IDENTIFYING TREATMENT EFFECTS IN THE PRESENCE OF CONFOUNDED TYPES

DÉSIRÉ KÉDAGNI

The Pennsylvania State University - Department of Economics

Abstract. In this paper, I consider identification of treatment effects when the treatment is endogenous. The use of instrumental variables is a popular solution to deal with endogeneity, but this may give misleading answers when the instrument is invalid. I show that when the instrument is invalid due to correlation with the first stage unobserved heterogeneity, a second (also possibly invalid) instrument allows to partially identify not only the local average treatment effect but also the entire potential outcomes distributions for compliers. I exploit the fact that the distribution of the observed outcome in each group defined by the treatment and the instrument is a mixture of the distributions of interest. I write the identified set in the form of conditional moment inequalities, and provide an easily implementable inference procedure. Under some (testable) tail restrictions, the potential outcomes distributions are point-identified for compliers. Finally, I illustrate my methodology on data from the National Longitudinal Survey of Young Men to estimate returns to college using college proximity as (potential) instrument. I find that a 95% level confidence set for the average return to college for compliers is [38%, 79%], while a two-stage least squares estimate would erroneously yield an 872% return to college.

Keywords: Potential outcome, instrumental variable, LATE, compliers, mixture models.

JEL subject classification: C16, C21, C25, C26.

Date: The present version is of November 2, 2017. I am deeply grateful to my advisors Marc Henry and Ismael Mourifié for their constant guidance, inspiration, and encouragement. I am also grateful to Victor Aguirregabiria, Andres Aradillas-Lopez, Christian Bontemps, Irene Botosaru, Ronald Gallant, Michael Gechter, Patrik Guggenberger, Keisuke Hirano, Sung Jae Jun, Kala Krishna, Kari Lock Morgan, Karim Nchare, Joris Pinkse, Neil Wallace for their helpful comments and discussions, as well as comments from participants at the Cornell/Penn State macroeconomics conference 2017, seminar audiences at Penn State and University of Toronto. This is my **job market paper**. All errors are mine.

Email address: duk30@psu.edu.

1. Introduction

In the literature on returns to education, the use of instrumental variables (IV) is a popular solution to deal with the endogeneity of education choice. But, this may give misleading answers when the instrument is invalid. Growing up in a county with a college located in it raises attendance and has been used as an instrument for schooling. At best, this instrument will pick up the causal effect of college attendance on the *set of agents induced to go to college because they live close to it*, the *compliers*. One of the requirements for this to be a good instrument is that the distribution of ability and taste is the same for individuals who grew up close to colleges and those who did not. However, because agents select where they live based on their preferences, this assumption will not hold. Indeed, agents' taste for college is correlated with their parents' taste, which influenced where they lived. This makes the college proximity instrument invalid.

Suppose now that we use another instrument that also raises college attendance, say that both parents live in the household. This instrument may also be invalid: parents who choose to stay together may be more concerned about their children's college attendance. Also, family structure affects a child's cognitive ability, which influences her educational attainment. In this paper, I show how to (partially) identify causal parameters using variation from two (potentially) invalid instruments.

Here is the intuition behind my identification strategy. In my canonical example, the treatment variable college graduation and the instrument college proximity partition the population into four unobserved groups, called the types. We distinguish the individuals that go to college regardless of the presence of a college in their county of residence (always-takers), those who will not go no matter how close they are to a college (never-takers), people who would not go if they lived close to a college, but would go if they did not (defiers), and the compliers. Individuals who graduated from college and lived near a college are either compliers or always-takers. Hence, for individuals who graduated from college and lived near a college, the identified distribution of earnings is a mixture of the earnings distributions for compliers and always-takers, respectively. Under my assumptions, mixture weights depend on the additional instrument, but mixture component distributions do not. I can then use variation in mixture weights to derive sharp bounds on distributions.

Contribution of the paper. First, I show that with the help of a second (invalid) instrument, one can derive sharp bounds on the potential outcomes distributions for the compliers. Indeed, I show that each of these distributions is identified up to a scale parameter that is partially identified. Therefore, I derive bounds on the local quantile treatment effects (LQTE), the LATE, the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATUT), as well as the average treatment effect (ATE) on the whole population of interest. It is worth pointing out that if the two invalid instruments are binary, they may help each other in partially identifying the potential outcomes distributions for their respective compliers. Furthermore, I show that under some

testable tail restrictions, the potential outcomes distributions are point-identified for the compliers. Thus, the LATE and the LQTE are point-identified under these restrictions. To the best of my knowledge, it is the first time point-identification of the LATE is shown with an invalid instrument. I am not aware of a paper that shows point-identification of the LATE under a different set of assumptions other than the so-called LATE assumptions. Moreover, my paper is the first to relax full independence to conditional independence given type.

Second, I show that my results generalize to settings where the LATE monotonicity assumption does not hold. Relaxing monotonicity only increases the dimensionality of the parameters to be (partially) identified. This assumption states that there are no defiers. In the return to schooling example, it means there are no individuals who will not go to college when they live close to a college, but will go when they live far. While this assumption seems reasonable in this example, it could be too restrictive in some circumstances.

Third, I extend the model to account for the sample selection problem that arises in the return to schooling literature and many other settings, since not all individuals are working and the wage is only observed for those who are working (Heckman, 1979). The difficulty comes from the fact that the endogenous variable college education affects both the employment status and the wage. I partially identify the LATE and the LQTE for individuals who are always employed regardless of their college education and who are induced to go to college by a change in educational institutions. I also show how to allow for misclassified treatment as the schooling variable could be mismeasured.

Fourth, I show how inference on mixture distributions can be conducted using the intersection bounds framework of Chernozhukov, Lee and Rosen (2013) or Andrews and Shi (2013). One can therefore use the stata packages developed by Chernozhukov, Kim, Lee and Rosen (2015) or Andrews, Kim and Shi (2016) to construct a confidence set on the potential outcomes distributions for the compliers.

Finally, I illustrate my methodology on data from the National Longitudinal Survey of Young Men (NLSYM), previously used by Card (1995), to estimate returns to college education using college proximity as instrument. As in Ginther (2000), I use family structure (presence of both parents at home at age 14) as a second instrument, and I find that getting a college degree has a positive effect on the log hourly wage. I find that a college degree increases the average hourly wage of the compliers by 38–79%, while the two-stage least squares (2SLS) estimate is an 872% increase. These findings suggest that the college proximity instrument is invalid, but helps provide meaningful information about the causal effect of college degree on wages.

Related literature. Imbens and Angrist (1994) introduced the concept of LATE, the average treatment effect for individuals whose treatment status is influenced by a change in an instrument that satisfies the following so-called LATE assumptions: it is independent of all latent variables (potential outcomes and potential treatments), and the treatment is monotone in it (also known as

no-defiers assumption). The set of such individuals is called the compliers in the language of Angrist, Imbens and Rubin (1996). Under these LATE assumptions, Heckman and Vytlacil (1999, 2001, 2005) introduced in the presence of a continuous instrument the local instrumental variable (LIV) estimand to identify the marginal treatment effect (MTE), defined as the ATE for the subpopulation at the margin. Heckman, Tobias and Vytlacil (2001), Carneiro, Heckman and Vytlacil (2010, 2011), Carneiro and Lee (2009), among many others used the LIV estimator to study the return to college using multiple instruments like tuition fees and distance to college, which Card (2001) argues are invalid.

