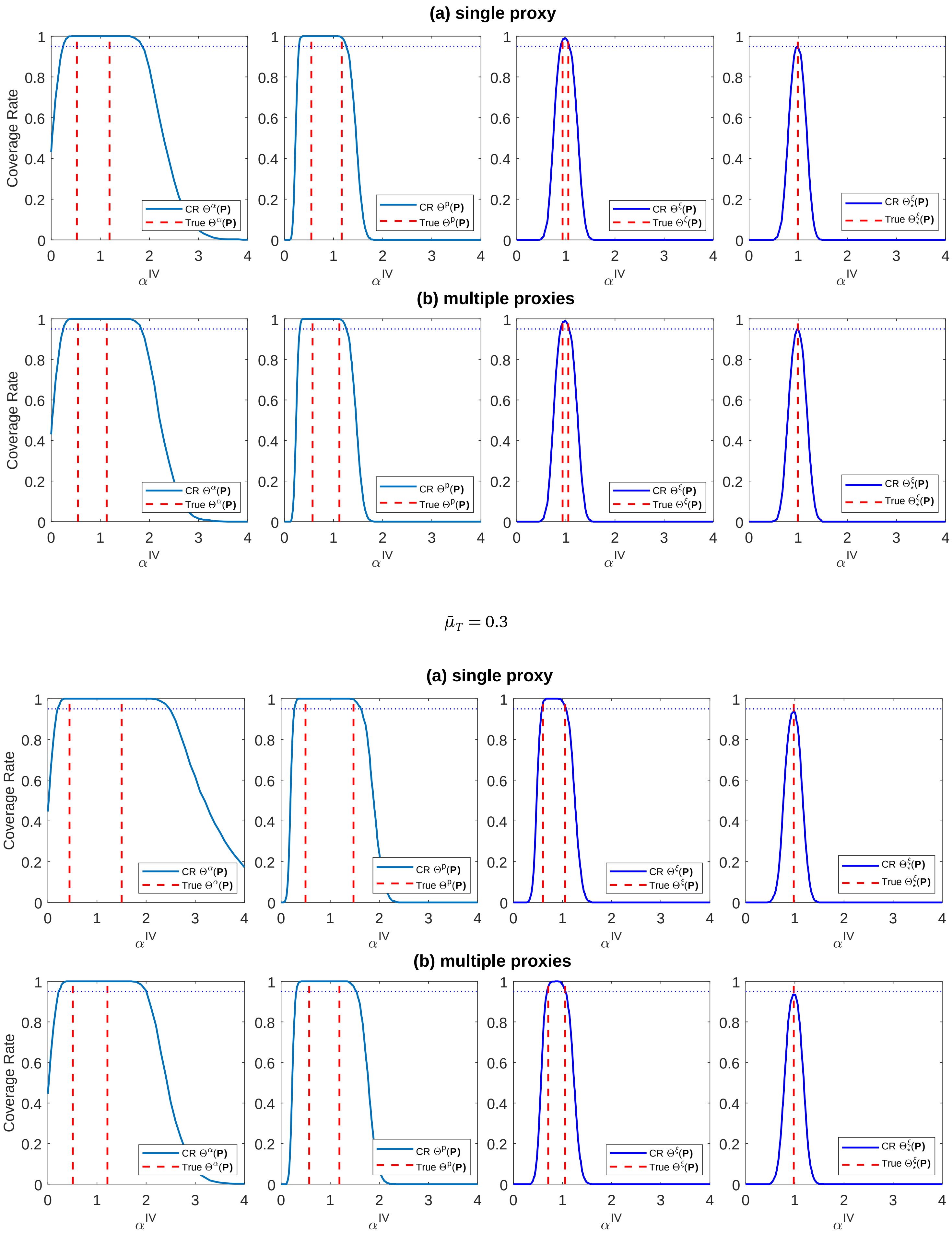
$\bar{\mu}_T = 0.3$



Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1.5$, $\underline{\mu}_T = 0.05$ and stratification $X = 1$. We vary $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A11: Coverage Rate of Confidence Intervals ($\gamma_1 = 1.5, \underline{\mu}_T = 0.05, X = 2$)

