Heterogeneous treatment effects with mismeasured endogenous treatment

TAKUYA URA
Department of Economics, University of California, Davis

This paper studies the identifying power of an instrumental variable in the non-parametric heterogeneous treatment effect framework when a binary treatment is mismeasured and endogenous. Using a binary instrumental variable, I characterize the sharp identified set for the local average treatment effect under the exclusion restriction of an instrument and the deterministic monotonicity of the true treatment in the instrument. Even allowing for general measurement error (e.g., the measurement error is endogenous), it is still possible to obtain finite bounds on the local average treatment effect. Notably, the Wald estimand is an upper bound on the local average treatment effect, but it is not the sharp bound in general. I also provide a confidence interval for the local average treatment effect with uniformly asymptotically valid size control. Furthermore, I demonstrate that the identification strategy of this paper offers a new use of repeated measurements for tightening the identified set.

KEYWORDS. Local average treatment effect, instrumental variable, nonclassical measurement error, endogenous measurement error, partial identification.

JEL CLASSIFICATION. C21, C26.

1. INTRODUCTION

Treatment effect analyses often entail a measurement error problem as well as an endogeneity problem. For example, Black, Sanders, and Taylor (2003) documented a substantial measurement error in educational attainments in the 1990 U.S. Census. At the same time, educational attainments are endogenous as treatment variables in return-to-schooling analyses, because unobserved individual ability affects both schooling decisions and wages (Card (2001)). The econometric literature, however, has offered only a few solutions for addressing the two problems at the same time. An instrumental variable is a standard technique for correcting endogeneity and measurement error (e.g., Angrist and Krueger (2001)). To the best of my knowledge, however, no existing research has explicitly investigated the identifying power of an instrumental variable for

the heterogeneous treatment effect when the treatment is both mismeasured and en-
dogenous.[1]

I consider a mismeasured treatment in the framework of Imbens and Angrist (1994)
and Angrist, Imbens, and Rubin (1996), and focus on the local average treatment ef-
fect as a parameter of interest. My analysis studies the identifying power of a binary in-
strumental variable under the following two assumptions: (i) the instrument affects the
outcome and the measured treatment only through the true treatment (the exclusion
restriction of an instrument), and (ii) the instrument weakly increases the true treat-
ment (the deterministic monotonicity of the true treatment in the instrument). These
assumptions are an extension of Imbens and Angrist (1994) and Angrist, Imbens, and
Rubin (1996) into the framework with mismeasured treatment. The local average treat-
ment effect is the average treatment effect for the compliers, that is, the subpopulation
whose true treatment status is strictly affected by an instrument. Focusing on the local
average treatment effect is meaningful for a few reasons. First, the local average treat-
ment effect has been a widely-used parameter to investigate the heterogeneous treat-
ment effect with endogeneity. My analysis offers a tool for a robustness check to those
who have already investigated the local average treatment effect. Second, the local aver-
age treatment effect can be used to extrapolate to the average treatment effect or other
parameters of interest. Imbens (2010) emphasizes the utility of reporting the local aver-
age treatment effect in addition to the other parameters of interest because the extrap-
olation often requires additional assumptions and can be less credible than the local
average treatment effect.

The mismeasured treatment prevents the local average treatment effect from being
point-identified. As in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996),
the local average treatment effect is the ratio of the intent-to-treat effect over the size of
compliers.[2] Since the measured treatment is not the true treatment, however, the size of
compliers is not identified and, therefore, the local average treatment effect is not iden-
tified. The underidentification for the local average treatment effect is a consequence of
the underidentification for the size of compliers; if I assumed no measurement error, I
could compute the size of compliers based on the measured treatment and, therefore,
the local average treatment effect would be the Wald estimand.[3]

I take a worst case scenario approach against the measurement error and allow for
a general form of measurement error. The only assumption concerning the measure-
ment error in this paper is its independence of the instrumental variable. (Section 3.4

---

[1]Many existing methods, including Mahajan (2006) and Lewbel (2007), allow for the treatment effect to
be heterogeneous due to observed variables. In this paper, I focus on the heterogeneity due to unobserved
variables by considering the local average treatment effect framework.

[2]The intent-to-treat effect is defined as the mean difference of the outcome between the two groups
defined by the instrument. The size of compliers is the probability of being a complier, and it is the mean
difference of the true treatment (Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996)).

[3]The Wald estimand in this paper is defined as the ratio of the intent-to-treat effect over the mean dif-
ference of the measured treatment between the two groups defined by the instrument. Note that the Wald
estimand is identified because it uses the measured treatment, but it is not the local average treatment
effect because it does not use the true treatment.

(a) When the intent-to-treat effect is positive

(b) When the intent-to-treat effect is zero

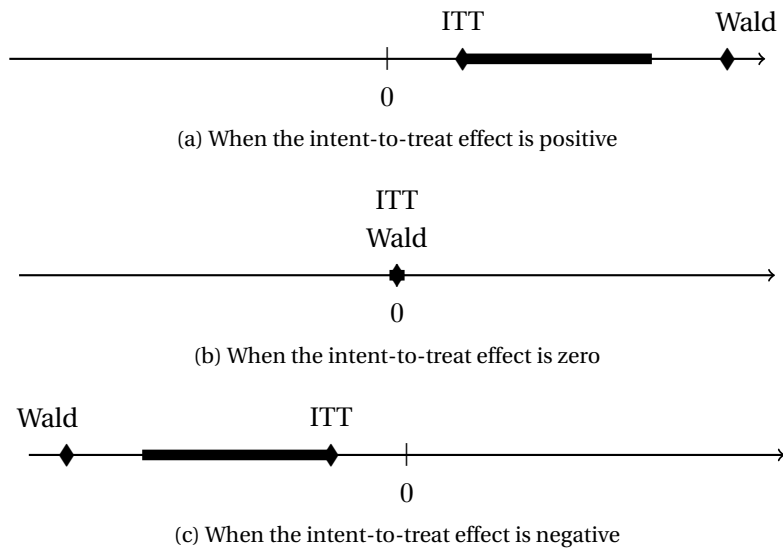(c) When the intent-to-treat effect is negative

FIGURE 1. Identified set for the local average treatment effect. ITT is the intent-to-treat effect and Wald is the Wald estimand. The thick line is the identified set for the local average treatment effect. Note that the identified set is {0} when the intent-to-treat effect is zero.

dispenses with this assumption and shows that it is still possible to bound the local average treatment effect.) The following properties of measurement error are practically and theoretically relevant. First, the measurement error is nonclassical; that is, it can be dependent on the true treatment. The measurement error for a discrete variable is always nonclassical. It is because the measurement error cannot be negative (positive) when the true variable takes the lowest (highest) value. Second, I allow the measurement error to be endogenous (or differential); that is, the measured treatment can be dependent on the outcome conditional on the true treatment. For example, as Black, Sanders, and Taylor (2003) argued, the measurement error for educational attainment depends on the familiarity with the educational system in the U.S., and immigrants may have a higher rate of measurement error. At the same time, the familiarity with the U.S. educational system can be related to the English language skills, which can affect the labor market outcomes. Bound, Brown, and Mathiowetz (2001) also argue that measurement error is likely to be endogenous in some empirical applications. (In Appendix C in the Supplementary Material (Ura (2018)), I explore the identifying power of the exogeneity assumption on the measurement error. The additional assumption yields a tighter sharp identified set, but I still cannot point identify the local average treatment effect in general.) Third, there is no assumption concerning the marginal distribution of the measurement error. It is not necessary to assume anything about the accuracy of the measurement.

    In the presence of measurement error, I derive the identified set for the local average treatment effect (Theorem 4). Figure 1 describes the relationship among the identified set for the local average treatment effect, the intent-to-treat effect, and the Wald estimand. First, the intent-to-treat effect has the same sign as the local average treatment

effect. Figure 1 has three subfigures according to the sign of the intent-to-treat effect: (a) positive, (b) zero, and (c) negative. Second, the intent-to-treat effect is the sharp lower bound on the local average treatment effect in absolute value. Third, the Wald estimand is an upper bound on the local average treatment effect in absolute value. The Wald estimand is the probability limit of the instrumental variable estimator in my framework, which ignores the measurement error but controls only for the endogeneity. This point implies that an upper bound on the local average treatment effect is obtained by ignoring the measurement error. Frazis and Loewenstein (2003) obtain a similar result in the homogeneous treatment effect model. Last, but most importantly, the sharp upper bound in absolute value can be smaller than the Wald estimand. It is a potential cost of ignoring the measurement error and using the Wald estimand. Even for analyzing only an upper bound on the local average treatment effect, it is recommended to take the measurement error into account, which can yield a smaller upper bound than the Wald estimand. Section 3.1 investigates when the Wald estimand coincides with the sharp upper bound.

I extend the identification analysis to incorporate covariates other than the treatment variable. In this setting, the instrumental variable satisfies the exclusion restriction after conditioning covariates. Based on the insights from Abadie (2003) and Frölich (2007), I show that the identification strategy of this paper works in the presence of covariates.

I construct a confidence interval for the local average treatment effect. To construct the confidence interval, first, I approximate the identified set by discretizing the support of the outcome where the discretization becomes finer as the sample size increases. The approximation for the identified set resembles many moment inequalities in Chernozhukov, Chetverikov, and Kato (2014), who consider a finite but divergent number of moment inequalities. I apply a bootstrap method in Chernozhukov, Chetverikov, and Kato (2014) to construct a confidence interval with uniformly asymptotically valid asymptotic size control. The confidence interval also rejects parameter values which do not belong to the sharp identified set. An empirical application and a Monte Carlo simulation demonstrate a finite sample property of the proposed inference method. The empirical exercise is based on Abadie (2003), who studies the effects of 401(k) participation on financial savings, and I consider a misclassification of the 401(k) participation.[4]

As an extension, I consider the dependence between the instrument and the measurement error. In this case, there is no assumption on the measurement error, and the measured treatment has no information on the local average treatment effect. Even without using the measured treatment, however, I can still apply the same identification strategy and obtain finite (but less tight) bounds on the local average treatment effect.

Moreover, I offer a new use of repeated measurements as additional sources for identification. The existing practice of repeated measurements uses one of them as an instrumental variable, as in Hausman, Newey, Ichimura, and Powell (1991), Hausman, Newey,

---

[4]The pension type is subject to a measurement error. See, for example, Gustman, Steinmeier, and Tabatabai (2008) for the pension-type misclassification in the Health and Retirement Study.

and Powell (1995), Mahajan (2006), and Hu (2008).[5] However, when the true treatment is endogenous, the repeated measurements are likely to be endogenous and are not good candidates for an instrumental variable. My identification strategy demonstrates that those variables are useful for bounding the local average treatment effect in the presence of measurement error, even if none of the repeated measurements are valid instrumental variables.

The remainder of this paper is organized as follows. Section 1.1 explains several empirical examples motivating mismeasured endogenous treatments and Section 1.2 reviews the related econometric literature. Section 2 introduces mismeasured treatments in the framework of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Section 3 constructs the identified set for the local average treatment effect. I also discuss two extensions. One extension describes how repeated measurements tighten the identified set, and the other dispenses with independence between the instrument and the measurement error. Section 4 proposes an inference procedure for the local average treatment effect. Section 5 conducts an empirical illustrations. Section 6 concludes. The Appendix collects proofs, remarks, and Monte Carlo simulations.

1.1 *Examples for mismeasured endogenous treatments*

I introduce several empirical examples in which binary treatments can be both endogenous and mismeasured at the same time. The first example is the return to schooling, in which the outcome is wages, and the treatment is educational attainment, for example, whether a person has completed college or not. Unobserved individual ability affects both the schooling decision and wage determination, which leads to the endogeneity of educational attainment (e.g., Card (2001)). Moreover, survey datasets record educational attainments based on the interviewee's answers, and these self-reported educational attainments are subject to measurement error. For example, Black, Sanders, and Taylor (2003) estimated that the 1990 Decennial Census had a 17.7% false positive rate of reporting a doctoral degree.

The second example is labor supply response to welfare program participation, in which the outcome is employment status and the treatment is welfare program participation. Self-reported welfare program participation in survey datasets can be mismeasured (Hernandez and Pudney (2007)). The psychological cost for welfare program participation, welfare stigma, affects job search behavior and welfare program participation simultaneously; that is, welfare stigma may discourage individuals from participating in a welfare program, and, at the same time, affect an individual's effort in the labor market. Moreover, the welfare stigma gives welfare recipients some incentive not to reveal their participation status to the survey, which causes endogenous measurement error in that the unobserved individual heterogeneity affects both the measurement error and the outcome.

The third example is the effect of a job training program on wages. As it is similar to the return to schooling, unobserved individual ability plays a crucial role in this example. Self-reported completion of job training program is also subject to measurement

---

[5]It is worthwhile to mention that Lewbel (2007) allows for a certain form of the endogeneity in a repeated measurement, under which a repeated measurement still satisfies some exclusion restriction.

error (Bollinger (1996)). Frazis and Loewenstein (2003) develop a methodology for evaluating a homogeneous treatment effect with mismeasured endogenous treatment, and apply their methodology to evaluate the effect of a job training program on wages.

The last example is the effect of maternal drug use on infant birth weight. Kaestner, Joyce, and Wehbeh (1996) estimate that a mother tends to underreport her drug use, but at the same time, she tends to report it correctly if she is a heavy user. When the degree of drug addiction cannot be observed, it becomes an individual unobserved heterogeneity which affects infant birth weight and the measurement in addition to the drug use.

1.2 *Literature review*

Here, I summarize the related econometric literature. Mahajan (2006), Lewbel (2007), and Hu (2008) used an instrumental variable to correct for measurement error in a binary (or discrete) treatment in the homogeneous treatment effect framework and they achieve nonparametric point identification of the average treatment effect. They assume that the true treatment is exogenous, whereas I allow it to be endogenous.