Angrist et al. (1996) discussed the sensitivity of the IV estimand to the LATE assumptions, and showed that whenever the monotonicity assumption is violated, it is equal to the LATE plus a bias that depends on the proportion of defiers: the smaller this proportion, the smaller the bias. Later on, de Chaisemartin (2017) has shown conditions (called the compliers-defiers: CD) that allow the presence of defiers in the population, and under which the IV estimand identifies the LATE for a specific subset of the compliers. Indeed, the CD conditions state that there exists a subpopulation of compliers that has the same proportion of individuals and the same LATE as the defiers.

The above two papers discuss relaxing the LATE monotonicity assumption. This paper studies the violation of the LATE independence assumption in the sense that the type is confounded. Moreover, it also discusses relaxing the LATE monotonicity. The paper also complements the work of Kitagawa (2015), Mourifié and Wan (2017), and Kédagni and Mourifié (2015) as it shows how one can relax the IV full independence assumption when the statistical tests derived in those papers reject the instrument validity. Thus, this article also fits in the imperfect instrument literature as it derives informative bounds on the causal parameters of interest when the instrument is invalid.

Nevo and Rosen (2012) characterized the identified set of the parameters of a single linear regression model in the presence of an endogenous regressor when the IV condition fails. They assumed that the IV has the same direction of correlation with the error as the endogenous regressor, but is less correlated with the error than is this regressor. On the other hand, Manski and Pepper (2000, 2009) derived bounds on the ATE under the monotone IV assumption that the expectation of each potential outcome variable conditional on the instrument is monotone. See also Altonji, Elder, and Taber (2005), Altonji, Conley, Elder, and Taber (2011), Conley, Hansen, and Rossi (2012), Hotz, Mullin, and Sanders (1997). Recently, Kédagni and Mourifié (2016) derived testable implications of the IV zero-covariance assumption, and showed that whenever an implication is rejected, the magnitude of its violation can help bound the ATE.

Lee (2009) showed that in the presence of sample selection, even with the aid of a randomized treatment with full compliance, researchers can only partially identify the average treatment effect for a subpopulation, the *always-employed*. Chen and Flores (2015) extended the model by allowing for noncompliance. They derived bounds on the average treatment effect for a subpopulation that they

called the *always-employed compliers*. Their papers maintain the unconfounded type assumption. In this article, I relax this assumption and still partially identify the average treatment effect for the always-employed compliers.

The identification approach that I develop in this paper relies on the existence of a second instrument. The use of a second instrument for identification purpose is not new in the literature. Mahajan (2006) used an additional instrument, which he called "instrument-like variable (ILV)," to nonparametrically identify the regression function in models with a misclassified binary regressor. Lewbel (2007) also used a second instrument to nonparametrically identify the ATE when the treatment is misclassified. Recently, Fricke et al. (2015) have developed a nonparametric method for evaluating treatment effects in the presence of both treatment endogeneity and attrition/non-response bias, using two instrumental variables. With the help of a discrete instrument for the treatment and a continuous instrument for non-response/attrition, they identify the LATE as well as the ATE under some restrictions. Note that all these papers need valid instruments, while I allow for invalid ones in my identification strategy. Kolesár et al. (2015) study identification and inference in the homogeneous effect setup in the presence of many invalid instruments that have direct effects on the outcome. They instead assume that the direct effects of the instruments are uncorrelated with their effects on the treatment. In this paper, I consider a different type of invalid instruments in the heterogeneous effects framework: the instruments are excluded from the outcome equation, but they are correlated with the first stage unobserved heterogeneity.

Finally, the main driver of my identification strategy is a mixture reformulation of the problem where mixture weights vary with the instrument, but mixture component distributions do not. Henry, Kitamura and Salanié (2014), and Jochmans, Henry and Salanié (2017) also consider nonparametric partial identification of finite mixtures with varying weights and fixed component distributions.

Outline. The remainder of the paper is organized as follows. Section 2 presents the model and discusses the problem. In Section 3, I discuss the main identification results. Section 4 discusses some empirical illustrations. Section 5 concludes. Proofs of the main results and extensions are discussed in the appendix.

2. Analytical Framework

Consider the following triangular system

$$\begin{cases} Y &= g(D,U) \\ D &= h(Z,W,V) \end{cases} \tag{2.1}$$

where Y is the outcome variable taking values in $\mathcal{Y} \subset \mathbb{R}$, D is a binary treatment variable, $Z \in \{0,1\}$ and $W \in \mathcal{W} \subset \mathbb{R}^{d_w}$ (d_w is the dimension of W) are observed variables excluded from the outcome

equation, U and V are unobserved heterogeneity variables whose dimensions are unrestricted. The functions g and h are unknown. Instead of assuming that Z is independent of the vector (U, V) as is usually the case, I assume that the vector of instruments (Z, W) is independent of U conditional on V, while Z and W can be both dependent on V. The goal of this paper is to (partially) identify traditional causal parameters: LATE, LQTE, ATE, ATT, ATUT. For the sake of clarity, I drop exogenous covariates from the model. All results derived in the paper hold conditionnally on covariates.

Here are some economic examples that support the above model.

Example 1 (**leading example**). *Consider the model of optimal schooling choice by individuals discussed by Card (2001). In this case, the variable Y denotes log earnings (or wage), D an indicator for college education, and Z a college proximity dummy variable. Card (2001) assumes that the relationship between the observed earnings and college education D takes the form*

$$Y = U_1 + U_2 D,$$

where there is heterogeneity in both the level of earnings of people without a college education U_1 and the return to schooling U_2. Assume further that the marginal cost of schooling conditional on $Z = z$, which subsumes all considerations other than the economic returns to education, is given by

$$C(z) = V_1 + V_2(1 - z),$$

where $V_2 > 0$ almost surely.

Suppose that it is optimal to go to college if the marginal return to college education is greater than its marginal cost plus an unobserved (dis)taste for schooling V_0 (which can be seen as a psychological cost/benefit), i.e.,

$$D = 1\{U_2 > C(Z) + V_0\}.$$

This is an extended Roy model with essential heterogeneity, in the terminology of Heckman, Urzua and Vytlacil (2006), in which agents make their choices based on the gain from treatment. In this example, U and V in (2.1) are $U = (U_1, U_2)$ and $V = (V_0, V_1, V_2, U_2)$. According to Card (2001), the exposure to educational institutions or college proximity Z is likely to be correlated with the unobserved schooling taste V_0, which in turn is likely to be correlated with ability factor U_2. As I illustrate in Appendix B.1, ability as measured by IQ appears to be affected by the college proximity instrument in the NLSYM data. Thus, the independence between college proximity (Z) and the unobserved heterogeneity (U, V) is unlikely to hold.