$\bar{\mu}_T = 0.1$



(a) single proxy

(b) multiple proxies

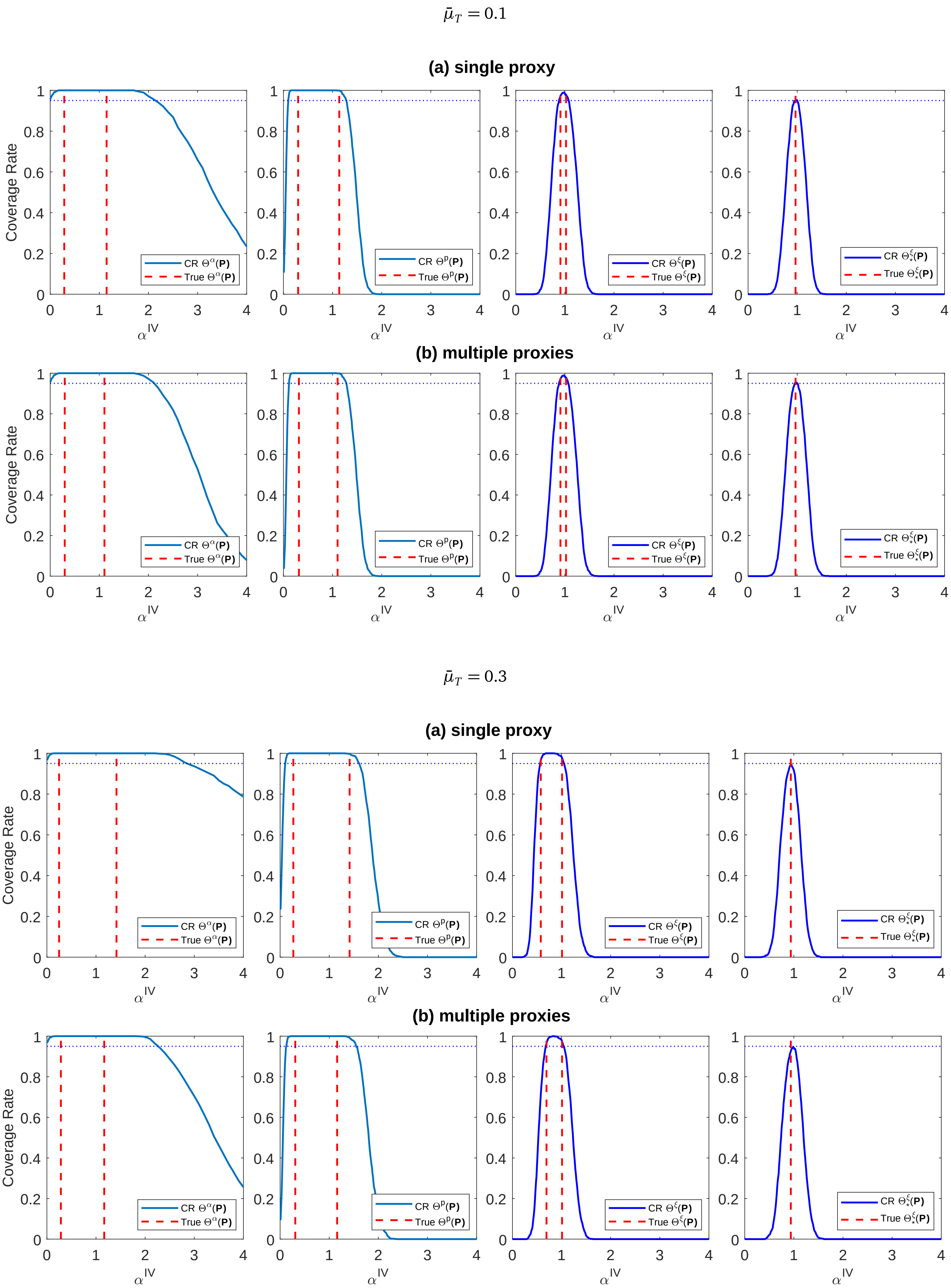
$\bar{\mu}_T = 0.3$

(a) single proxy

(b) multiple proxies

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1.5$, $\underline{\mu}_T = 0.05$ and stratification $X = 2$. We vary $\bar{\mu}_T = 0.1$ and $\bar{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^{\alpha}(\beta^{\alpha})$ in the sense that, compared to $\mathscr{C}^{\alpha}(\beta^{\alpha})$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \bar{\xi}^T]$ is given, $\mathscr{C}^{\xi}(\beta^{\xi})$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.

Figure A12: Coverage Rate of Confidence Intervals ($\gamma_1 = 1.5, \underline{\mu}_T = 0.05, X = 3$)



$\bar{\mu}_T = 0.1$

$\bar{\mu}_T = 0.3$

Notes: This figure plots the coverage rates of the confidence intervals associated with $\gamma_1 = 1.5$, $\underline{\mu}_T = 0.05$ and stratification $X = 3$. We vary $\overline{\mu}_T = 0.1$ and $\overline{\mu}_T = 0.3$. The blue curve is the coverage rate of the confidence intervals, and the red vertical dashed-lines are the ending points of true identified sets. The blue horizontal dotted-line is the 95%. We can see that all the confidence intervals cover their corresponding true identified sets with probability at least 95%. $\mathscr{C}^p(\beta^p)$ significantly outperforms $\mathscr{C}^\alpha(\beta^\alpha)$ in the sense that, compared to $\mathscr{C}^\alpha(\beta^\alpha)$, the coverage rate of $\mathscr{C}^p(\beta^p)$ drops dramatically faster for those parameter values outside the identified set. In addition, when extra information of the range $[\underline{\xi}^T, \overline{\xi}^T]$ is given, $\mathscr{C}^\xi(\beta^\xi)$ gives the least conservative 95% confidence interval which includes the true value of α^{IV}. Besides, all confidence intervals become less conservative when multiple proxies are available.