Finite mixture models are related to my analysis. I consider the unobserved binary treatment, whereas finite mixture models deal with unobserved type. Henry, Kitamura, and Salanié (2014) and Henry, Jochmans, and Salanié (2015) are the most closely related. They investigate the identification problem in finite mixture models, by using the exclusion restriction in which an instrumental variable only affects the mixing distribution of a type without affecting the component distribution (i.e., the conditional distribution given the type). If I applied their approach directly to my framework, their exclusion restriction would imply conditional independence between the instrumental variable and the outcome given the true treatment. Instead of applying the approaches in Henry, Kitamura, and Salanié (2014) and Henry, Jochmans, and Salanié (2015), I use a different exclusion restriction in which the instrumental variable does not affect the outcome or the measured treatment directly.

A few papers have applied an instrumental variable to a mismeasured binary regressor in the homogenous treatment effect framework. They include Aigner (1973), Kane, Rouse, and Staiger (1999), Bollinger (1996), Black, Berger, and Scott (2000), Frazis and Loewenstein (2003), and DiTraglia and García-Jimeno (2015). Frazis and Loewenstein (2003) and DiTraglia and García-Jimeno (2015) are the most closely related among them, since they allow for endogeneity. Here, I allow for heterogeneous treatment effects, and I contribute to the heterogeneous treatment effect literature by investigating the consequences of the measurement errors in the treatment.

Kreider and Pepper (2007), Molinari (2008), Imai and Yamamoto (2010), and Kreider, Pepper, Gundersen, and Jolliffe (2012) applied a partial identification strategy for the average treatment effect to the mismeasured binary regressor problem by utilizing the knowledge of the marginal distribution for the true treatment. Those papers use auxiliary datasets to obtain the marginal distribution for the true treatment. The framework in Kreider et al. (2012) is the most closely related to this paper, in that they allow for both treatment endogeneity and endogenous measurement error. My instrumental variable approach can be an an alternative strategy to deal with mismeasured endogenous treatment. It is worthwhile because, as mentioned in Schennach (2013), the availability of an

auxiliary dataset is limited in empirical research. Furthermore, it is not always the case that the results from auxiliary datasets is transported into the primary dataset (Carroll, Ruppert, Stefanski, and Crainiceanu (2012, p. 10)),

Some papers investigate mismeasured endogenous continuous variables, instead of binary variables. Amemiya (1985), Hsiao (1989), Lewbel (1998), Song, Schennach, and White (2015) considered nonlinear models with mismeasured continuous explanatory variables. The continuity of the treatment is crucial for their analysis, because they assume classical measurement error. The treatment in my analysis is binary and, therefore, the measurement error is nonclassical. Hu, Shiu, and Woutersen (2015) considered mismeasured endogenous continuous variables in single index models. However, their approach depends on taking derivatives of the conditional expectations with respect to the continuous variable. It is not clear if it can be extended to binary variables. Song (2015) considered the semiparametric model when endogenous continuous variables are subject to nonclassical measurement error. He assumes conditional independence between the instrumental variable and the outcome given the true treatment, which would impose some structure on the outcome equation when a treatment is binary. Instead I propose an identification strategy without assuming any structure on the outcome equation.

Chalak (2017) investigated the consequences of measurement error in the instrumental variable instead of the treatment. He assumed that the treatment is perfectly observed, whereas I allow for it to be measured with error. Since I assume that the instrumental variable is perfectly observed, my analysis is not overlapped with Chalak (2017).

Manski (2003), Blundell, Gosling, Ichimura, and Meghir (2007), and Kitagawa (2010) have similar identification strategy in the context of sample selection models. These papers also use the exclusion restriction of the instrumental variable for their partial identification results. Particularly, Kitagawa (2010) derived the integrated envelope from the exclusion restriction, which is similar to the total variation distance in my analysis because both of them are characterized as a supremum over the set of the partitions. First and the most importantly, I consider mismeasurement of the treatment, whereas the sample selection model considers truncation of the outcome. It is not straightforward to apply their methodologies in sample selection models into mismeasured treatment problem. Second, I offer an inference method with uniform size control, but Kitagawa (2010) derived only point-wise size control. Last, Blundell et al. (2007) and Kitagawa (2010) used their result for specification test, but I cannot use it to carry out a specification test because the sharp identified set of my analysis is always non-empty.

Finally, Calvi, Lewbel, and Tommasi (2017) and Yanagi (2017) have recently discussed identification issues of the local average treatment effect in the presence of a measurement error in the treatment variable. They are built on results in the previous draft of this paper to derive novel and important results when there are additional variables in a dataset: multiple measurements of the true treatment variable (Calvi, Lewbel, and Tommasi (2017)) or multiple instrumental variables (Yanagi (2017)). In contrast, the results of this paper are valid without these additional variables and only requires the assumptions in Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).

outcome Y $\longleftarrow$ true treatment T^* $\longleftarrow$ instrument Z
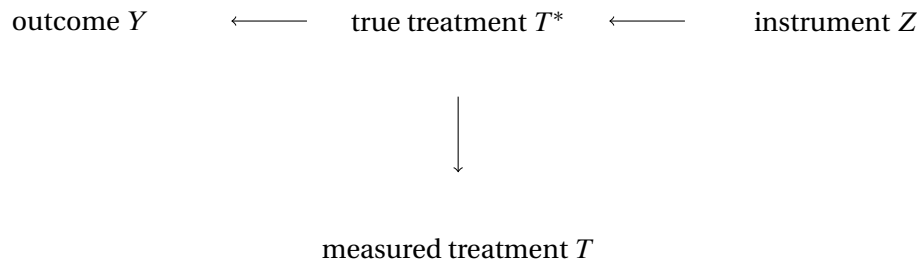
$\downarrow$

measured treatment T

FIGURE 2. Graphical representation of dependencies among variables.

2. LOCAL AVERAGE TREATMENT EFFECT FRAMEWORK WITH MISCLASSIFICATION

My analysis considers a mismeasured treatment in the framework of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). The objective is to evaluate the causal effect of a binary treatment $T^* \in \{0, 1\}$ on an outcome Y, where $T^* = 0$ represents the control group and $T^* = 1$ represents the treatment group. To deal with endogeneity of T^*, I use a binary instrumental variable $Z \in \{0, 1\}$ which shifts T^* exogenously without any direct effect on Y. The treatment T^* of interest is not directly observed, and instead there is a binary measurement $T \in \{0, 1\}$ for T^*. I put the $*$ symbol on T^* to emphasize that the true treatment T^* is unobserved. I allow Y to be discrete, continuous or mixed; Y is only required to have some known dominating finite measure μ_Y on the real line. For example, μ_Y can be the Lebesgue measure or the counting measure. Let **Y** be the support for the random variable Y and $\mathbf{T} = \{0, 1\}$ be the support for T.

To describe the data generating process, I consider the counterfactual variables. T_z^* is the counterfactual true treatment when $Z = z$. Y_{t^*} is the counterfactual outcome when $T^* = t^*$. T_{t^*} is the counterfactual measured treatment when $T^* = t^*$. The individual treatment effect is $Y_1 - Y_0$. It is not directly observed; Y_0 and Y_1 cannot be observed at the same time. Only Y_{T^*} is observable. Using the notation, the observed variables (Y, T, Z) are generated by the three equations:

$$T = T_{T^*}, \tag{1}$$

$$Y = Y_{T^*}, \tag{2}$$

$$T^* = T_Z^*. \tag{3}$$

Figure 2 graphically describes the relationship among the instrument Z, the (unobserved) true treatment T^*, the measured treatment T, and the outcome Y. (1) is the measurement equation, which is the arrow from T^* to T in Figure 2. $T - T^*$ is the measurement error; $T - T^* = 1$ represents a false positive and $T - T^* = -1$ represents a false negative. Equations (2) and (3) are the same as Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Equation (2) is the outcome equation, which is the arrow from T^* to Y in Figure 2. Equation (3) is the treatment assignment equation, which is the arrow from Z to T^* in Figure 2. A potentially nonzero correlation between (Y_0, Y_1) and (T_0^*, T_1^*) causes an endogeneity problem.

In a return to schooling analysis, Y is wages, T^* is the true indicator for college completion, Z is the proximity to college, and T is the self-reported college completion. The treatment effect $Y_1 - Y_0$ in the return to schooling is the effect of college completion T^* on wages Y. The college completion is not correctly measured in a survey dataset, such that only the self report T is observed.

This section and Section 3 impose only the following assumption.

ASSUMPTION 1. (i) *For each $t^* = 0, 1$, Z is independent of $(T_{t^*}, Y_{t^*}, T_0^*, T_1^*)$.* (ii) $T_1^* \geq T_0^*$ *almost surely.* (iii) $0 < P(Z = 1) < 1$.

Assumption 1(i) is the exclusion restriction and I consider stochastic independence instead of mean independence. Although it is stronger than the minimal conditions for the identification for the local average treatment effect without measurement error, a large part of the existing applied papers assume stochastic independence (Huber and Mellace (2015, p. 405)). Z is also independent of T_{t^*} conditional on (Y_{t^*}, T_0^*, T_1^*), which is the only assumption on the measurement error for the identified set in Section 3. (Section 3.4 even dispenses with this assumption.) Assumption 1(ii) is the monotonicity condition for the instrument, in which the instrument Z increases the value of T^* for all the individuals. de Chaisemartin (2016) relaxed the monotonicity condition, and it can be shown in Appendix D in the Supplementary Material (Ura (2018)) that the identification results in my analysis still holds with a slight modification under the complier-defiers-for-marginals condition in de Chaisemartin (2016). Note that Assumption 1 does not include a relevance condition for the instrumental variable. The standard relevance condition $T_1^* \neq T_0^*$ does not affect the identification results in my analysis. I will discuss the relevance condition in my framework after Theorem 4. Assumption 1(iii) excludes that Z is constant.

As I emphasized in the Introduction, the framework here does not assume anything on measurement error T_{t^*} except for its independence from Z. Assumption 1 does not impose any restriction on the marginal distribution of the measurement error T_{t^*} or on the relationship between the measurement error T_{t^*} and (Y_{t^*}, T_0^*, T_1^*). Particularly, the measurement error can be endogenous, that is, T_{t^*} and (Y_{t^*}, T_0^*, T_1^*) can be correlated.[6]

I focus on the local average treatment effect, which is defined by

$$\theta = E\big[Y_1 - Y_0 \mid T_0^* < T_1^*\big].$$

The local average treatment effect is the average of the treatment effect $Y_1 - Y_0$ over the subpopulation (the compliers) whose treatment status is strictly affected by the instrument. Imbens and Angrist (1994, Theorem 1) showed that the local average treatment

---

[6]Although it has not been supported in validation data studies (e.g., Black, Sanders, and Taylor (2003)), a majority of the literature on measurement error has assume that the measurement error is exogenous (Bound, Brown, and Mathiowetz (2001)). I also explore the identifying power of the exogenous measurement error assumption in Appendix C available in the Supplementary Material (Ura (2018)).

effect equals

$$\frac{\Delta E[Y \mid Z]}{\Delta E[T^* \mid Z]},$$

where I define $\Delta E[X \mid Z] = E[X \mid Z = 1] - E[X \mid Z = 0]$ for a random variable X. Note that $\Delta E[Y \mid Z]$ is the intent-to-treat effect, that is, the regression of Y on Z. The treatment is measured with error and, therefore, the fraction $\Delta E[Y \mid Z]/\Delta E[T^* \mid Z]$ is not the Wald estimand

$$\frac{\Delta E[Y \mid Z]}{\Delta E[T \mid Z]}.$$

Since $\Delta E[T^* \mid Z]$ is not identified, I cannot identify the local average treatment effect. The failure of point identification comes purely from the measurement error, because the local average treatment effect would be point-identified under $T = T^*$. In fact, my proposed methodology in this paper is essentially a bounding strategy of $\Delta E[T^* \mid Z]$ and I use the bound to construct the sharp identified set for the local average treatment effect.

3. Identified set for the local average treatment effect

This section shows how the instrumental variable partially identifies the local average treatment effect in the framework of Section 2. Before defining the identified set, I characterize the local average treatment effect as a function of the underlying distribution P^* of $(Y_0, Y_1, T_0, T_1, T_0^*, T_1^*, Z)$. I use the $*$ symbol on P^* to clarify that P^* is the distribution of the unobserved variables. I denote the expectation operator E by E_{P^*} when I need to clarify the underlying distribution. The local average treatment effect is a function of the unobserved distribution P^*:

$$\theta(P^*) = E_{P^*}[Y_1 - Y_0 \mid T_0^* < T_1^*].$$

I denote by Θ the parameter space for the local average treatment effect θ, that is, the set of $\int y f_1(y) \, d\mu_Y(y) - \int y f_0(y) \, d\mu_Y(y)$ where f_0 and f_1 are density functions dominated by the known probability measure μ_Y. For example, $\Theta = [-1, 1]$ when Y is binary.

The identified set is the set of parameter values for the local average treatment effect which is consistent with the distribution of the observed variables. I use P for the distribution of the observed variables (Y, T, Z). Equations (1), (2), and (3) induce the distribution of the observables (Y, T, Z) from the unobserved distribution P^*, and I denote by $P^*_{(Y,T,Z)}$ the induced distribution. When the distribution of (Y, T, Z) is P, the set of P^* which induces P is $\{P^* \in \mathcal{P}^* : P = P^*_{(Y,T,Z)}\}$, where $\mathcal{P}^*$ is the set of P^*'s satisfying Assumptions 1. For every distribution P of (Y, T, Z), the (sharp) identified set for the local average treatment effect is defined as $\Theta_I(P) = \{\theta(P^*) \in \Theta : P^* \in \mathcal{P}^* \text{ and } P = P^*_{(Y,T,Z)}\}$.