Although the college proximity instrument is controversial, it has been widely used in the literature: Card (1995), Kane and Rouse (1995), Kling (2001), Currie and Moretti (2003), Cameron and Taber (2004), and more recently Carneiro and Lee (2009), Carneiro, Heckman and Vytlacil (2011), among many others.

In this example, the variable W could be local earnings in the county of residence at age 17 (Cameron and Heckman, 1998; Cameron and Taber, 2004), local unemployment at 17 (Cameron and Heckman, 1998; Carneiro and Lee, 2009; Carneiro, Heckman and Vytlacil, 2011), local tuition in public four-year colleges at 17 (Kane and Rouse, 1995; Carneiro and Lee, 2009; Carneiro, Heckman and Vytlacil, 2011), presence of both parents at home, or school-quality measures such as teacher-to-student ratio, percent of teachers with advanced degrees, beginning teacher salaries (Ginther, 2000), etc.

The unobserved heterogeneity (U) in the earnings equation represents post-graduate shocks, and can then be interpreted as a function of shocks in the schooling decision (V) and exogenous shocks (ϵ) that are unrelated to all variables determining the schooling decision (Z, W, V). This interpretation supports the assumption that U is independent of (Z, W) given V. □

Example 2. *Suppose that a researcher wants to estimate the effect of a given change in prices (D) on the demand (Y) for a differentiated product. The variable U is an unobserved demand shock, Z a change in prices in the nearest market, W advertising, and V unobserved marginal cost shocks. If the variable V is a combination of common and idiosyncratic shocks, then prices in other markets are likely to be correlated with common shocks, implying that Z is potentially correlated with V. However, after conditioning on common shocks, prices in other markets are likely to be unrelated to unobserved demand shocks.* □

Example 3. *Suppose that a policy maker is interested in measuring a given change in current interest rate (D) on investment (Y). Since interest rate is endogenous, she uses lagged interest rate as instrument (Z). The variable W could be exchange rate or inflation. It could be the case that the lagged interest rate is endogenous. For example, lagged interest rate could be correlated with shocks (V) in the current interest rate. However, the policy maker can assume that conditional on those current shocks V, lagged interest rate is independent of investment shocks (U).* □

Example 4. *Suppose that a researcher wants to estimate the causal effect of smoking (D) on health outcome (Y) like heart disease, lung cancer, stroke, etc. She observes a network (N) of connections between agents and she assumes that the health outcome Y is a function of the smoking decision D, unobserved social characteristics V (which may include local network features, preference shocks, gregariousness) and some exogenous idiosyncratic health shock ϵ (genetic characteristics). By setting $U = (V, \epsilon)$, I have $Y = g(D, U)$. The researcher also assumes that the network N is a function of the social characteristics V and some exogenous idiosyncratic shock η so that $N = \varphi(V, \eta)$. Let Z be the indicator that an individual's "best" friend smokes and W be his number of friends (degree). The degree W depends on the agent's taste for having many friends (gregariousness) and probably so does her decision to smoke. In this model, the only variable that creates endogeneity is the social characteristics V. Therefore, conditioning on V would remove endogeneity. Thus, the assumption that $(Z, W) \perp\!\!\!\perp U|V$ is plausible in this example.* □

My identification methodology relies on the variation in V induced by W, so that we can mimic conditioning on V (which is infeasible) using W. However, given that the treatment variable D and the instrument Z are binary, we can work with a more general version of the structural model (2.1) known as the potential outcome model (POM):

$$\begin{cases} Y &= Y_1 D + Y_0(1-D) \\ D &= D_1 Z + D_0(1-Z) \end{cases} \tag{2.2}$$

where $Y_1 = g(1,U)$, $Y_0 = g(0,U)$, $D_1 = h(1,W,V)$, and $D_0 = h(0,W,V)$. The potential treatments D_0 and D_1 divide the population into four unobserved groups, commonly known as the *types* (as in Angrist, Imbens and Rubin, 1996) or *strata* (as in Frangakis and Rubin, 2002, which built on Hirano et al., 2000). The fact that the instrument Z is correlated with the unobservable V makes the types endogenous. The relationship between the treatment variable D and the instrument Z describes the types as follows:

- $D = 1$ for any Z: always-takers (a)
- $D = 0$ for any Z: never-takers (n)
- $D = Z$: compliers (c)
- $D = 1 - Z$: defiers (df)

Let T denote the random type of an individual with support $\{a, c, n, df\}$. For instance, $(D_0, D_1) = (1,1)$ means $T = a$.

Example 1 (continued)**.** *In this example, the always-takers are individuals that go to college regardless of the presence of a college in their county of residence at age 17. The compliers are those who go to college only because of the presence of a college in their county, while the never-takers are people who will not go no matter how close they are to a college. It is likely that there are no defiers in this example, that is, people who would not go to college if they lived in an area with a college, but they would go if they did not (this can be seen if distance to college is a cost as it is in this example).*

As I argued above, the type is confounded as Z is correlated with V. In this example, we have: $D_1 = 1\{U_2 > V_1 + V_0\}$ and $D_0 = 1\{U_2 > V_1 + V_2 + V_0\}$. The always-takers are the set of individuals for whom $Y_1 - Y_0 > V_1 + V_2 + V_0$, the never-takers are those such that $Y_1 - Y_0 \leq V_1 + V_0$, and the compliers are those that satisfy $V_1 + V_0 < Y_1 - Y_0 \leq V_1 + V_2 + V_0$. Therefore, the always-takers can be interpreted as "high" return people, the never-takers as "low" return ones, and the compliers can be seen as "marginal" individuals (See Figure 1). □

Since the population is partitioned into four types, the independence between the vector (Z, W) and the potential outcome Y_d $(d = 0, 1)$ conditional on the types is sufficient to derive my identification results. More precisely, I need the following assumption.

Assumption 1 (Conditional independence (CI))**.** *The vector (Z, W) is independent of Y_d given the type T, i.e., $(Z, W) \perp\!\!\!\perp Y_d \mid T$, for both $d = 0$ and $d = 1$.* □

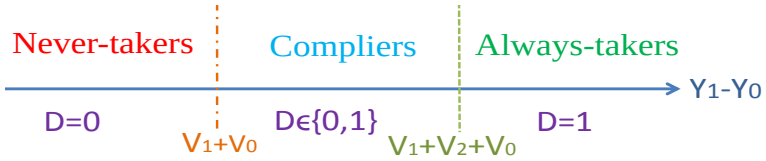

FIGURE 1. Types.

This is the main assumption in this paper. The following lemma shows that the assumption $(Z, W) \perp\!\!\!\perp U|V$ on the structural equation (2.1) is a sufficient condition for Assumption 1 to hold.

Lemma 1. *In the structural model (2.1), the following implication holds:*

$$(Z, W) \perp\!\!\!\perp U \mid V \Longrightarrow (Z, W) \perp\!\!\!\perp Y_d \mid T.$$

□

For the rest of the paper, I consider the potential outcome model described in Equation (2.2).