Imbens and Angrist (1994, Theorem 1) provided a relationship between $\Delta E[Y \mid Z]$ and the local average treatment effect:

$$\theta(P^*)P^*(T_0^* < T_1^*) = \Delta E_{P^*}[Y \mid Z]. \tag{4}$$

This equation gives the two pieces of information of $\theta(P^*)$. First, the sign of $\theta(P^*)$ is the same as $\Delta E_{P^*}[Y \mid Z]$. Second, the absolute value of $\theta(P^*)$ is at least the absolute value of $\Delta E_{P^*}[Y \mid Z]$. The following lemma summarizes these two pieces.

Lemma 2. *Under Assumption* 1,

$$\theta(P^*)\Delta E_{P^*}[Y \mid Z] \geq 0,$$

$$\left|\theta(P^*)\right| \geq \left|\Delta E_{P^*}[Y \mid Z]\right|.$$

I derive a new implication from the exclusion restriction for the instrumental variable in order to obtain an upper bound on $\theta(P^*)$ in absolute value. To explain the new implication, I introduce the total variation distance, which is the L^1 distance between the distribution f_1 and f_0: For any random variable X, define

$$TV_X = \frac{1}{2} \int \left|f_{X|Z=1}(x) - f_{X|Z=0}(x)\right| d\mu_X(x),$$

where μ_X is a dominating measure for the distribution of X.

Lemma 3. *Under Assumption* 1,

$$TV_{(Y,T)} \leq TV_{T^*} = P^*\left(T_0^* < T_1^*\right).$$

The first term, $TV_{(Y,T)}$, in Lemma 3 reflects the dependency of $f_{(Y,T)|Z=z}(y,t)$ on z, and it can be interpreted as the magnitude of the distributional effect of Z on (Y, T). The second and third terms, TV_{T^*} and $P^*(T_0^* < T_1^*)$, are the effect of the instrument Z on the true treatment T^*. Based on Lemma 3, the magnitude of the effect of Z on T^* is no smaller than the magnitude of the effect of Z on (Y, T).

The new implication in Lemma 3 gives a lower bound on $P^*(T_0^* < T_1^*)$ and, therefore, yields an upper bound on the local average treatment effect in absolute value, combined with equation (4). In other words,

$$\left|\theta(P^*)\right| = \frac{\left|\Delta E_{P^*}[Y \mid Z]\right|}{P^*\left(T_0^* < T_1^*\right)} \leq \frac{\left|\Delta E_{P^*}[Y \mid Z]\right|}{TV_{(Y,T)}}$$

as long as $TV_{(Y,T)} > 0$.

In summary, I can characterize the sharp identified set for the local average treatment effect as in the following theorem.

Theorem 4. *Suppose that Assumption* 1 *holds, and consider an arbitrary data distribution* P *of* (Y, T, Z). *The identified set* $\Theta_I(P)$ *for the local average treatment effect is characterized as follows*: $\Theta_I(P) = \Theta$ *if* $TV_{(Y,T)} = 0$; *otherwise,*

$$\Theta_I(P) = \begin{cases} \left[\Delta E_P[Y \mid Z], \dfrac{\Delta E_P[Y \mid Z]}{TV_{(Y,T)}}\right] & \text{if } \Delta E_P[Y \mid Z] > 0, \\[3mm] \{0\} & \text{if } \Delta E_P[Y \mid Z] = 0, \\[3mm] \left[\dfrac{\Delta E_P[Y \mid Z]}{TV_{(Y,T)}}, \Delta E_P[Y \mid Z]\right] & \text{if } \Delta E_P[Y \mid Z] < 0. \end{cases}$$

The total variation distance $TV_{(Y,T)}$ plays two roles in determining the sharp identified set in this theorem. First, $TV_{(Y,T)}$ measures the strength of the instrumental variable, that is, $TV_{(Y,T)} > 0$ is the relevance condition in my identification analysis. When $TV_{(Y,T)} > 0$, the interval in the above theorem is a strict subset of the whole parameter space, which implies that Z has some identifying power for the local average treatment effect. By contrast, $TV_{(Y,T)} = 0$ means that the instrumental variable Z does not affect Y and T, in which case Z has no identifying power for the local average treatment effect. In this case, $f_{(Y,T)|Z=1} = f_{(Y,T)|Z=0}$ almost everywhere over (y, t) and particularly $\Delta E_P[Y \mid Z] = 0$. Note that all the three inequalities in Theorem 4 have no restriction on θ in this case. Second, $TV_{(Y,T)}$ determines the length of the sharp identified set. The length is $|\Delta E_P[Y \mid Z]|(TV_{(Y,T)}^{-1} - 1)$, which is a decreasing function in $TV_{(Y,T)}$.

In general, the lower and upper bounds of the sharp identified set are not equal to the local average treatment effect. The lower bound is weakly smaller (in the absolute value) than the local average treatment effect, because the size of the compliers is weakly smaller than one. The upper bound is weakly larger (in the absolute value) than the local average treatment effect, because $TV_{(Y,T)}$ is weakly smaller than the size of the compliers due to the mismeasurement of the treatment variable.

The standard relevance condition $\Delta E_P[T \mid Z] \neq 0$ is not required in Theorem 4. $\Delta E_P[T \mid Z] \neq 0$ is a necessary condition to define the Wald estimand, but the sharp identified set does not depend directly on the Wald estimand. In fact, $TV_{(Y,T)} > 0$ in Theorem 4 is weaker than $\Delta E_P[T \mid Z] \neq 0$.

Note that the sharp identified set is always nonempty. There is no testable implications on the distribution of the observed variables and, therefore, it is impossible to conduct a specification test for Assumption 1.

3.1 *Wald estimand and the identified set*

The Wald estimand $\Delta E_P[Y \mid Z]/\Delta E_P[T \mid Z]$ can be outside the identified set. One necessary and sufficient condition for the Wald estimand to be included in the identified set is given as follows.

Lemma 5. *The Wald estimand is in the identified set if and only if*

$$f_{(Y,T)|Z=1}(y, 1) \geq f_{(Y,T)|Z=0}(y, 1) \quad and \quad f_{(Y,T)|Z=1}(y, 0) \leq f_{(Y,T)|Z=0}(y, 0). \tag{5}$$

This condition in (5) are the testable implications from the the local average treatment effect framework without measurement error (Balke and Pearl (1997) and Heckman and Vytlacil (2005)). The recent papers by Huber and Mellace (2015), Kitagawa (2015), and Mourifié and Wan (2016) proposed the testing procedures for (5). Based on the results in Theorem 4, their testing procedures are reinterpreted as a test for the null hypothesis that the Wald estimand is inside the sharp upper bound on the local average treatment effect.

The condition (5) cannot be used for testing the existence of a measurement error. Even if there is nonzero measurement error, (5) can still hold. One example is

$$Z \sim \text{Bernoulli}(0.5),$$
$$T^* = Z,$$
$$Y = U_1,$$
$$T = T^* + (1 - 2T^*)1\{U_2 \leq 1/3\},$$

where, conditional on Z, U_1 and U_2 are independently drawn from the uniform distribution. This data generating process P^* satisfies Assumption 1. Furthermore, (5) holds, because

$$f_{(Y,T)|Z=1}(y, 1) = f_{U_1}(y)P(U_2 > 1/3) = 2/3,$$
$$f_{(Y,T)|Z=0}(y, 1) = f_{U_1}(y)P(U_2 \leq 1/3) = 1/3,$$
$$f_{(Y,T)|Z=1}(y, 0) = f_{U_1}(y)P(U_2 \leq 1/3) = 1/3,$$
$$f_{(Y,T)|Z=0}(y, 0) = f_{U_1}(y)P(U_2 > 1/3) = 2/3$$

for every $y \in [0, 1]$.

3.2 *Conditional exogeneity of the instrumental variable*

As in Abadie (2003) and Frölich (2007), this section considers the conditional exogeneity of the instrumental variable Z in which Z is exogenous given a set of covariates V.

ASSUMPTION 6. *There is some variable V taking values in a set* **V** *satisfying the following properties.* (i) *For each $t^* = 0, 1$, Z is conditionally independent of the variables $(T_{t^*}, Y_{t^*}, T_0^*, T_1^*)$ given V.* (ii) *$T_1^* \geq T_0^*$ almost surely.* (iii) *$0 < P(Z = 1 \mid V) < 1$.*

I define the V-conditional total variation distance by

$$TV_{X|V} = \frac{1}{2} \int \left| f_{X|Z=1,V}(x) - f_{X|Z=0,V}(x) \right| d\mu_X(x).$$

Note that $TV_{X|V}$ is a random variable as a function of V. Under the conditional exogeneity of Z, Theorem 4 becomes as follows.

THEOREM 7. *Suppose that Assumption 6 holds, and consider an arbitrary data distribution P of (Y, T, Z, V). The identified set $\Theta_I(P)$ for the local average treatment effect is characterized as follows:* $\Theta_I(P) = \Theta$ *if* $E_P[TV_{(Y,T)|V}] = 0$; *otherwise,*

$$\Theta_I(P) = \begin{cases} \left[E_P\big[\Delta E_P[Y \mid Z, V]\big], \dfrac{E_P\big[\Delta E_P[Y \mid Z, V]\big]}{E_P[TV_{(Y,T)|V}]} \right] & \text{if } E_P\big[\Delta E_P[Y \mid Z, V]\big] > 0, \\ \{0\} & \text{if } E_P\big[\Delta E_P[Y \mid Z, V]\big] = 0, \\ \left[\dfrac{E_P\big[\Delta E_P[Y \mid Z, V]\big]}{E_P[TV_{(Y,T)|V}]}, E_P\big[\Delta E_P[Y \mid Z, V]\big] \right] & \text{if } E_P\big[\Delta E_P[Y \mid Z, V]\big] < 0. \end{cases}$$

3.3 *Identifying power of repeated measurements*

The identification strategy in the above analysis offers a new use of repeated measurements as additional sources for identification. The use of repeated measurements (e.g., Hausman et al. (1991)) is a popular approach in the literature on measurement error, but they cannot be instrumental variables in this framework. This is because the true treatment T^* is endogenous and it is natural to suspect that a measurement of T^* is also endogenous. The more accurate the measurement is, the more likely it is to be endogenous. Nevertheless, the identification strategy of this paper incorporates repeated measurements as an additional information to tighten the identified set for the local average treatment effect, when they are coupled with the instrumental variable Z. Unlike the other paper on repeated measurements, I do not need to assume the independence of measurement errors among multiple measurements. The strategy also benefits from having more than two measurements unlike Hausman et al. (1991) who achieve point identification with two measurements.

Consider a repeated measurement R for T^*. Like $T = T_{T^*}$, I model R using the counterfactual outcome notations. R_1 is a counterfactual second measurement when the true treatment T^* is 1, and R_0 is a counterfactual second measurement when the true treatment T^* is 0. Then the data generation of R is

$$R = R_{T^*}.$$

I strengthen Assumption 1 by assuming that the instrumental variable Z is independent of R_{t^*} conditional on $(Y_{t^*}, T_{t^*}, T_0^*, T_1^*)$.

ASSUMPTION 8. (i) Z is independent of $(R_{t^*}, T_{t^*}, Y_{t^*}, T_0^*, T_1^*)$ for each $t^* = 0, 1$. (ii) $T_1^* \geq T_0^*$ almost surely. (iii) $0 < P(Z = 0) < 1$.

Note that I do not assume the independence between R_{t^*} and T_{t^*}, where the independence between the measurement errors is a key assumption when the repeated measurement is an instrumental variable. Assumption 8 tightens the identified set for the local average treatment effect as follows.

The requirement on R does not restrict R to have the same support as T^*. In fact, R can be any variable which depends on T^*. For example, R can be an outcome variable other than Y.

THEOREM 9. *Suppose that Assumption 8 holds, and consider an arbitrary data distribution P of (R, Y, T, Z). The identified set $\Theta_I(P)$ for the local average treatment effect is characterized as follows*: $\Theta_I(P) = \Theta$ *if* $TV_{(R,Y,T)} = 0$; *otherwise,*

$$\Theta_I(P) = \begin{cases} \left[\Delta E_P[Y \mid Z], \dfrac{\Delta E_P[Y \mid Z]}{TV_{(R,Y,T)}} \right] & \textit{if } \Delta E_P[Y \mid Z] > 0, \\[2mm] \{0\} & \textit{if } \Delta E_P[Y \mid Z] = 0, \\[2mm] \left[\dfrac{\Delta E_P[Y \mid Z]}{TV_{(R,Y,T)}}, \Delta E_P[Y \mid Z] \right] & \textit{if } \Delta E_P[Y \mid Z] < 0. \end{cases}$$

The identified set in Theorem 9 is weakly smaller than the identified set in Theorem 4. The total variation distance $TV_{(R,Y,T)}$ in Theorem 9 is weakly larger than that in Theorem 4, because, using the triangle inequality,

$$
\begin{aligned}
TV_{(R,Y,T)} &= \frac{1}{2} \sum_{t=0,1} \iint \left| (f_{(R,Y,T)|Z=1} - f_{(R,Y,T)|Z=0})(r,y,t) \right| d\mu_R(r)\, d\mu_Y(y) \\
&\geq \frac{1}{2} \sum_{t=0,1} \int \left| \int (f_{(R,Y,T)|Z=1} - f_{(R,Y,T)|Z=0})(r,y,t)\, d\mu_R(r) \right| d\mu_Y(y) \\
&= \frac{1}{2} \sum_{t=0,1} \int \left| (f_{(Y,T)|Z=1} - f_{(Y,T)|Z=0})(y,t) \right| d\mu_Y(y) \\
&= TV_{(Y,T)}
\end{aligned}
$$

and the strict inequality holds unless the sign of $(f_{(R,Y,T)|Z=1} - f_{(R,Y,T)|Z=0})(r,y,t)$ is constant in r for every (y,t). Therefore, it is possible to test whether the repeated measurement R has additional information, by testing whether the sign of $(f_{(R,Y,T)|Z=1} - f_{(R,Y,T)|Z=0})(r,y,t)$ is constant in r.

3.4 Dependence between measurement error and instrumental variable

It is still possible to apply the same identification strategy and obtain finite (but less tight) bounds on the local average treatment effect, even without the independence between the instrumental variable and the measurement error. (Assumption 1(i) implies that Z is independent of T_{t^*} for each $t^* = 0, 1$.) Instead Assumption 1 is weakened to allow for the measurement error T_{t^*} to be correlated with the instrumental variable Z.

ASSUMPTION 10. (i) Z is independent of (Y_{t^*}, T_0^*, T_1^*) for each $t^* = 0, 1$. (ii) $T_1^* \geq T_0^*$ almost surely. (iii) $0 < P(Z=0) < 1$.

Theorem 11 shows that the above observations characterize the identified set for the local average treatment effect under Assumption 10.

THEOREM 11. Suppose that Assumption 10 holds, and consider an arbitrary data distribution P of (Y, T, Z). The identified set $\Theta_I(P)$ for the local average treatment effect is characterized as follows: $\Theta_I(P) = \Theta$ if $TV_Y = 0$; otherwise,

$$
\Theta_I(P) = \begin{cases} \left[\Delta E_P[Y \mid Z], \dfrac{\Delta E_P[Y \mid Z]}{TV_Y} \right] & \text{if } \Delta E_P[Y \mid Z] > 0, \\[3mm] \{0\} & \text{if } \Delta E_P[Y \mid Z] = 0, \\[3mm] \left[\dfrac{\Delta E_P[Y \mid Z]}{TV_Y}, \Delta E_P[Y \mid Z] \right] & \text{if } \Delta E_P[Y \mid Z] < 0. \end{cases}
$$

The difference from Theorem 4 is that Theorem 11 does not depend on the measured treatment T. Although it is observed in the dataset, T does not have any information on the local average treatment effect because Assumption 10 does not restrict T. When $TV_Y > 0$, there are nontrivial upper and lower bounds on the local average treatment effect even without using the measured treatment T.

4. Inference

Based on the sharp identified set in the presence of covariates (Theorem 7), this section constructs a confidence interval for the local average treatment effect based on an i.i.d. sample $\{W_i : 1 \leq i \leq n\}$ of $W = (Y, T, Z, V)$. The confidence interval described below controls the asymptotic size uniformly over a class of data generating processes, and rejects all the fixed alternatives.

The identified set in Theorem 7 is characterized by moment inequalities as follows.

LEMMA 12. *Let P be an arbitrary data distribution of $W = (Y, T, Z, V)$. Under Assumption 6, $\Theta_I(P)$ is the set of $\theta \in \Theta$ in which*

$$E_P\left[-\frac{Z - \pi(V)}{\pi(V)\bigl(1 - \pi(V)\bigr)}\operatorname{sgn}(\theta)Y\right] \leq 0, \tag{6}$$

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)\bigl(1 - \pi(V)\bigr)}\operatorname{sgn}(\theta)Y - |\theta|\right] \leq 0, \tag{7}$$

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)\bigl(1 - \pi(V)\bigr)}\bigl(|\theta|h(Y, T, V) - \operatorname{sgn}(\theta)Y\bigr)\right] \leq 0 \quad \text{for all } h \in \mathbf{H}, \tag{8}$$

where $\pi(V) = P(Z = 1 \mid V)$, $\mathbf{H}$ is the set of measurable functions on $\mathbf{Y} \times \mathbf{T} \times \mathbf{V}$ taking a value in $\{-0.5, 0.5\}$ and $\operatorname{sgn}(x) \equiv \mathbb{1}\{x \geq 0\} - \mathbb{1}\{x < 0\}$.

The first condition in (6) states that the local average treatment effect θ has the same sign as the intent-to-treat effect

$$E_P\bigl[\Delta E_P[Y \mid Z, V]\bigr] = E_P\left[Y\frac{Z - E_P[Z \mid V]}{E_P[Z \mid V]E_P[1 - Z \mid V]}\right].$$

The second condition in (7) is $|\theta| \geq |E_P[\Delta E_P[Y \mid Z, V]]|$. The last condition in (8) corresponds to

$$|\theta| \leq \frac{\bigl|E_P\bigl[\Delta E_P[Y \mid Z, V]\bigr]\bigr|}{E_P[TV_{(Y,T)|V}]} = \frac{\bigl|E_P\bigl[\Delta E_P[Y \mid Z, V]\bigr]\bigr|}{\sup\limits_{h \in \mathbf{H}} E_P\bigl[\Delta E_P\bigl[h(Y, T, V) \mid Z, V\bigr]\bigr]},$$

where I use $TV_{(Y,T)|V} = \sup_{h \in \mathbf{H}} \Delta E_P[h(Y, T, V) \mid Z, V]$. The derivations are found in the proof of Lemma 12 in Appendix B.

I construct a $(1 - \alpha)$-confidence interval for the local average treatment effect θ with treating π as a nuisance parameter for given $\alpha \in (0, 0.5)$. I assume that a $(1 - \delta)$-confidence interval $\mathcal{C}_{\pi,n}(\delta)$ for π is available for researchers for given $\delta \in (0, \alpha)$. Given $\mathcal{C}_{\pi,n}(\delta)$, I construct the $(1 - \alpha - \delta)$-confidence interval $\mathcal{C}_{\theta,n}(\alpha + \delta)$ for the local average treatment effect as

$$\mathcal{C}_{\theta,n}(\alpha + \delta) = \bigcup_{\pi \in \mathcal{C}_{\pi,n}(\delta)} \bigl\{\theta \in \Theta : T(\theta, \pi) \leq c(\alpha, \theta, \pi)\bigr\},$$

where $T(\theta, \pi)$ and $c(\alpha, \theta, \pi)$ are defined below using the bootstrap-based testing (Chernozhukov, Chetverikov, and Kato (2014)).

The number of the moment inequalities in Lemma 12 can be finite or infinite, which determines whether some of the existing methods can be applied directly to the inference on the local average treatment effect. When (Y, V) has finite supports and, therefore, **H** is finite, the sharp identified set is characterized by a finite number of inequalities and, therefore, I can apply inference methods based on unconditional moment inequalities. To the best of my knowledge, however, inference for the local average treatment effect in my framework does not fall directly into the existing moment inequality models when either Y or V is continuous. When either Y or V is continuous, the sharp identified set is characterized by an uncountably infinite number of inequalities. In the current literature on the partially identified parameters, an infinite number of moment inequalities are mainly considered in the context of conditional moment inequalities. The identified set in this paper is not characterized by conditional moment inequalities.

I consider a sequence of finite sets $\mathbf{H}_n$ which converges to **H** as a sample size increases. (The convergence is formally defined in Assumption 14, and an example for $\mathbf{H}_n$ appears after Assumption 14.) Note that, when **H** is finite, $\mathbf{H}_n$ can be equal to **H**. If **H** is replaced with $\mathbf{H}_n$ in Lemma 12, the number of the moment inequalities becomes finite. At the same time, as $\mathbf{H}_n$ approaches to **H**, the approximation error from using $\mathbf{H}_n$ converges to zero, and the number of the inequalities can be increasing, particularly diverging to the infinity when **H** includes infinite elements. The approximated identified set is characterized by a finite number of the following moment inequalities:

$$E_P\left[-\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} \operatorname{sgn}(\theta)Y\right] \leq 0, \tag{9}$$

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} \operatorname{sgn}(\theta)Y - |\theta|\right] \leq 0, \tag{10}$$

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))}\big(|\theta|h(Y, T, V) - \operatorname{sgn}(\theta)Y\big)\right] \leq 0 \quad \text{for all } h \in \mathbf{H}_n. \tag{11}$$

Denote by p_n the resulting number of moment inequalities, that is, the number of elements in $\mathbf{H}_n$ plus 2.

For the size $\alpha \in (0, 0.5)$, I construct a test statistic $T(\theta, \pi)$ and a critical value $c(\alpha, \theta, \pi)$ via the two-step multiplier bootstrap in Chernozhukov, Chetverikov, and Kato (2014), where Appendix A describes the procedures. Chernozhukov, Chetverikov, and Kato (2014) studies the testing problem for moment inequality models in which the number of the moment inequalities is finite but growing. Since the number of the moment inequalities in (9)–(11) is finite but growing, their results are applicable to construct a confidence interval based on (9)–(11).

ASSUMPTION 13. *Given positive constants C_2 and η, the class of data generating processes, denoted by $\mathcal{P}_0$, and the parameter spaces $\Theta \times \Pi$ satisfy the following properties*: (i) $\max\{E_P[Y^3]^{2/3}, E_P[Y^4]^{1/2}\} < C_2$, (ii) $\Theta \subset \mathbb{R}$ *is bounded*, (iii) *The random vari-*

able inside E_P in (9)–(11) *has a nonzero variance for every* $j = 1, \ldots, p_n$ *and every* $\theta \in \Theta$, (iv) $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_0} P(\pi \in \mathcal{C}_{\pi,n}(\delta)) \geq 1 - \delta$, *and* (v) $\eta < \pi(V) < 1 - \eta$ *for every* $\pi \in \Pi$.

The first assumption (i) is a regularity condition about the moments of Y. The second assumption (ii) requires researchers to know ex ante upper and lower bounds on the parameter. The third assumption (iii) guarantees that the test statistic is well-defined. The fourth assumption (iv) is that the confidence interval for π controls the size uniformly over $\mathcal{P}_0$. The last assumption (v) is that the propensity score $\pi(v) = P(Z = 1 \mid V = v)$ is bounded away from zero and one.

In this paper, I assume that $\{\mathbf{H}_n\}$ satisfies the following conditions.

ASSUMPTION 14. (i) $\mathbf{H}_n \subset \mathbf{H}_{n+1}$. (ii) *The convergence*

$$\sup_{h\in\mathbf{H}} E_P\left[\frac{Z - \pi(V)}{\pi(V)\left(1 - \pi(V)\right)} h(Y, T, V)\right] - \max_{h\in\mathbf{H}_n} E_P\left[\frac{Z - \pi(V)}{\pi(V)\left(1 - \pi(V)\right)} h(Y, T, V)\right] \to 0 \quad (12)$$

holds uniformly over $\pi \in \Pi$ *and* $P \in \mathcal{P}_0$. (iii) *The number of elements in* $\mathbf{H}_n$ *satisfies*

$$\log^{7/2}(p_n n) \leq C_1 n^{1/2 - c_1} \quad and \quad \log^{1/2} p_n \leq C_1 n^{1/2 - c_1} \quad (13)$$

for some $c_1 \in (0, 1/2)$ *and* $C_1 > 0$.

An example of $\mathbf{H}_n$ is obtained by discretizing $\mathbf{Y} \times \mathbf{T} \times \mathbf{V}$. Partition $\mathbf{Y} \times \mathbf{T} \times \mathbf{V}$ by $I_{n,1}, \ldots, I_{n,K_n}$, in which the intervals $I_{n,k}$ and the grid size K_n depend on the sample size n. Let $h_{n,j}$ be a generic function of $\mathbf{Y} \times \mathbf{T} \times \mathbf{V}$ into $\{-0.5, 0.5\}$ that is constant over $I_{n,k}$ for every $1 \leq k \leq K_n$. Let $\mathbf{H}_n = \{h_{n,1}, \ldots, h_{n,2^{K_n}}\}$ be the set of all such functions. Lemma 15 shows that this construction of $\mathbf{H}_n$ satisfies equation (12) under conditions on $I_{n,k}$ and $f_{(Y,T)|Z=z}$. The conditions in Lemma 15 guarantee that the approximation error from the discretization vanishes as the sample size n increases.

LEMMA 15. *Assumption* 14 *holds if* (i) *the partition* $I_{n+1,1}, \ldots, I_{n+1,K_{n+1}}$ *is a refinement of the partition* $I_{n,1}, \ldots, I_{n,K_n}$, (ii) $p_n = 2^{K_n} + 2$ *satisfies* (13), (iii) *there is a positive constant* D_1 *such that* $I_{n,k}$ *is a subset of some open ball with radius* D_1/K_n *in* $\mathbf{Y} \times \mathbf{T} \times \mathbf{V}$, *and* (iv) *the density function* $f_{(Y,T)|Z=z,V}$ *is Hölder continuous in* (y, t, v) *with the Hölder constant* D_0 *and exponent* d.

Theorem 16 shows asymptotic properties of the confidence interval $\mathcal{C}_{\theta,n}(\alpha + \delta)$. The first result (i) is the uniform asymptotic size control and the second result (ii) is the consistency against all the fixed alternatives.

THEOREM 16. *Construct* $T(\theta, \pi)$ *and* $c(\alpha, \theta, \pi)$ *via the two-step multiplier bootstrap as in Appendix* A. *Suppose that Assumptions* 13 *and* 14 *hold.* (i) *The confidence interval controls the asymptotic size uniformly*:

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_0, \theta\in\Theta_I(P)} P\big(\theta \in \mathcal{C}_{\theta,n}(\alpha + \delta)\big) \geq 1 - \alpha - \delta.$$

(ii) *If equation* (12) *holds, the confidence interval excludes all the fixed alternatives*:

$$\lim_{n\to\infty} P\big(\theta \in \mathcal{C}_{\theta,n}(\alpha + \delta)\big) = 0 \quad for\ every\ (\theta, P) \in \Theta \times \mathcal{P}_0\ with\ \theta \notin \Theta_I(P).$$

5. Empirical illustrations

This section studies the effects of the 401(k) participation on financial savings using the inference method in Section 4. I introduce a measurement error problem to the analysis of Poterba, Venti, and Wise (1995) and Abadie (2003), which investigate the local average treatment effect using the eligibility for a 401(k) program. The robustness to misclassification is empirically relevant because the retirement pension plan type is subject to a measurement error in survey datasets. Using the Health and Retirement Study, for example, Gustman, Steinmeier, and Tabatabai (2008) estimated that around one-fourth of the survey respondents misclassified their pension plan type.

The dataset in my analysis is from the Survey of Income and Program Participation (SIPP) of 1991. I follow the data construction in Abadie (2003). The sample consists of households in which at least one person is in employment, which has no income from self-employment, and whose annual family income is between $10,000 and $200,000. The resulting sample size is 9275.