3. Identification results

3.1. **Main result.** In this subsection, I develop an identification strategy when the type is confounded, but the monotonicity assumption below (no defiers) still holds. I relax the monotonicity assumption in Appendix C.1. I show below that with the help of a second instrument (the variable I called W, which is also potentially invalid), one can derive sharp bounds on the distributions of the potential outcomes Y_0 and Y_1 for compliers. I then derive sharp bounds on the LATE, the LQTE, the ATT, the ATUT, as well as the ATE.

Assumption 2 (Monotonicity (MON))**.** *There are no defiers, i.e.,* $T \in \{a, n, c\}$. □

Notation 1. *Let* $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$ *be the measurable space for the outcome* Y*,* Ψ *the space of all probability distributions on* $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$*. Denote* $F_Y(. \mid X = x)$ *the conditional distribution of* Y *given* $X = x$*, and* $F_{dt}(.) \equiv F_{Y_d}(. \mid T = t)$*,* $d \in \{0, 1\}$*,* $t \in \{a, c, df, n\}$*. Let* $F(y|d, z, w)$ *be the cumulative distribution function (cdf) of* Y *conditional on* $(D = d, Z = z, W = w)$*, and* $F(y|d, z)$ *that of* Y *conditional on* $(D = d, Z = z)$*. Denote* $f(y|d, z, w)$ *and* $f(y|d, z)$ *the density or probability mass functions of* Y *conditional on* $(D = d, Z = z, W = w)$ *and* $(D = d, Z = z)$*, respectively, depending on whether* Y *is continuous, discrete or mixed. Define* $\alpha^1(w) \equiv \mathbb{P}(T = c \mid D = 1, Z = 1, W = w)$*,* $\alpha^0(w) \equiv \mathbb{P}(T = c \mid D = 0, Z = 0, W = w)$. □

My strategy in order to bound relevant parameters will be to start from sharp bounds on potential outcomes distributions. I observe the data (Y, D, Z, W) and want to identify the potential outcome distributions for the compliers, i.e., F_{1c} and F_{0c}. The main idea of my identification strategy is

the following: I first show that under CI (Assumption 1), the identified conditional distribution $F(y|1,1,w)$ is a mixture of two distributions of interest F_{1c} and F_{1a}, where only the weights depend on w, not the component distributions. Afterwards, I exploit variations in w to characterize the distributions F_{1c} and F_{1a} as functions of two parameters that are partially identified. Finally, I use the fact F_{1a} is point-identified under MON (Assumption 2) to pin down one of the two parameters. Thus, F_{1c} is identified up to a scale parameter that is partially identified. A similar reasoning using the conditional distribution $F(y|0,0,w)$ leads to partial identification of F_{0c}.

Before stating the results, I now give a heuristic derivation following the structure outlined above. Suppose that CI holds. Then, I have:

$$\begin{aligned} F(y|1,1,w) &= \mathbb{P}(T=c|D=1,Z=1,W=w) * F_{Y_1}(y|T=c,Z=1,W=w) \\ &\qquad +\mathbb{P}(T=a|D=1,Z=1,W=w) * F_{Y_1}(y|T=a,Z=1,W=w), \\ &= \mathbb{P}(T=c|D=1,Z=1,W=w) * F_{1c}(y) \\ &\qquad +\mathbb{P}(T=a|D=1,Z=1,W=w) * F_{1a}(y), \end{aligned}$$

where the first equality follows from the law of iterated expectations using the fact that $\mathbb{P}(T=n|D=1,Z=1,W=w)=0$ and $\mathbb{P}(T=df|D=1,Z=1,W=w)=0$, and the second holds under CI. Similar expressions hold for $F(y|0,0,w)$, $F(y|0,1,w)$, and $F(y|1,0,w)$. Hence, I have four mixtures with two components each.

In addition, assume that MON holds. Then any individual for whom $D=1$ and $Z=0$ is an always-taker, and anyone for whom $D=0$ and $Z=1$ is a never-taker. Therefore, under CI the distribution of Y_1 for the always-takers is point-identified by the distribution of observed earnings for the subgroup $(D=1,Z=0)$, i.e., $F_{1a}(y)=F(y|1,0)$, and that of Y_0 for the never-takers is point-identified by the distribution of observed earnings for the subgroup $(D=0,Z=1)$, i.e., $F_{0n}(y)=F(y|0,1)$.

The remaining two mixtures are

$$F(y|1,1,w) = \alpha^1(w)F_{1c}(y) + \left(1-\alpha^1(w)\right)F_{1a}(y), \tag{3.1}$$

$$F(y|0,0,w) = \alpha^0(w)F_{0c}(y) + \left(1-\alpha^0(w)\right)F_{0n}(y), \tag{3.2}$$

where $\alpha^1(w)$ and $\alpha^0(w)$ defined in Notation 1 are mixture weights. Let me consider Equation (3.1) first. Equation (3.2) is treated analogously. This is a two-component mixture model in which the mixture distributions $F_{1c}(y)$ and $F_{1a}(y)$ do not depend on w, while only the weight $\alpha^1(w)$ does. If for some $\tilde{w}$ in the support $\mathcal{W}$, $\alpha^1(\tilde{w})=1$, then $F_{1c}(y)$ is point-identified: $F_{1c}(y)=F(y|1,1,\tilde{w})$. This identification strategy is known as identification at infinity. Now, I am going to use a different approach, which does not rely on a large support assumption. I use variation in the second instrument W as follows:

$$F(y|1,1,w)-F(y|1,1,w') = \left[\alpha^1(w)-\alpha^1(w')\right]\left[F_{1c}(y)-F_{1a}(y)\right], \tag{3.3}$$

for all $w,\ w' \in \mathcal{W}$, and $y \in \mathcal{Y}$. I assume that the instrument W affects the treatment variable D so that there exist w_1^1 and w_0^1 in the support $\mathcal{W}$ such that $\alpha^1(w_1^1) \neq \alpha^1(w_0^1)$. In other words, W affects the proportion of compliers within the subgroup defined by $D = 1$ and $Z = 1$.

Notation 2. *Define* $\theta^1 \equiv 1/(\alpha^1(w_1^1) - \alpha^1(w_0^1))$, $\eta^1 \equiv \alpha^1(w_0^1)/(\alpha^1(w_1^1) - \alpha^1(w_0^1))$, *and* $\Lambda^1(w) \equiv (\alpha^1(w) - \alpha^1(w_0^1))/(\alpha^1(w_1^1) - \alpha^1(w_0^1))$; θ^0, η^0 *and* Λ^0 *are defined similarly.* □

Equations (3.1) and (3.3) imply that F_{1a} and F_{1c} are identified up to the two parameters θ^1 and η^1:

$$\begin{aligned} F_{1a}(y) &= F(y|1,1,w_0^1) - \eta^1 \left[F(y|1,1,w_1^1) - F(y|1,1,w_0^1)\right], \\ F_{1c}(y) &= F(y|1,1,w_0^1) + \left(\theta^1 - \eta^1\right)\left[F(y|1,1,w_1^1) - F(y|1,1,w_0^1)\right], \\ \alpha^1(w) &= \frac{1}{\theta^1}\left(\eta^1 + \Lambda^1(w)\right). \end{aligned} \tag{3.4}$$

At this point, I have shown that the potential outcomes distributions are nonparametrically identified up to scalars θ^1 and η^1. So, from now on, I will be interested in identifying these two parameters. I know by definition that θ^1 and η^1 have the same sign and θ^1 belongs to $(-\infty, -1] \cup [1, +\infty)$.