The outcome variable Y is the net amount of financial assets in dollars, and the measured treatment variable T is the self-reported participation in a 401(k) program. The 401(k) participation can be endogenous because participants in a 401(k) program might be more informed or plan more about retirement savings than nonparticipants. To control for the endogeneity problem, I use the 401(k) eligibility, which is the indicator of whether an employer offers a 401(k) program, as an instrumental variable Z. I also use the participation in an individual retirement account (IRA) as the variable R in Section 3.3. The control variables V includes constant, family income, age, age squared, marital status, and family size. The summary statistics for (Y, T, Z, R) are in Table 1.

To implement the proposed method, I impose Assumption 6 in this empirical exercise. Assumption 6(i) requires that, given the covariates V, the 401(k) eligibility Z is exogenous for the self-report T as well as the saving Y and the true 401(k) participation T^*. Poterba, Venti, and Wise (1995) justified the exogeneity of the 401(k) eligibility given the income level, based on that an employer determines the 401(k) eligibility wheres an employee determines saving and pension plan. Using the same reasoning, the self-report T is an employee's choice and then the 401(k) eligibility Z is considered to be exogenous for the self-report T. Assumption 6(ii) holds because a 401(k) plan is not available unless an employee is eligible. Assumption 6(iii) requires that the 401(k) eligibility is not deterministic given the covariates V. Given that the 401(k) eligibility is an employer's

TABLE 1. Summary statistics for Y, T, Z, and R.

	Mean	Standard Deviation
Y: Family net financial assets	19,071	63,963
T: 401(k) participation	0.2762	0.4472
Z: 401(k) eligibility	0.3921	0.4883
R: IRA participation	0.2543	0.4355

decision and the covariates V are employee's demographic variables, this condition is likely to hold.

To construct the confidence interval $C_{\pi,n}(\delta)$ for π, I use the linear probability model for the regression of the instrumental variable Z on the control variables V. I estimate β_π in $E[Z \mid V] = \pi(V) = V'\beta_\pi$ using OLS and compute the confidence interval using a nonparametric bootstrap.

To compare the proposed method with the exiting methods, I compute the Wald estimator, \$16,290, with a 95% bootstrapped confidence interval [5976, 27,611]. (The Wald estimator is the sample analogue of $E[\frac{Z-\pi(V)}{\pi(V)(1-\pi(V))}Y]/E[\frac{Z-\pi(V)}{\pi(V)(1-\pi(V))}T]$ with π being estimated in the linear probability model.) The estimated intent-to-treat effect is \$10,981 with a 95% bootstrapped confidence interval [4169, 18,558].

For the construction of $I_{n,1}, \ldots, I_{n,K_n}$, I construct a partition only over $\mathbf{Y} \times \mathbf{T}$ to make the number of K_n reasonable.[7] Using an integer L_n, I discretize the outcome variables Y using $\{y_\tau : \tau = 0, 1/L_n, \ldots, (L_n - 1)/L_n, 1\}$, where y_τ is the unconditional τ-quantiles of Y. Define $K_n = 2L_n$ and, for every $l = 1, \ldots, L_n$, define $I_{n,l} = \{(y, t) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 0\}$ and $I_{n,l+L_n} = \{(y, t) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 1\}$. I use $L_n = 1, 2, 3, 4$ in the following analysis.[8] Although the data-driven choice of L_n is beyond the scope of this paper, a possible rule-of-thumb choice is the largest L_n for which each $I_{n,k}$ has at least a certain number of observations, for example, 30. In this empirical exercise, the corresponding choice is $L_n = 3$.

Table 2 shows the confidence intervals for the local average treatment effect, where I use 2000 draws for the bootstraps and set $\beta = 0.1\%$ for the moment selection and $\delta = 1\%$ for the estimation of β_π. The confidence intervals for the local average treatment effect are qualitatively similar to the confidence interval for the Wald estimand. The confidence intervals do not shrink as L_n increases from 1 to 4. It is possibly because the data generation process does not violate the conditions in (5) to a large extent and, therefore, the Wald estimand is close to the sharp upper bound for the local average treatment effect.

TABLE 2. 90% and 95% confidence intervals for the local average treatment effect for different L_n. (Based on Theorem 7).

L_n	90% CI	95% CI
1	[5743 25,287]	[4465 27,415]
2	[5748 25,891]	[4461 28,062]
3	[5707 26,081]	[4443 28,163]
4	[5713 26,122]	[4430 28,296]

[7]Otherwise the number of the observations in some $I_{n,k}$'s can be small. Note that the resulting approximated identified set is still weakly tighter than the identified set based on the Wald estimand. Particularly when $L_n = 1$, the condition in (11) is equivalent to using the Wald estimand as the upper bound for $|\theta|$.

[8]The results are similar for larger values of L_n. Further results are available from the author upon request.

TABLE 3. 90% and 95% confidence intervals with using R as in Theorem 9.

L_n	90% CI	95% CI
1	[5741　25,829]	[4431　28,081]
2	[5696　26,197]	[4417　28,588]
3	[5652　26,487]	[4422　28,713]
4	[5665　26,612]	[4418　28,846]

TABLE 4. 90% and 95% confidence intervals without T as in Theorem 11.

L_n	90% CI	95% CI
1	[5781　Inf]	[4485　Inf]
2	[5741　81,973]	[4483　89,204]
3	[5729　108,861]	[4455　118,218]
4	[5725　84,999]	[4433　92,032]

Table 3 summarizes the confidence intervals with the IRA participation R as an additional measurement discussed in Theorem 9.[9] It shows similar values to Table 2, and it can be because the IRA participation has only little identifying power on the local average treatment effect.

Table 4 summarizes the confidence intervals without using the measured treatment T, as in Theorem 11.[10] The lower bound of the confidence intervals does not change from those in Table 2, because the lower bound of the identified set does not depend on the measured treatment T. The upper bound is 3–4 times larger than those in Table 2, which is the cost of not using T. When $L_n = 1$, TV_Y becomes zero, and then there is no finite upper bound on the local average treatment effect.

In summary, the empirical results based on the proposed method are qualitatively similar to those results based on the Wald estimator. Given that this paper allows for a misclassification of the treatment variable, the empirical exercise has demonstrated the robustness of the existing results which ignore the measurement error (cf. Poterba, Venti, and Wise (1995) and Abadie (2003)).

6. Conclusion

This paper studies the identifying power of an instrumental variable in the heterogeneous treatment effect framework when a binary treatment is mismeasured and endogenous. The assumptions in this framework are the monotonicity of the instrumen-

---

[9]In Table 3, I construct $I_{n,1}, \ldots, I_{n,K_n}$ as follows. Define $K_n = 4L_n$ and, for every $l = 1, \ldots, L_n$, define $I_{n,l} = \{(y,t,r) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 0, r = 0\}$, $I_{n,l+L_n} = \{(y,t,r) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 1, r = 0\}$, $I_{n,l+2L_n} = \{(y,t,r) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 0, r = 1\}$, and $I_{n,l+3L_n} = \{(y,t,r) : y_{(l-1)/L_n} < y \leq y_{l/L_n}, t = 1, r = 1\}$.

[10]In Table 4, I construct $I_{n,1}, \ldots, I_{n,K_n}$ as follows. Define $K_n = L_n$ and, for every $l = 1, \ldots, L_n$, define $I_{n,l} = \{(y,t,r) : y_{(l-1)/L_n} < y \leq y_{l/L_n}\}$.

tal variable Z on the true treatment T^* and the exogeneity of Z. I use the total variation distance to characterize the identified set for the local average treatment effect $E[Y_1 - Y_0 \mid T_0^* < T_1^*]$. I also provide an inference procedure for the local average treatment effect. Unlike the existing literature on measurement error, the identification strategy does not reply on a specific structure of the measurement error; the only assumption on the measurement error is its independence of the instrumental variable.

There are several directions for future research. First, the choice of the partition $\mathbf{I}_n$ in Section 4, particularly the choice of K_n, is an interesting direction. To the best of my knowledge, the literature on many moment inequalities has not investigated how econometricians choose the numbers of the many moment inequalities. Second, it may be worthwhile to investigate the other parameter for the treatment effect. This paper has focused on the local average treatment effect, but the literature on heterogeneous treatment effect has emphasized the importance of choosing an adequate treatment effect parameter in order to answer relevant policy questions. Third, it is also interesting to investigate various assumptions on the measurement errors. In some empirical settings, for example, it may be reasonable to assume that the measurement error is one-directional (e.g., misclassification happens only when $T^* = 1$). Fourth, it is not trivial how the analysis of this paper can be extended to an instrumental variable taking more than two values. For a general instrumental variable, it is always possible to focus on two values of the instrumental variable and apply the analysis of this paper to the subpopulation with the instrumental variable taking these two values. However, different pairs of the values can have different compliers, so that the parameter of interest is not common across different pairs, as discussed in Heckman and Vytlacil (2005).

Appendix A: Multiplier bootstrap

This section describes the construction of $T(\theta, \pi)$ and $c(\alpha, \theta, \pi)$ via the two-step multiplier bootstrap in Chernozhukov, Chetverikov, and Kato (2014). The two-step procedures involves inequality selection in the first step and testing in the second step. The first step uses $\beta = \beta_n$ as the size for selecting the moments. Chernozhukov, Chetverikov, and Kato (2014) imposed $\beta < \alpha/3$ and $1/\beta_n \le C_1 \log(n)$ where C_1 is the constant in Assumption 12.

Define the moment functions based on (9)–(11). Define $\mathbf{H}_n = \{h_1, \ldots, h_{p_n-2}\}$ and

$$g_1(W, \theta, \pi) = -\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} \operatorname{sgn}(\theta)Y,$$

$$g_2(W, \theta, \pi) = \frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} \operatorname{sgn}(\theta)Y - |\theta|,$$

$$g_{2+j}(W, \theta, \pi) = \frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} \big(|\theta|h_j(Y, T, V) - \operatorname{sgn}(\theta)Y\big) \quad \text{for } 3 \le j \le p_n.$$

Then the approximated identified set is characterized by

$$\big\{(\theta, \pi) \in \Theta \times [0, 1] : E_P\big[g_j(W, \theta, \pi)\big] \text{ for every } j = 1, \ldots, p_n\big\}.$$

The test statistic for the true parameter values being (θ, π) is defined by

$$T(\theta, \pi) = \max_{1 \leq j \leq p_n} \frac{\sqrt{n}\hat{m}_j(\theta, \pi)}{\hat{\sigma}_j(\theta, \pi)},$$

where $\hat{m}_j(\theta, \pi)$ estimates $m_j(\theta, \pi) = E_P[g_j(W, \theta, \pi)]$, and $\hat{\sigma}_j^2(\theta, \pi)$ estimates the variance $\sigma_j^2(\theta, \pi)$ of $\sqrt{n}\hat{m}_j(\theta, \pi)$:

$$\hat{m}_j(\theta, \pi) = n^{-1} \sum_{i=1}^{n} g_j(W_i, \theta, \pi),$$

$$\hat{\sigma}_j^2(\theta, \pi) = n^{-1} \sum_{i=1}^{n} \big(g_j(W_i, \theta, \pi) - \hat{m}_j(\theta, \pi)\big)^2.$$

To conduct a multiplier bootstrap, generate n independent standard normal random variables $\epsilon_1, \ldots, \epsilon_n$. The centered bootstrap moments are

$$\hat{m}_j^B(\theta, \pi) = n^{-1} \sum_{i=1}^{n} \epsilon_i\big(g_j(W_i, \theta, \pi) - \hat{m}_j(\theta, \pi)\big).$$

The bootstrapped test statistic for inequality selection is defined by

$$T^{B,1}(\theta, \pi) = \max_{1 \leq j \leq p_n} \sqrt{n}\frac{\hat{m}_j^B(\theta, \pi)}{\hat{\sigma}_j(\theta, \pi)}.$$

The critical value $c^1(\beta, \theta, \pi)$ for inequality selection is defined as the conditional $(1-\beta)$-quantile of $T^{B,1}(\theta, \pi)$ given $\{W_i\}$. Define

$$\hat{J}_B = \left\{ j = 1, \ldots, p_n : \sqrt{n}\frac{\hat{m}_j^B(\theta, \pi)}{\hat{\sigma}_j(\theta, \pi)} > -2c^1(\beta, \theta, \pi) \right\}.$$

The bootstrapped test statistic is defined by

$$T^B(\theta, \pi) = \max_{j \in \hat{J}_B} \sqrt{n}\frac{\hat{m}_j^B(\theta, \pi)}{\hat{\sigma}_j(\theta, \pi)},$$

where $T^B(\theta, \pi)$ is defined as 0 if $\hat{J}_B = \emptyset$. The critical value $c(\alpha, \theta, \pi)$ is defined as the conditional $(1 - \alpha + 2\beta)$-quantile of $T^B(\theta, \pi)$ given $\{W_i\}$.

Appendix B: Proofs

Proof of Lemma 2

By equation (4), $\theta(P^*)\Delta E_{P^*}[Y \mid Z] = \theta(P^*)^2 P^*(T_0^* < T_1^*) \geq 0$, and $|\Delta E_{P^*}[Y \mid Z]| = |\theta(P^*)P^*(T_0^* < T_1^*)| \leq |\theta(P^*)|$.

Proof of Lemma 3

I obtain $f_{(Y,T)|Z=1} - f_{(Y,T)|Z=0} = P^*(T_0^* < T_1^*)(f_{(Y_1,T_1)|T_0^*<T_1^*} - f_{(Y_0,T_0)|T_0^*<T_1^*})$ by the same logic as Theorem 1 in Imbens and Angrist (1994):

$$
\begin{aligned}
f_{(Y,T)|Z=0} &= P^*\big(T_0^* = T_1^* = 1 \mid Z = 0\big) f_{(Y,T)|Z=0,T_0^*=T_1^*=1} \\
&\quad + P^*\big(T_0^* < T_1^* \mid Z = 0\big) f_{(Y,T)|Z=0,T_0^*<T_1^*} \\
&\quad + P^*\big(T_0^* = T_1^* = 0 \mid Z = 0\big) f_{(Y,T)|Z=0,T_0^*=T_1^*=0} \\
&= P^*\big(T_0^* = T_1^* = 1\big) f_{(Y_1,T_1)|T_0^*=T_1^*=1} \\
&\quad + P^*\big(T_0^* < T_1^*\big) f_{(Y_0,T_0)|T_0^*<T_1^*} \\
&\quad + P^*\big(T_0^* = T_1^* = 0\big) f_{(Y_0,T_0)|T_0^*=T_1^*=0}, \\
f_{(Y,T)|Z=1} &= P^*\big(T_0^* = T_1^* = 1\big) f_{(Y_1,T_1)|T_0^*=T_1^*=1} \\
&\quad + P^*\big(T_0^* < T_1^*\big) f_{(Y_1,T_1)|T_0^*<T_1^*} \\
&\quad + P^*\big(T_0^* = T_1^* = 0\big) f_{(Y_0,T_0)|T_0^*=T_1^*=0}.
\end{aligned}
$$

By the triangle inequality,

$$
\begin{aligned}
TV_{(Y,T)} &= \frac{1}{2} \sum_{t=0,1} \int \big| f_{(Y,T)|Z=1}(y,t) - f_{(Y,T)|Z=0}(y,t) \big| \, d\mu_Y(y) \\
&= \frac{1}{2} \sum_{t=0,1} \int \big| P^*\big(T_0^* < T_1^*\big)\big(f_{(Y_1,T_1)|T_0^*<T_1^*}(y,t) - f_{(Y_0,T_0)|T_0^*<T_1^*}(y,t)\big) \big| \, d\mu_Y(y) \\
&= P^*\big(T_0^* < T_1^*\big) \frac{1}{2} \sum_{t=0,1} \int \big| f_{(Y_1,T_1)|T_0^*<T_1^*}(y,t) - f_{(Y_0,T_0)|T_0^*<T_1^*}(y,t) \big| \, d\mu_Y(y) \\
&\leq P^*\big(T_0^* < T_1^*\big) \frac{1}{2} \sum_{t=0,1} \int \big(f_{(Y_1,T_1)|T_0^*<T_1^*}(y,t) + f_{(Y_0,T_0)|T_0^*<T_1^*}(y,t) \big) \, d\mu_Y(y) \\
&= P^*\big(T_0^* < T_1^*\big).
\end{aligned}
$$

Moreover, since $T_0^* \leq T_1^*$ almost surely,

$$
\begin{aligned}
P^*\big(T_0^* < T_1^*\big) &= \frac{1}{2}\big| P^*\big(T_0^* = 1\big) - P^*\big(T_1^* = 1\big) \big| + \frac{1}{2}\big| P^*\big(T_0^* = 0\big) - P^*\big(T_1^* = 0\big) \big| \\
&= \frac{1}{2} \sum_{t^*=0,1} \big| f_{T^*|Z=1}\big(t^*\big) - f_{T^*|Z=0}\big(t^*\big) \big| \\
&= TV_{T^*}.
\end{aligned}
$$

Proof of Lemma 5

Based on the triangle inequality,

$$
\begin{aligned}
TV_{(Y,T)} &= \frac{1}{2} \sum_{t=0,1} \int \left| f_{(Y,T)|Z=1}(y,t) - f_{(Y,T)|Z=0}(y,t) \right| d\mu_Y(y) \\
&\geq \frac{1}{2} \sum_{t=0,1} \left| \int \left(f_{(Y,T)|Z=1}(y,t) - f_{(Y,T)|Z=0}(y,t) \right) d\mu_Y(y) \right| \\
&= \frac{1}{2} \sum_{t=0,1} \left| f_{T|Z=1}(t) - f_{T|Z=0}(t) \right| \\
&= \left| f_{T|Z=1}(1) - f_{T|Z=0}(1) \right| \\
&= \left| \Delta E_P[T \mid Z] \right|.
\end{aligned}
$$

The equality holds if and only if the sign of $f_{(Y,T)|Z=1}(y,t) - f_{(Y,T)|Z=0}(y,t)$ is constant in y for every t. Since $|\Delta E_P[T \mid Z]| = \Delta E_P[T \mid Z]$ if and only if $f_{T|Z=1}(1) - f_{T|Z=0}(1)$ is positive, the condition in (5) is a necessary and sufficient condition for $TV_{(Y,T)} = \Delta E_P[T \mid Z]$, which is equivalent for the Wald estimand to belong to the identified set.

Proof of Theorems 4, 7, *and* 9

Theorems 4, 7, and 9 are special cases of the following theorem.

THEOREM 17. *There is some variable V taking values in a set* **V** *satisfying the following properties*:

   (i) *For each $t^* = 0, 1$, Z is conditionally independent of $(R_{t^*}, T_{t^*}, Y_{t^*}, T_0^*, T_1^*)$ given V.*
   (ii) $T_1^* \geq T_0^*$ *almost surely.*
   (iii) $0 < P(Z = 1 \mid V) < 1$.

Consider an arbitrary data distribution P of (R, Y, T, Z, V). The identified set $\Theta_I(P)$ for the local average treatment effect is the set of $\theta \in \Theta$ satisfying the following three inequalities:

$$
\begin{aligned}
|\theta| E_P[TV_{(R,Y,T)|V}] &\leq \left| E_P\big[\Delta E_P[Y \mid Z, V] \big] \right|, \\
\theta E_P\big[\Delta E_P[Y \mid Z, V] \big] &\geq 0, \\
|\theta| &\geq \left| E_P\big[\Delta E_P[Y \mid Z, V] \big] \right|.
\end{aligned}
$$

To show Theorem 17, I consider the two cases separately: $E_P[TV_{(R,Y,T)|V}] = 0$ or $E_P[TV_{(R,Y,T)|V}] > 0$.

Case 1: $E_P[TV_{(R,Y,T)|V}] = 0$.   In this case, $f_{(R,Y,T)|Z=z,V} = f_{(R,Y,T)|V}$ a.s. for every $z = 0, 1$. Let f_1 and f_0 be any pair of density functions for Y dominated by μ_Y. Define the

data generating process $P^*_{f_1,f_0}$:

$$(Z,V) \sim f_{(Z,V)},$$

$$\left(T_0^*, T_1^*\right) \mid Z, V = (1,1),$$

$$(R_1, Y_1, T_1) \mid \left(T_0^*, T_1^*, Z, V\right) \sim \begin{cases} f_{(R,Y,T)}(r,y,t) & \text{if } T_0^* = T_1^*, \\ f_1(y)f_{(R,T)}(r,t) & \text{if } T_0^* \neq T_1^*, \end{cases}$$

$$(R_0, Y_0, T_0) \mid \left(T_0^*, T_1^*, Z\right) \sim f_0(y)f_{(R,T)}(r,t).$$

Theorem 17 follows from the following three observations:

 (i) $P^*_{f_1,f_0}$ satisfies all the assumptions in Theorem 17,
 (ii) $P^*_{f_1,f_0}$ generates the data distribution P,
(iii) under $P^*_{f_1,f_0}$, the local average treatment effect is $\int y(f_1(y) - f_0(y)) \, d\mu_Y(y)$.

(i) $P^*_{f_1,f_0}$ satisfies the independence between Z and $(R_{t^*}, T_{t^*}, Y_{t^*}, T_0^*, T_1^*)$ given V for each $t^* = 0, 1$. Furthermore, $P^*_{f_1,f_0}$ satisfies $T_1^* \geq T_0^*$ almost surely.
(ii) Denote by f^* the density function of $P^*_{f_1,f_0}$. Then

$$\begin{aligned} f^*_{(R,Y,T)|Z=0,V} &= f^*_{(R_1,Y_1,T_1)|Z=0,V} \\ &= f_{(R,Y,T)|V} \\ &= f_{(R,Y,T)|Z=0,V}, \\ f^*_{(R,Y,T)|Z=1,V} &= f^*_{(R_1,Y_1,T_1)|Z=0,V} \\ &= f_{(R,Y,T)|V} \\ &= f_{(R,Y,T)|Z=1,V}, \end{aligned}$$

where the last equality uses $f_{(R,Y,T)|Z=0,V} = f_{(R,Y,T)|Z=1,V}$.
(iii) The local average treatment effect under $P^*_{f_1,f_0}$ is

$$\begin{aligned} E_{P^*_{f_1,f_0}}\left[Y_1 - Y_0 \mid T_0^* < T_1^*\right] &= E_{P^*_{f_1,f_0}}\left[Y_1 \mid T_0^* < T_1^*\right] - E_{P^*_{f_1,f_0}}\left[Y_0 \mid T_0^* < T_1^*\right] \\ &= \int y f_1(y) \, d\mu_Y(y) - \int y f_0(y) \, d\mu_Y(y). \end{aligned}$$

Case 2: $E_P[TV_{(R,Y,T)|V}] > 0$.   Lemma 3 is modified into the framework of Theorem 17.

LEMMA 18. *Under Assumption 8,* $E_P[TV_{(R,Y,T)|V}] \leq P^*(T_0^* < T_1^*)$.

PROOF. The proof is the same as Lemma 3 and this lemma follows from $f_{(R,Y,T)|Z=1,V} - f_{(R,Y,T)|Z=0,V} = P^*(T_0^* < T_1^* \mid V)(f_{(R_1,Y_1,T_1)|T_0^* < T_1^*,V} - f_{(R_0,Y_0,T_0)|T_0^* < T_1^*,V})$.   □

From Lemmas 2 and 18 and equation (4), all the three inequalities in Theorem 9 are satisfied for the true value of the local average treatment effect. To complete Theorem 17,

it suffices to show that, for any data generating process P, any point θ satisfying the three inequalities in Theorem 17 is the local average treatment effect under some data generating process whose data distribution is equal to P.

Define the two data generating processes: P_L^* and P_U^*. First, P_L^* is defined by

$$(Z, V) \sim f_{(Z,V)},$$

$$(T_0^*, T_1^*) \mid Z, V = (0, 1),$$

$$(R_0, Y_0, T_0) \mid (T_0^*, T_1^*, Z, V) \sim f_{(R,Y,T)\mid Z=0, V},$$

$$(R_1, Y_1, T_1) \mid (T_0^*, T_1^*, Z, V) \sim f_{(R,Y,T)\mid Z=1, V}.$$

Second, P_U^* is defined as follows. Using $\operatorname{sgn}(x) = \mathbb{1}\{x \geq 0\} - \mathbb{1}\{x < 0\}$, define $H = 0.5 \times \operatorname{sgn}(\Delta f_{(R,Y,T)\mid Z,V}(R, Y, T))$, and define P_U^* as

$$(Z, V) \sim f_{(Z,V)},$$

$$(T_0^*, T_1^*) \mid Z, V = \begin{cases} (0, 1) & \text{with probability } \Delta E_P[H \mid Z, V], \\ (0, 0) & \text{with probability } P(H = -0.5 \mid Z = 1, V), \\ (1, 1) & \text{with probability } P(H = 0.5 \mid Z = 0, V), \end{cases}$$

$$(R_1, Y_1, T_1) \mid (T_0^*, T_1^*, Z, V) \sim \begin{cases} \dfrac{\Delta f_{(R,Y,T,H)\mid Z,V}(r, y, t, 0.5)}{\Delta E_P[H \mid Z, V]} & \text{if } T_0^* < T_1^*, \\ f_{(R,Y,T)\mid H=0.5, Z=0, V}(r, y, t) & \text{if } T_0^* = T_1^*, \end{cases}$$

$$(R_0, Y_0, T_0) \mid (T_0^*, T_1^*, Z, V) \sim \begin{cases} -\dfrac{\Delta f_{(R,Y,T,H)\mid Z,V}(r, y, t, -0.5)}{\Delta E_P[H \mid Z, V]} & \text{if } T_0^* < T_1^*, \\ f_{(R,Y,T)\mid H=-0.5, Z=1, V}(y, t) & \text{if } T_0^* = T_1^*. \end{cases}$$

LEMMA 19. *If $E_P[TV_{(R,Y,T)\mid V}] > 0$, then* (i) *$P_L^*$ generates the data distribution P and the local average treatment effect under P_L^* is $E_P[\Delta E_P[Y \mid Z, V]]$, and* (ii) *$P_U^*$ generates the data distribution P and the local average treatment effect under P_U^* is $E_P[\Delta E_P[Y \mid Z]]/E_P[TV_{(R,Y,T)\mid V}]$.*

PROOF. (i) Denote by f^* the density function of P_L^*. The data generating process P_L^* generates the data distribution P:

$$f_{(R,Y,T)\mid Z=0, V}^*(r, y, t) = f_{(R,Y,T)\mid T^*=0, Z=0, V}^*(r, y, t)$$

$$= f_{(R_0, Y_0, T_0)\mid T^*=0, Z=0, V}^*(r, y, t)$$

$$= f_{(R,Y,T)\mid Z=0, V}(r, y, t),$$

$$f_{(R,Y,T)\mid Z=1, V}^*(r, y, t) = f_{(R,Y,T)\mid T^*=1, Z=1, V}^*(r, y, t)$$

$$= f_{(R_1, Y_1, T_1)\mid T^*=1, Z=1, V}^*(r, y, t)$$

$$= f_{(R,Y,T)\mid Z=1, V}(r, y, t),$$

where the first equality uses $T^* = Z$. Under P_L^*, the local average treatment effect is $E_P[\Delta E_P[Y \mid Z, V]]$:

$$
\begin{aligned}
E_{P_L^*}[Y_1 - Y_0 \mid T_0^* < T_1^*] &= E_{P_L^*}[Y_1] - E_{P_L^*}[Y_0] \\
&= E_P[E_P[Y \mid Z = 1, V]] - E_P[E_P[Y \mid Z = 0, V]] \\
&= E_P[\Delta E_P[Y \mid Z, V]].
\end{aligned}
$$