If $F(y|1,1,w_1^1) = F(y|1,1,w_0^1)$ for all y, then the distributions F_{1c} and F_{1a} are identical and therefore point-identified: $F_{1c}(y) = F_{1a}(y) = F(y|1,1)$. Consider now the case where for some $y_1^1 \in \mathcal{Y}$, $F(y_1^1|1,1,w_1^1) \neq F(y_1^1|1,1,w_0^1)$. Then $\Lambda^1(w)$ is identified as follows:

$$\Lambda^1(w) = \frac{F(y_1^1|1,1,w) - F(y_1^1|1,1,w_0^1)}{F(y_1^1|1,1,w_1^1) - F(y_1^1|1,1,w_0^1)}.$$

Let me use the fact that F_{1a} is point-identified. Then, η^1 is identified and

$$F_{1c}(y) = F(y|1,0) + \theta^1\left[F(y|1,1,w_1^1) - F(y|1,1,w_0^1)\right].$$

It is easy to see that the right-hand side of the above equality is right-continuous, has 0 as limit at $-\infty$ and 1 as limit at ∞. The only constraint that remains for this right-hand side to be a cdf is the following monotonicity condition:

$$f(y|1,0) + \theta^1\left[f(y|1,1,w_1^1) - f(y|1,1,w_0^1)\right] \geq 0, \tag{3.5}$$

where $f(y|d,z)$ and $f(y|d,z,w)$ are defined in Notation 1. Also, because the weight function $\alpha^1(w)$ is nonnegative and less than 1, we must have

$$0 \leq \frac{1}{\theta^1}\left(\eta^1 + \Lambda^1(w)\right) \leq 1 \tag{3.6}$$

for all $w \in \mathcal{W}$. At this point, all quantities in (3.4) are point identified except θ^1 and inequalities (3.5) and (3.6) characterize the identified set for θ^1, as stated below and formally proven in Appendix A.2.

Now, I am going to summarize the above discussion in a theorem. I state the following rank and relevance assumptions:

Assumption 3.a (Relevance (REL)). *For $d \in \{0,1\}$, there exist w_0^d and w_1^d in the support $\mathcal{W}$ such that $\alpha^d(w_1^d) \neq \alpha^d(w_0^d)$.* □

Assumption 3.b (Rank (RAN)). *For $d \in \{0,1\}$, there exist w_0^d and w_1^d in the support $\mathcal{W}$ and y_d^d in the support $\mathcal{Y}$ such that $F(y_d^d|d,d,w_0^d) \neq F(y_d^d|d,d,w_1^d)$.* □

Under the CI assumption, the rank condition RAN (Assumption 3.b) implies the relevance assumption REL (Assumption 3.a). RAN is a testable sufficient condition for REL.

Theorem 1. *Under CI and MON, the distribution of the potential outcome Y_1 is point-identified for the always-takers, while that of Y_0 is point-identified for the never-takers: $F_{1a}(y) = F(y|1,0)$ and $F_{0n}(y) = F(y|0,1)$.*

Under CI, MON and REL, the distribution of the potential outcome Y_d for the compliers satisfies the following: $F_{dc}(y) = F(y|d,1-d) + \theta^d \left[F(y|d,d,w_1^d) - F(y|d,d,w_0^d)\right]$, where θ^d is set-identified:

$$\theta^d \in \Theta^d = \begin{cases} \left[\theta_\ell^d, \theta_u^d\right] & \text{if RAN holds} \\ (-\infty,-1] \cup [1,+\infty) & \text{otherwise,} \end{cases}$$

with θ_ℓ^d and θ_u^d defined in Notation 3 in Appendix A.2. The set Θ^d is the (sharp) identified set. □

The theorem shows nonparametric partial identification of potential outcome distributions for compliers, where the latter are given in closed form as a scalar parameter family. As you can see from the closed form expression, F_{dc} is fixed once the value of θ^d is.

Comments.

(1) If assumptions CI, MON and REL hold and the RAN assumption does not, then the distribution of the potential outcome Y_d is point-identified for compliers: $F_{dc}(y) = F(y|d,d)$, and the following must hold: $F(y|d,d) = F(y|d,1-d)$.

(2) If either Θ^1 or Θ^0 is empty, then at least one of the assumptions CI, MON and REL is violated.

(3) If the type is unconfounded conditional on W, i.e., $Z \perp\!\!\!\perp T|W$, the weight $\alpha^1(w)$ is identified, and so is the parameter θ^1 since $\theta^1 = \left[\alpha^1(w_1^1) - \alpha^1(w_0^1)\right]^{-1}$. Therefore, the distribution $F_{1c}(y)$ is also identified. In this case, one can test this conditional unconfoundedness type assumption by checking whether the point-identified θ^1 lies within the identified set Θ^1. More generally, the unconfoundedness assumption $Z \perp\!\!\!\perp T$ is testable. Indeed, under this assumption combined with CI, MON and REL, the distributions F_{1c} and F_{0c} are point-identified. Therefore, these point-identified distributions must lie within their respective identified sets.

I transform Θ^d into an equivalent moment inequality model. It is more conducive to inference. For the sake of clarity of exposition, suppose first that the second instrument W is discrete. Define $c_0 = 1/\mathbb{E}[D(1-Z)]$, $c_1 = 1/\mathbb{E}[DZ\mathbb{1}\left\{W = w_1^1\right\}]$, and $c_2 = 1/\mathbb{E}[DZ\mathbb{1}\left\{W = w_0^1\right\}]$. The following corollary holds.

Corollary 1. *Under CI, MON and RAN, the identified set Θ^1 is equal to the set of θ^1 satisfying:*

$$\begin{cases} \inf_{y\in\mathcal{Y}} \mathbb{E}[m_0^1(\theta^1, D, Z, W)|Y = y] & \geq & 0 \\ \inf_{w\in\mathcal{W}} \mathbb{E}[m_1^1(\theta^1, Y)|D = 1, Z = 1, W = w] & \geq & 0 \end{cases} \tag{3.7}$$

where $m_0^1(\theta^1, D, Z, W) = c_0 D(1-Z) + \theta^1\left(c_1 DZ\mathbb{1}\left\{W = w_1^1\right\} - c_2 DZ\mathbb{1}\left\{W = w_0^1\right\}\right)$ *and*

$$m_1^1(\theta^1, Y) = \begin{bmatrix} sign(\theta^1)\left(\theta^1 - \frac{\mathbb{1}\left\{Y\leq y_1^1\right\} - \mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=1,W=w_0^1]}\right) \\ sign(\theta^1)\frac{\mathbb{1}\left\{Y\leq y_1^1\right\} - \mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\left\{Y\leq y_1^1\right\}|D=1,Z=1,W=w_0^1]} \end{bmatrix},$$

and symmetrically for Θ^0. □

Comments. When W is continuous, the event $\left\{W = w_\ell^1\right\}$ $(\ell = 0, 1)$ in the above derivation could be replaced by the event $\left\{W \in A_\ell^1\right\}$, where A_ℓ^1 is any mesurable set such that $F(y_1^1|1, 1, A_0^1) - F(y_1^1|1, 1, A_1^1) \neq 0$ and $\mathbb{P}(W \in A_\ell^1) > 0$.