(ii) Note that

$$
\begin{aligned}
\Delta E_P[H \mid Z, V] &= \sum_{t=0,1} \iint \frac{|\Delta f_{(R,Y,T)|Z,V}(r,y,t)|}{2} \, d\mu_Y(y) \, d\mu_R(r) \\
&= TV_{(R,Y,T)|V}.
\end{aligned}
$$

Denote by f^* the density function of P_U^*. Notice that $f^*_{(R_{t^*}, Y_{t^*}, T_{t^*}) | (T_0^*, T_1^*, Z, V)}$ is well-defined because

$$
\Delta E_P[H \mid Z, V] = TV_{(R,Y,T)|V} > 0,
$$
$$
\Delta f_{(R,Y,T,H)|Z,V}(r,y,t,0.5) = \Delta f_{(R,Y,T)|Z,V}(r,y,t)\mathbb{1}\{\Delta f_{(R,Y,T)|Z,V}(r,y,t) \geq 0\} \geq 0,
$$
$$
\Delta f_{(R,Y,T,H)|Z,V}(r,y,t,-0.5) = \Delta f_{(R,Y,T)|Z,V}(r,y,t)\mathbb{1}\{\Delta f_{(R,Y,T)|Z,V}(r,y,t) < 0\} < 0.
$$

P_U^* generates the data distribution P:

$$
\begin{aligned}
f^*_{(R,Y,T)|Z=0,V}(r,y,t) &= P_U^*(T_0^* < T_1^* \mid Z = 0, V) f^*_{(R,Y,T)|T_0^* < T_1^*, Z=0, V}(r,y,t) \\
&\quad + P_U^*(T_1^* = T_0^* = 0 \mid Z = 0, V) f^*_{(R,Y,T)|T_1^* = T_0^* = 0, Z=0, V}(r,y,t) \\
&\quad + P_U^*(T_1^* = T_0^* = 1 \mid Z = 0, V) f^*_{(R,Y,T)|T_1^* = T_0^* = 1, Z=0, V}(r,y,t) \\
&= P_U^*(T_0^* < T_1^* \mid V) f^*_{(R_0, Y_0, T_0)|T_0^* < T_1^*, V}(r,y,t) \\
&\quad + P_U^*(T_1^* = T_0^* = 0 \mid V) f^*_{(R_0, Y_0, T_0)|T_1^* = T_0^* = 0, V}(r,y,t) \\
&\quad + P_U^*(T_1^* = T_0^* = 1 \mid V) f^*_{(R_1, Y_1, T_1)|T_1^* = T_0^* = 1, V}(r,y,t) \\
&= -\Delta E_P[H \mid Z, V] \frac{\Delta f_{(R,Y,T,H)|Z,V}(r,y,t,-0.5)}{\Delta E_P[H \mid Z, V]} \\
&\quad + P(H = -0.5 \mid Z = 1, V) f_{(R,Y,T)|H=-0.5, Z=1, V}(r,y,t) \\
&\quad + P(H = 0.5 \mid Z = 0, V) f_{(R,Y,T)|H=0.5, Z=0, V}(r,y,t) \\
&= -\big(f_{(R,Y,T,H)|Z=1,V}(r,y,t,-0.5) - f_{(R,Y,T,H)|Z=0,V}(r,y,t,-0.5)\big) \\
&\quad + f_{(R,Y,T,H)|Z=1,V}(r,y,t,-0.5) + f_{(R,Y,T,H)|Z=0,V}(r,y,t,0.5) \\
&= f_{(R,Y,T)|Z=0,V}(r,y,t),
\end{aligned}
$$

and similarly $f^*_{(R,Y,T)|Z=1,V}(r,y,t) = f_{(R,Y,T)|Z=1,V}(r,y,t)$. Under P^*_U, the local average treatment effect is

$$
\begin{aligned}
\theta(P^*_U) &= E_{P^*_U}\big[Y_1 - Y_0 \mid T^*_0 < T^*_1\big]\\[4pt]
&= E_{P^*_U}\big[E_{P^*_U}[Y_1 \mid T^*_0 < T^*_1, V] \mid T^*_0 < T^*_1\big]\\[2pt]
&\quad - E_{P^*_U}\big[E_{P^*_U}[Y_0 \mid T^*_0 < T^*_1, V] \mid T^*_0 < T^*_1\big]\\[4pt]
&= E_{P^*_U}\Bigg[\sum_{t=0,1}\iint y\,\frac{\Delta f_{(R,Y,T,H)|Z,V}(r,y,t,0.5)}{\Delta E_P[H \mid Z,V]}\,d\mu_Y(y)\,d\mu_R(r)\;\Big|\;T^*_0 < T^*_1\Bigg]\\[4pt]
&\quad + E_{P^*_U}\Bigg[\sum_{t=0,1}\iint y\,\frac{\Delta f_{(R,Y,T,H)|Z,V}(r,y,t,-0.5)}{\Delta E_P[H \mid Z,V]}\,d\mu_Y(y)\,d\mu_R(r)\;\Big|\;T^*_0 < T^*_1\Bigg]\\[4pt]
&= E_{P^*_U}\Bigg[\sum_{t=0,1}\iint y\,\frac{\Delta f_{(R,Y,T)|Z,V}(r,y,t)}{\Delta E_P[H \mid Z,V]}\,d\mu_Y(y)\,d\mu_R(r)\;\Big|\;T^*_0 < T^*_1\Bigg]\\[4pt]
&= E_{P^*_U}\Bigg[\sum_{t=0,1}\iint y\,\frac{\Delta f_{(R,Y,T)|Z,V}(r,y,t)}{\Delta E_P[H \mid Z,V]}\,d\mu_Y(y)\,d\mu_R(r)\,\frac{P^*_U(T^*_0 < T^*_1 \mid V)}{E_{P^*_U}[P^*_U(T^*_0 < T^*_1 \mid V)]}\Bigg]\\[4pt]
&= E_{P^*_U}\Bigg[\sum_{t=0,1}\iint y\,\frac{\Delta f_{(R,Y,T)|Z,V}(r,y,t)}{E_{P^*_U}[P^*_U(T^*_0 < T^*_1 \mid V)]}\,d\mu_Y(y)\,d\mu_R(r)\Bigg]\\[4pt]
&= \frac{E_{P^*_U}\Bigg[\displaystyle\sum_{t=0,1}\iint y\Delta f_{(R,Y,T)|Z,V}(r,y,t)\,d\mu_Y(y)\,d\mu_R(r)\Bigg]}{E_{P^*_U}[P^*_U(T^*_0 < T^*_1 \mid V)]}\\[6pt]
&= \frac{E_P\big[\Delta E_P[Y \mid Z,V]\big]}{E_P\big[\Delta E_P[H \mid Z,V]\big]}\\[6pt]
&= \frac{E_P\big[\Delta E_P[Y \mid Z,V]\big]}{E_P[TV_{(R,Y,T)|V}]},
\end{aligned}
$$

where the fifth equality comes from Bayes' theorem.　　　　　　　　　　□

Theorem 17 follows from the next lemma.

LEMMA 20. *If $E_P[TV_{(R,Y,T)|V}] > 0$, then, for every $\lambda \in [0,1]$, (i) the mixture distribution $\lambda P^*_L + (1-\lambda)P^*_U$ satisfies Assumption 1, (ii) $\lambda P^*_L + (1-\lambda)P^*_U$ generates the data distribution P, (iii) under $\lambda P^*_L + (1-\lambda)P^*_U$, the local average treatment effect is*

$$
\lambda E_P\big[\Delta E_P[Y \mid Z,V]\big] + (1-\lambda)\frac{E_P\big[\Delta E_P[Y \mid Z,V]\big]}{E_P[TV_{(R,Y,T)|V}]}.
$$

PROOF. (i) Under both P^*_L and P^*_U, Z is independent of $(R_{t^*}, T_{t^*}, Y_{t^*}, T^*_0, T^*_1)$ given V for each $t^* = 0, 1$. Furthermore, P^*_L and P^*_U have the same marginal distribution for (Z, V): $f_{Z,V}$. Therefore, the mixture of P^*_L and P^*_U also satisfies the independence. Since both P^*_L

and P_U^* satisfy $T_1^* \geq T_0^*$ almost surely, so does the mixture. (ii) By Lemma 19, both P_L^* and P_U^* generate the data distribution P and so does the mixture. (iii) It follows from the last statement in Lemma 19. □

Proof of Theorem 11

The proof of Theorems 11 is similar to Theorem 17. Only the difference is to change the definition of P_U^* as follows. Define $H = 0.5 \times \mathrm{sgn}(\Delta f_{Y|Z}(Y))$ and define P_U^* as

$$Z \sim P(Z = z),$$

$$\left(T_0^*, T_1^*\right) \mid Z = \begin{cases} (0,1) & \text{with probability } \Delta E_P[H \mid Z], \\ (0,0) & \text{with probability } P(H = -0.5 \mid Z = 1), \\ (1,1) & \text{with probability } P(H = 0.5 \mid Z = 0), \end{cases}$$

$$(Y_1, T_1) \mid \left(T_0^*, T_1^*, Z = z\right) \sim \begin{cases} f_{T|Y=y, Z=z}(t) \dfrac{\Delta f_{(Y,H)|Z}(y, 0.5)}{\Delta E_P[H \mid Z]} & \text{if } T_0^* < T_1^*, \\ f_{T|Y=y, Z=z}(t) f_{Y|H=0.5, Z=0}(y) & \text{if } T_0^* = T_1^*, \end{cases}$$

$$(Y_0, T_0) \mid \left(T_0^*, T_1^*, Z = z\right) \sim \begin{cases} -f_{T|Y=y, Z=z}(t) \dfrac{\Delta f_{(Y,H)|Z}(y, -0.5)}{\Delta E_P[H \mid Z]} & \text{if } T_0^* < T_1^*, \\ f_{T|Y=y, Z=z}(t) f_{Y|H=-0.5, Z=1}(y) & \text{if } T_0^* = T_1^*. \end{cases}$$

Proof of Lemma 12

First, note that

$$TV_{(Y,T)|V} = \sup_{h \in \mathbf{H}} \Delta E_P\big[h(Y, T, V) \mid Z, V\big]. \tag{14}$$

This is verified as follows. For every $h \in \mathbf{H}$, I have

$$\begin{aligned} &\Delta E_P\big[h(Y, T, V) \mid Z, V\big] \\ &\quad = E_P\big[h(Y, T, V) \mid Z = 1, V\big] - E_P\big[h(Y, T, V) \mid Z = 0, V\big] \\ &\quad = \sum_{t=0,1} \int h(y, t, V) \Delta f_{(Y,T)|Z,V}(y, t)\, d\mu_Y(y) \\ &\quad \leq \frac{1}{2} \sum_{t=0,1} \int \big|\Delta f_{(Y,T)|Z,V}(y, t)\big|\, d\mu_Y(y) \\ &\quad = TV_{(Y,T)|V}. \end{aligned}$$

Moreover, the above inequality becomes an equality when $h(y, t, v) = 0.5$ if and only if $\Delta f_{(Y,T)|Z,V=v}(y, t)$ is positive.

Abadie (2003) and Frölich (2007) showed that

$$E_P\big[\Delta E_P[X \mid Z, V]\big] = E_P\left[\frac{Z - \pi(V)}{\pi(V)\big(1 - \pi(V)\big)}X\right] \tag{15}$$

for any random variable X.[11]

By Theorem 17 and equation (14), $\Theta_I(P)$ is characterized by

$$|\theta|E_P\big[\Delta E_P\big[h(Y,T,V) \mid Z, V\big]\big] \le \big|E_P\big[\Delta E_P[Y \mid Z, V]\big]\big| \quad \text{for every } h \in \mathbf{H},$$

$$\theta E_P\big[\Delta E_P[Y \mid Z, V]\big] \ge 0,$$

$$|\theta| \ge \big|E_P\big[\Delta E_P[Y \mid Z, V]\big]\big|.$$

Since the second condition implies $\mathrm{sgn}(\theta) = \mathrm{sgn}(E_P[\Delta E_P[Y \mid Z, V]])$, the above three conditions become

$$|\theta|E_P\big[\Delta E_P\big[h(Y,T,V) \mid Z, V\big]\big] \le \mathrm{sgn}(\theta)E_P\big[\Delta E_P[Y \mid Z, V]\big] \quad \text{for every } h \in \mathbf{H},$$

$$\mathrm{sgn}(\theta)E_P\big[\Delta E_P[Y \mid Z, V]\big] \ge 0,$$

$$|\theta| \ge \mathrm{sgn}(\theta)E_P\big[\Delta E_P[Y \mid Z, V]\big].$$

By equation (15), $\Theta_I(P)$ is characterized as in Lemma 12.