Inference can be performed using the intersection bounds framework of Chernozhukov, Lee and Rosen (2013, CLR). One can use the stata packages of Chernozhukov, Kim, Lee and Rosen (2015, CKLR) or Andrews, Kim and Shi (2016) built on Andrews and Shi (2013).

Throughout the rest of the paper, I use the following example to illustrate my results.

Numerical illustration. I specify the joint distribution $p(z, w)$ of the instruments Z and W (both binary),

$$p(1,1) = 0.3, \quad p(1,0) = 0.2, \quad p(0,1) = 0.2,$$

the conditional distribution of the types $p(t|z, w)$,

$$\begin{array}{llllll} p(a|1,1) = 0.2, & p(c|1,1) = 0.5, & p(n|1,1) = 0.3, \\ p(a|1,0) = 0.1, & p(c|1,0) = 0.4, & p(n|1,0) = 0.5, \\ p(a|0,1) = 0.4, & p(c|0,1) = 0.2, & p(n|0,1) = 0.4, \\ p(a|0,0) = 0.35, & p(c|0,0) = 0.45, & p(n|0,0) = 0.20, \end{array}$$

and the potential outcomes distributions conditional on the types and the instruments $F_d(y|t, z, w)$ over the support $\mathcal{Y} = [0, \infty)$,

$$\begin{array}{llllll} F_1(y|a,z,w) = \frac{y}{y+1}, & F_1(y|c,z,w) = \frac{y^2}{y^2+1}, & F_1(y|n,z,w) = \frac{y^3}{y^3+1}, \\ F_0(y|a,z,w) = \frac{y^2}{y^2+1}, & F_0(y|c,z,w) = \frac{y^3}{y^3+1}, & F_0(y|n,z,w) = \frac{y^{3/2}}{y^{3/2}+1}. \end{array}$$

More details about this data generating process (DGP) are shown in Appendix A.9.

It can be shown that the type is confounded. For instance, $\mathbb{P}(T = c|Z = 1) \neq \mathbb{P}(T = c|Z = 0)$. It is easy to see that CI holds. I have an uncountable number of distributions F_{1c} and F_{0c}. I discretize the identified sets Θ^1 and Θ^0. Figure 2 displays bounds on distributions of the potential outcomes Y_0 and Y_1 for compliers. As you can see, even with binary instruments, the bounds on the compliers' potential outcomes distributions are tight.

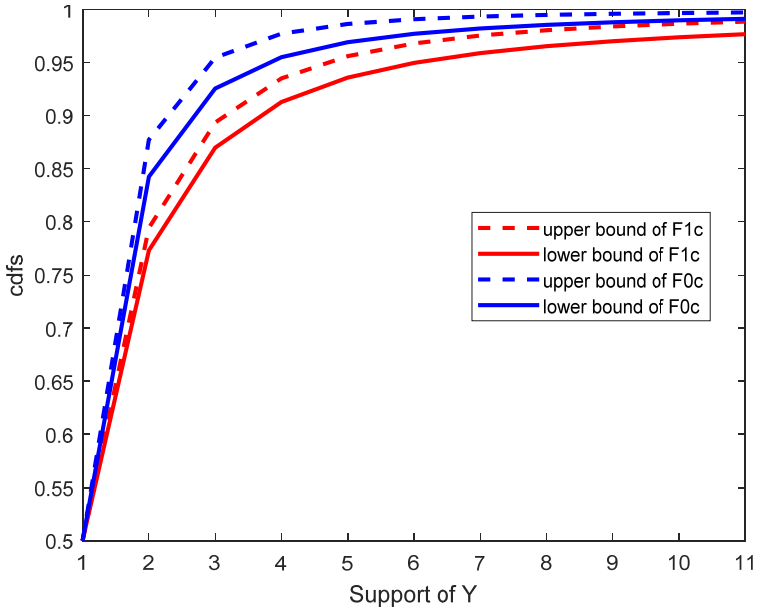

FIGURE 2. Bounds on the distributions F_{1c} and F_{0c}.

I obtain nontrivial bounds on expectations even without bounded support assumption on the outcome Y. Indeed, under RAN, the identified set Θ^d for the parameter θ^d is compact, and the distribution $F_{dc}(y)$ is linear in θ^d, hence continuous in θ^d. Now, I am going to derive bounds on other commonly used parameters of interest in the treatment effect literature.

3.2. **Sharp bounds on LATE and LQTE.** In this subsection, I derive bounds on the local average treatment effect, $LATE = \mathbb{E}[Y_1 - Y_0|T = c]$, and the local quantile treatment effects for compliers, $LQTE(p) = F_{1c}^{-1}(p) - F_{0c}^{-1}(p)$, for $p \in (0, 1)$, where F_{dc}^{-1} denotes the generalized inverse of the cdf F_{dc}. Before I proceed, I explain why the standard IV can no longer be interpreted as the LATE in the framework of this paper.

Under the LATE assumptions (unconfounded type: $Z \perp\!\!\!\perp T$, conditional independence: $Z \perp\!\!\!\perp Y_d|T$, and monotonicity: no defiers), the distributions of the potential outcomes Y_0 and Y_1 are point-identified for compliers, and LATE is identified and equal to the standard IV estimand α_{IV}.

When the type is confounded, the IV estimand may become a biased estimator for LATE (even asymptotically). LATE retains a causal interpretation while the IV estimand α_{IV} does not. See Appendix A.10 for details.

I now show the bounds that I derive for the LATE and the LQTE. As in the previous subsection, Θ^d, $d \in \{0,1\}$, is the sharp identified set for θ^d. Each parameter θ^d in Θ^d uniquely characterizes a distribution function in the identified set for F_{dc}, which I denote $F_{dc}^{\theta^d}$. I was interested in identifying a distribution function, which is an infinite-dimensional object. Now, I turn this into a scalar parameter identification problem. This helps derive closed-form expressions for the bounds on the LATE and the LQTE. Let $\mu_{dc}^{\theta^d}$ denote the expectation of the distribution $F_{dc}^{\theta^d}$. Then the following proposition holds.