Proof of Lemma 15

Condition (i) implies Assumption 14(i). Condition (ii) implies Assumption 14(iii). The rest of the proof is going to show Assumption 14(ii). Define $h_P^*(y, t, v) = 0.5 \times \mathrm{sgn}(\Delta f_{(Y,T)|Z,V=v}(y,t))$ and $h_{P,n}^* = \mathrm{argmax}_{h \in \mathbf{H}_n} E_P[\Delta E_P[h(Y,T) \mid Z, V]]$. Then $h_P^* =$

---

[11]The proof is as follows:

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)\big(1 - \pi(V)\big)}X\right] = E_P\left[\frac{\big(1 - \pi(V)\big)Z + (Z - 1)\pi(V)}{\pi(V)\big(1 - \pi(V)\big)}X\right]$$

$$= E_P\left[\frac{Z}{\pi(V)}X\right] - E_P\left[\frac{1 - Z}{\big(1 - \pi(V)\big)}X\right]$$

$$= E_P\left[\frac{E_P[ZX \mid V]}{\pi(V)}\right] - E_P\left[\frac{E_P\big[(1 - Z)X \mid V\big]}{\big(1 - \pi(V)\big)}\right]$$

$$= E_P\left[\frac{\pi(V)E_P[X \mid Z = 1, V]}{\pi(V)}\right]$$

$$\quad - E_P\left[\frac{\big(1 - \pi(V)\big)E_P[X \mid Z = 0, V]}{\big(1 - \pi(V)\big)}\right]$$

$$= E_P\big[E_P[X \mid Z = 1, V] - E_P[X \mid Z = 0, V]\big]$$

$$= E_P\big[\Delta E_P[X \mid Z, V]\big].$$

$\mathrm{argmax}_{h \in \mathbf{H}} E_P[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} h(Y, T, V)]$. By the Hölder continuity of $f_{(Y,T)|Z,V}$,

$$\max_{(y,t,v) \in I_{n,k}} \Delta f_{(Y,T)|Z,V=v}(y, t) - \min_{(y,t,v) \in I_{n,k}} \Delta f_{(Y,T)|Z,V=v}(y, t) \leq 2D_0 \left(2 \frac{D_1}{K_n} \right)^d.$$

Define $D_n = 2D_0 (2\frac{D_1}{K_n})^d$. For $I_{n,k}$ with $\sup_{(y,t,v) \in I_{n,k}} |\Delta f_{(Y,T)|Z,V=v}(y, t)| > D_n$, the above inequality implies that $\mathrm{sgn}(\Delta f_{(Y,T)|Z,V=v}(y, t))$ is constant on $I_{n,k}$. For those $I_{n,k}$, $h^* = h_n^*$ on $I_{n,k}$. Then, on every $I_{n,k}$, either $h^* = h_n^*$ or $|\Delta f_{(Y,T)|Z,V=v}(y, t)| \leq D_n$. Therefore,

$$h_{P,n}^*(y, t, v) \Delta f_{(Y,T)|Z,V=v}(y, t) \geq h_P^*(y, t, v) \Delta f_{(Y,T)|Z,V=v}(y, t) + 0.5 \times D_n.$$

Since

$$\begin{aligned}
& E_P \left[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} h(Y, T, V) \right] \\
&= E_P \left[\Delta E_P [h(Y, T, V) \mid Z, V] \right] \\
&= \iiint h(y, t, v) \Delta f_{(Y,T)|Z,V=v}(y, t) f_V(v) \mu_Y(dy) \mu_T(dt) \mu_V(dv) \\
&= \sum_{k=1}^{K_n} \iiint_{I_{n,k}} h(y, t, v) \Delta f_{(Y,T)|Z,V=v}(y, t) f_V(v) \mu_Y(dy) \mu_T(dt) \mu_V(dv),
\end{aligned}$$

it follows that

$$E_P \left[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} h_{P,n}^*(Y, T, V) \right] \geq E_P \left[\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))} h_P^*(Y, T, V) \right] + D_n.$$

Since D_n converges to zero uniformly over P, Assumption 14(ii) holds.

Proof of Theorem 16

The following theorem is taken from Corollary 5.1 and Theorem 6.1 in Chernozhukov, Chetverikov, and Kato (2014).

THEOREM 21. *Given $\varepsilon_n > 0$ with $\varepsilon_n \to 0$ and $\varepsilon_n \sqrt{\log p_n} \to \infty$, denote by $\mathcal{H}_{1,n}$ the set of $(\theta, \pi, P) \in \Theta \times \Pi \times \mathcal{P}_0$ for which $\pi = P(Z \mid V = \cdot)$ and*

$$\max_{j=1,\dots,p_n} m_j(\theta)/\sigma_j(\theta) \geq (1 + \varepsilon_n) \sqrt{2 \log(p_n/\alpha)/n}. \tag{16}$$

Under the assumptions in Theorem 16,

(i) $\displaystyle \liminf_{n \to \infty} \inf_{(\theta, \pi, P) \in \Theta \times \Pi \times \mathcal{P}_0 \text{ s.t. } \theta \in \Theta_I(P) \text{ and } \pi = P(Z|V)} P(T(\theta, \pi) \leq c(\alpha, \theta, \pi)) \geq 1 - \alpha;$

(ii) $\displaystyle \lim_{n \to \infty} \sup_{(\theta, P) \in \mathcal{H}_{1,n}} P(T(\theta, \pi) \leq c(\alpha, \theta, \pi)) = 0.$

Theorem 16(i) follows from

$$P\big(\theta \notin \mathcal{C}_{\theta,n}(\alpha + \delta)\big) \leq P\big(\theta \notin \mathcal{C}_{\theta,n}(\alpha + \delta) \& \pi \in \mathcal{C}_{\pi,n}(\delta)\big) + P\big(\pi \notin \mathcal{C}_{\pi,n}(\delta)\big)$$
$$\leq P\big(T(\theta, \pi) > c(\alpha, \theta, \pi) \& \pi \in \mathcal{C}_{\pi,n}(\delta)\big) + P\big(\pi \notin \mathcal{C}_{\pi,n}(\delta)\big)$$
$$= P\big(T(\theta, \pi) > c(\alpha, \theta, \pi)\big) + P\big(\pi \notin \mathcal{C}_{\pi,n}(\delta)\big) \leq \alpha + \delta,$$

where the last inequality comes from Theorem 21(i) and Assumption 13(iv).

Theorem 16(ii) is shown as follows. Denote by D_2 a constant for which $\sigma_j(\cdot) < D_2$. Let (θ, P) be any element of $\Theta \times \mathcal{P}_0$ with $\theta \notin \Theta_I(P)$. It suffices to show that (16) holds for sufficiently large n. If either (6) or (7) is violated, then (16) holds for sufficiently large n. In the rest of the proof, I focus on the case where (8) is violated. That is,

$$\sup_{h \in \mathbf{H}} E_P\left[\frac{Z - \pi(V)}{\pi(V)\big(1 - \pi(V)\big)}\big(|\theta|h(Y, T, V)\big) - \mathrm{sgn}(\theta)Y\right] > 0.$$

Since $\mathbf{H}_n$ converges to $\mathbf{H}$ in the sense of (12) and $\frac{Z - \pi(V)}{\pi(V)(1 - \pi(V))}|\theta|$ is bounded, it follows that, for sufficiently large n, there is $h \in \mathbf{H}_n$ such that

$$E_P\left[\frac{Z - \pi(V)}{\pi(V)\big(1 - \pi(V)\big)}\big(|\theta|h(Y, T, V) - \mathrm{sgn}(\theta)Y\big)\right] > 0.$$

Denoted by κ the value of the left-hand side in the above inequality. For sufficiently large n, there is $j = 3, \ldots, p_n$ such that $m_j(\theta) > \kappa$. For such j, $m_j(\theta)/\sigma_j(\theta) > \kappa/D_2$. Therefore, (16) holds for sufficiently large n.

References

Abadie, A. (2003), "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics*, 113, 231–263. [1338, 1347, 1353, 1355, 1365]

Aigner, D. J. (1973), "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics*, 1, 49–59. [1340]

Amemiya, Y. (1985), "Instrumental variable estimator for the nonlinear errors-in-variables model." *Journal of Econometrics*, 28, 273–289. [1341]

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996), "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association*, 91, 444–455. [1336, 1339, 1341, 1342]

Angrist, J. D. and A. B. Krueger (2001), "Instrumental variables and the search for identification: From supply and demand to natural experiments." *Journal of Economic Perspectives*, 15, 69–85. [1335]

Balke, A. and J. Pearl (1997), "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American Statistical Association*, 92, 1171–1176. [1346]

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. [1335, 1337, 1339, 1343]

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. [1340]

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007), "Changes in the distribution of male and female wages accounting for employment composition using bounds." *Econometrica*, 75, 323–363. [1341]

Bollinger, C. R. (1996), "Bounding mean regressions when a binary regressor is mismeasured." *Journal of Econometrics*, 73, 387–399. [1340]

Bound, J., C. Brown, and N. Mathiowetz (2001), "Measurement error in survey data." In *Handbook of Econometrics*, Vol. 5, Chapter 59 (J. Heckman and E. Leamer, eds.), 3705–3843, Elsevier. [1337, 1343]

Calvi, R., A. Lewbel, and D. Tommasi (2017), "LATE with mismeasured or misspecified treatment: An application to women's empowerment in India." Working paper. [1341]

Card, D. (2001), "Estimating the return to schooling: Progress on some persistent econometric problems." *Econometrica*, 69, 1127–1160. [1335, 1339]

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2012), *Measurement Error in Nonlinear Models: A Modern Perspective*, second edition. Chapman & Hall/CRC, Boca Raton. [1341]

Chalak, K. (2017), "Instrumental variables methods with heterogeneity and mismeasured instruments." *Econometric Theory*, 33, 69–104. [1341]

Chernozhukov, V., D. Chetverikov, and K. Kato (2014), "Testing many moment inequalities." Working paper. [1338, 1351, 1356, 1366]

de Chaisemartin, C. (2016), "Tolerating defiance? Local average treatment effects without monotonicity." Quantitative Economics. (forthcoming). [1343]

DiTraglia, F. J. and C. García-Jimeno (2015), "On mis-measured binary regressors: New results and some comments on the literature." Working paper. [1340]

Frazis, H. and M. A. Loewenstein (2003), "Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables." *Journal of Econometrics*, 117, 151–178. [1338, 1340]

Frölich, M. (2007), "Nonparametric IV estimation of local average treatment effects with covariates." *Journal of Econometrics*, 139, 35–75. [1338, 1347, 1365]

Gustman, A. L., T. Steinmeier, and N. Tabatabai (2008), "Do workers know about their pension plan type? Comparing workers' and employers' pension information." In *Overcoming the Saving Slump; how to Increase the Effectiveness of Financial Education and Saving Programs* (A. Lusardi, ed.), 47–81, University of Chicago Press, Chicago. [1338, 1353]

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. [1338, 1348]

Hausman, J. A., W. K. Newey, and J. L. Powell (1995), "Nonlinear errors in variables estimation of some engel curves." *Journal of Econometrics*, 65, 205–233. [1339]

Heckman, J. J. and E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation." *Econometrica*, 73, 669–738. [1346, 1356]

Henry, M., K. Jochmans, and B. Salanié (2015), "Inference on two-component mixtures under tail restrictions." Econometric Theory. (forthcoming). [1340]

Henry, M., Y. Kitamura, and B. Salanié (2014), "Partial identification of finite mixtures in econometric models." *Quantitative Economics*, 5, 123–144. [1340]

Hernandez, M. and S. Pudney (2007), "Measurement error in models of welfare participation." *Journal of Public Economics*, 91, 327–341. [1339]

Hsiao, C. (1989), "Consistent estimation for some nonlinear errors-in-variables models." *Journal of Econometrics*, 41, 159–185. [1341]

Hu, Y. (2008), "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics*, 144, 27–61. [1339, 1340]

Hu, Y., J.-L. Shiu, and T. Woutersen (2015), "Identification and estimation of single index models with measurement error and endogeneity." *The Econometrics Journal*, 18, 347–362. [1341]

Huber, M. and G. Mellace (2015), "Testing instrument validity for LATE identification based on inequality moment constraints." *Review of Economics and Statistics*, 97, 398–411. [1343, 1346]

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. [1340]

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. [1336]

Imbens, G. W. and J. D. Angrist (1994), "Identification and estimation of local average treatment effects." *Econometrica*, 62, 467–475. [1336, 1339, 1341, 1342, 1343, 1344, 1358]

Kaestner, R., T. Joyce, and H. Wehbeh (1996), "The effect of maternal drug use on birth weight: Measurement error in binary variables." *Economic Inquiry*, 34, 617–629. [1340]

Kane, T. J., C. E. Rouse, and D. Staiger (1999), "Estimating returns to schooling when schooling is misreported." Working paper. [1340]

Kitagawa, T. (2010), "Testing for instrument independence in the selection model." Working paper. [1341]

Kitagawa, T. (2015), "A test for instrument validity." *Econometrica*, 83, 2043–2063. [1346]

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. [1340]

Kreider, B., J. V. Pepper, C. Gundersen, and D. Jolliffe (2012), "Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported." *Journal of the American Statistical Association*, 107, 958–975. [1340]

Lewbel, A. (1998), "Semiparametric latent variable model estimation with endogenous or mismeasured regressors." *Econometrica*, 66, 105–121. [1341]

Lewbel, A. (2007), "Estimation of average treatment effects with misclassification." *Econometrica*, 75, 537–551. [1336, 1339, 1340]

Mahajan, A. (2006), "Identification and estimation of regression models with misclassification." *Econometrica*, 74, 631–665. [1336, 1339, 1340]

Manski, C. F. (2003), *Partial Identification of Probability Distributions*. Springer-Verlag, New York. [1341]

Molinari, F. (2008), "Partial identification of probability distributions with misclassified data." *Journal of Econometrics*, 144, 81–117. [1340]

Mourifié, I. Y. and Y. Wan (2016), "Testing local average treatment effect assumptions." Review of Economics and Statistics. (forthcoming). [1346]

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. [1353, 1355]

Schennach, S. M. (2013), "Measurement error in nonlinear models—a review." In *Advances in Economics and Econometrics: Economic Theory*, Vol. 3 (D. Acemoglu, M. Arellano, and E. Dekel, eds.), 296–337, Cambridge University Press. [1340]

Song, S. (2015), "Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors." *Journal of Econometrics*, 185, 95–109. [1341]

Song, S., S. M. Schennach, and H. White (2015), "Estimating nonseparable models with mismeasured endogenous variables on separable models with mismeasured endogenous variables." *Quantitative Economics*, 6, 749–794. [1341]

Ura, T. (2018), "Supplement to 'Heterogeneous treatment effects with mismeasured endogenous treatment'." *Quantitative Economics Supplemental Material*, 86, https://doi.org/10.3982/QE886. [1337, 1343]

Yanagi, T. (2017), "Inference on local average treatment effects for misclassified treatment." Working paper. [1341]