Proposition 1. *Under CI, MON and REL, we have the following bounds for the average and quantile treatment effects for compliers:*

$$\inf_{\theta^1 \in \Theta^1} \mu_{1c}^{\theta^1} - \sup_{\theta^0 \in \Theta^0} \mu_{0c}^{\theta^0} \leq \mathbb{E}\left[Y_1 - Y_0 | T = c\right] \leq \sup_{\theta^1 \in \Theta^1} \mu_{1c}^{\theta^1} - \inf_{\theta^0 \in \Theta^0} \mu_{0c}^{\theta^0},$$

and

$$\inf_{\theta^1 \in \Theta^1} \left(F_{1c}^{\theta^1}\right)^{-1}(p) - \sup_{\theta^0 \in \Theta^0} \left(F_{0c}^{\theta^0}\right)^{-1}(p) \leq \left(F_{1c}^{-1} - F_{0c}^{-1}\right)(p) \leq \sup_{\theta^1 \in \Theta^1} \left(F_{1c}^{\theta^1}\right)^{-1}(p) - \inf_{\theta^0 \in \Theta^0} \left(F_{0c}^{\theta^0}\right)^{-1}(p),$$

for all $p \in (0,1)$, where $\left(F_{dc}^{\theta^d}\right)^{-1}$ denotes the generalized inverse of the cdf $F_{dc}^{\theta^d}$ $(d = 0,1)$, and Θ^d is defined in Theorem 1.

These bounds are sharp. □

Comments. Proposition 1 provides bounds on the average treatment effect for the subpopulation whose treatment status is affected by the instrument Z, namely the compliers. It is not always a parameter of interest for a policy maker. Below, I explain that the LATE may be of interest in Example 1.

Example 1 (continued)**.** *Suppose that a policy maker notices that the average income is low in some region of the country, and thinks that one possible reason could be the fact that most people in this region did not get a college degree. Then, she wants to build a campus in this region to encourage people to go to college. In this case, she might be interested in the effect of college degree on the earnings of individuals who obtained a degree only because they lived close to college (i.e., the LATE).* □

Numerical illustration (continued). In the numerical example, the identified set for the LATE is $[0.18, 0.92]$. The IV estimand is -1.82 (negative), while the actual LATE is 0.38 (positive). As can be seen, the IV estimand does not belong to the identified set of the LATE, meaning the LATE assumptions are violated and that the IV estimand has no causal interpretation. A researcher that ignores the fact that the type is confounded would make the wrong inference that the causal effect

is negative while it is, in fact, positive. An analyst that takes this information into account and uses my methodology will accurately infer the direction of the effect of the treatment on the outcome (as the lower bound is positive).

Now, I change the conditional probabilities of the type T so that it is unconfounded given W:

$$\begin{array}{llllllll} p(a|1,1) & = & 0.2, & p(c|1,1) & = & 0.5, & p(n|1,1) & = & 0.3, \\ p(a|1,0) & = & 0.1, & p(c|1,0) & = & 0.4, & p(n|1,0) & = & 0.5, \\ p(a|0,1) & = & 0.2, & p(c|0,1) & = & 0.5, & p(n|0,1) & = & 0.3, \\ p(a|0,0) & = & 0.1, & p(c|0,0) & = & 0.4, & p(n|0,0) & = & 0.5. \end{array}$$

The IV estimand is now 0.38, which is exactly the LATE. The sharp bounds on the LATE are $[0.21, 1.01]$. The sign of the causal effect is still identified in this case using my identification approach, which does not use the information that type is unconfounded.

3.3. **Sharp bounds on the ATT, the ATUT and the ATE.** In this section, I show that the counterfactual distributions $F_{Y_0}(y|D=1,W=w)$ and $F_{Y_1}(y|D=0,W=w)$ are partially identified. From there, I derive sharp bounds on the ATT, the ATUT, and the ATE conditional on W. I summarize the results in the following proposition.

Proposition 2. *Under CI, MON and REL, the identified sets for the counterfactual distributions are defined by:*

$$\begin{aligned} F_{Y_0}(y|D=1,W=w) &= \left[\mathbb{P}(Z=0|D=1,W=w)+\mathbb{P}(Z=1|D=1,W=w)(1-\alpha^1(w))\right]\psi^0(y) \\ &\quad + \mathbb{P}(Z=1|D=1,W=w)\alpha^1(w)F_{0c}^{\theta^0}(y), \\ F_{Y_1}(y|D=0,W=w) &= \left[\mathbb{P}(Z=1|D=0,W=w)+\mathbb{P}(Z=0|D=0,W=w)\alpha^0(w)\right]\psi^1(y) \\ &\quad + \mathbb{P}(Z=0|D=0,W=w)(1-\alpha^0(w))F_{1c}^{\theta^1}(y), \end{aligned}$$

where $\psi^d \in \Psi$, $\alpha^1(w)$ is defined in Equation (3.4), $\alpha^0(w)$ is defined similarly, and $\theta^d \in \Theta^d$, $d=0,1$.

These bounds are sharp and bounds on the conditional potential outcomes distributions of Y_0 and Y_1 are obtained trivially.

□

Bounds on $F_{Y_1}(y)$ and $F_{Y_0}(y)$ can be obtained by integrating out bounds on $F_{Y_1}(y|W=w)$ and $F_{Y_0}(y|W=w)$ with respect to the distribution of W, respectively.

Unlike the case of $LATE$, we now need bounds on the support of Y for nontrivial bounds on ATT, $ATUT$, and ATE. Denote $\mathcal{Y} = \left[y^\ell, y^u\right]$ the support of the outcome Y, where y^ℓ and y^u are lower and upper bounds of the support, respectively. Note that y^ℓ and y^u could be infinite. I partially identify the average treatment effect on the treated $ATT(w) = \mathbb{E}[Y|D=1,W=w] - \mathbb{E}[Y_0|D=1,W=w]$, the average treatment effect on the untreated $ATUT(w) = \mathbb{E}[Y_1|D=0,W=w] - \mathbb{E}[Y|D=0,W=w]$, and the average treatment effect $ATE(w) = \mathbb{P}(D=1|W=w)ATT(w) + \mathbb{P}(D=0|W=w)ATUT(w)$.

On can easily obtain bounds on ATT and $ATUT$ by integrating out the bounds on $ATT(w)$ and $ATUT(w)$ with respect to the distribution of W. Bounds on ATE are then obtained from those on ATT and $ATUT$. The following proposition holds.

Corollary 2. *Under CI, MON and REL, the ATT, the identified sets for the ATUT and the ATE are defined by:*

$$\begin{aligned} ATT(w) &= \mathbb{E}[Y|D=1,W=w] - \left[\mathbb{P}(Z=0|D=1,W=w) + \mathbb{P}(Z=1|D=1,W=w)(1-\alpha^1(w))\right]\delta^1 \\ &\quad -\mathbb{P}(Z=1|D=1,W=w)\alpha^1(w)\mu_{0c}^{\theta^0}, \\ ATUT(w) &= \left[\mathbb{P}(Z=1|D=0,W=w) + \mathbb{P}(Z=0|D=0,W=w)\alpha^0(w)\right]\delta^0 \\ &\quad +\mathbb{P}(Z=0|D=0,W=w)(1-\alpha^0(w))\mu_{1c}^{\theta^1} - \mathbb{E}[Y|D=0,W=w], \end{aligned}$$

where $\theta^d \in \Theta^d$ *and* $\delta^d \in \left[y^\ell, y^u\right]$, $d=0,1$.

These bounds are sharp. □

3.4. **Point-identification under tail restrictions.** So far, I have shown under assumptions CI, MON and REL that each of the distributions F_{0c} and F_{1c} is identified up to a parameter for which I derived sharp bounds. I need one more constraint to point-identify this parameter. I show that under the following tail restrictions, the potential outcomes distributions F_{0c} and F_{1c} for the compliers are point-identified.

Assumption 3 (Tail restrictions (TR))**.** $\lim_{y\downarrow y^\ell} \frac{F_{0c}(y)}{F_{0n}(y)} = 0$ *and* $\lim_{y\uparrow y^u} \frac{1-F_{1c}(y)}{1-F_{1a}(y)} = 0$*, where* y^ℓ *and* y^u *are lower and upper bounds of the support* $\mathcal{Y}$*, repectively.* □

Example 1 (continued)**.** *In this example, an interpretation of the assumption* $\lim_{y\uparrow y^u} \frac{1-F_{1c}(y)}{1-F_{1a}(y)} = 0$ *is that among people who went to college, the high earners among the always-takers (high return individuals) earn an order of magnitude more than high earners among the compliers (marginal individuals).* □

Numerical illustration (continued). The DGP in the numerical illustration satisfies this assumption. For instance, $\lim_{y\uparrow\infty} \frac{1-F_{1c}(y)}{1-F_{1a}(y)} = \lim_{y\uparrow\infty} \frac{f_{1c}(y)}{f_{1a}(y)} = 0$: f_{1c} goes to zero faster than f_{1a} as seen in Figure 3.

The first constraint imposed by Assumption TR (Assumption 3) fixes θ^0 and the second pins down θ^1. The following proposition summarizes these results.

Theorem 2. *Under CI, MON, REL and TR, the distributions* $F_{1c}(y)$ *and* $F_{0c}(y)$ *are point-identified as follows:*

$$\begin{aligned} F_{0c}(y) &= F(y|0,0,w_0^0) + \tfrac{1}{1-\zeta^0(w_1^0,w_0^1)}\left[F(y|0,0,w_1^0) - F(y|0,0,w_0^0)\right], \\ F_{1c}(y) &= F(y|1,1,w_0^1) + \tfrac{1}{1-\pi^1(w_1^1,w_0^1)}\left[F(y|1,1,w_1^1) - F(y|1,1,w_0^1)\right], \end{aligned} \tag{3.8}$$

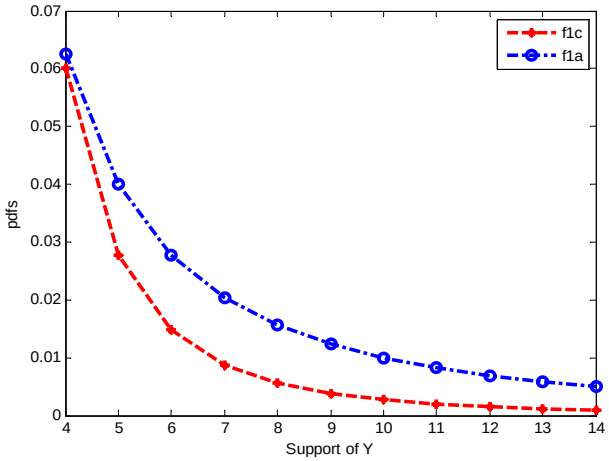

FIGURE 3. Density plots.

where

$$\zeta^0(w_1^0, w_0^0) = \lim_{y \downarrow y^\ell} \frac{F(y|0,0,w_1^0)}{F(y|0,0,w_0^0)}, \quad \textit{and} \quad \pi^1(w_1^1, w_0^1) = \lim_{y \uparrow y^u} \frac{1 - F(y|1,1,w_1^1)}{1 - F(y|1,1,w_0^1)}.$$

□

Comments. From Theorem 2, the LATE and the LQTE are point-identified under the additional TR assumption. Note that the monotonicity assumption is not necessary for this theorem to hold. However, the interpretation I give for the TR assumption makes more sense under this monotonicity assumption.

Assumption TR is testable under the maintained assumptions CI, MON and REL. Indeed, under these four assumptions, the results of Theorem 2 as well as those of Theorem 1 hold. Hence, the parameters θ^1 and θ^0 are point-identified, which should lie within the identified set Θ^0 and Θ^1 defined in Theorem 1.

3.4.1. *Bounds on ATT, ATUT and ATE under conditional monotone treatment response.* Under Assumption TR, the ATT, the ATUT and the ATE are still partially identified. The bounds on these parameters are the same as those derived in Corollary 2, except that θ^d and $\alpha^d(w)$ $(d = 0, 1)$ are point-identified. These bounds could still be wide depending on the application. I discuss below some assumptions under which the bounds can be tightened.

Assumption 4 (CMTR)**.** *For each type* $t \in \{a, c, n\}$, $\mathbb{E}[Y_1|T = t] \geq \mathbb{E}[Y_0|T = t]$. □

This assumption is identical to Chen, Flores and Flores-Lagunes's (2012) monotonicity in D of average outcomes of types, except that I assume here that the direction of the monotonicity is known. It is a weak version of Manski's (1997) monotone treatment response (MTR): $Y_1 \geq Y_0$. For

instance, in the leading example 1, if $V_1 + V_0 > 0$, the Manski (1997) MTR is a sufficient condition for the CMTR assumption to hold. In this example, the CMTR assumption is equivalent to saying that the average gross return to college is positive for each type.

Under TR, $\alpha^1(w)$, $\alpha^0(w)$, F_{0c} and F_{1c} are identified, as are θ^1 and θ^0. Assumption CMTR implies

$$\begin{aligned} \mathbb{E}[Y_0|T=a] &\leq \mathbb{E}[Y_1|T=a] = \mathbb{E}[Y|D=1,Z=0], \\ \mathbb{E}[Y_1|T=n] &\geq \mathbb{E}[Y_0|T=n] = \mathbb{E}[Y|D=0,Z=1]. \end{aligned}$$

Therefore, the following proposition holds:

Proposition 3. *Under CI, MON, REL, TR and CMTR, tighter bounds on ATT and ATUT are obtained from nontrivial bounds on the parameters δ^1 and δ^0 in Corollary 2: $\delta^1 \in \left[y^\ell, \mathbb{E}[Y|D=1,Z=0]\right]$ and $\delta^0 \in [\mathbb{E}[Y|D=0,Z=1], y^u]$.* □

The proof of this proposition is straightforward from Proposition 2 and is therefore omitted. An implication of the CMTR assumption is $\mathbb{E}[F_{0c}] \leq \mathbb{E}[F_{1c}]$, which is testable in this case since F_{0c} and F_{1c} are identified. The CMTR assumption tightens the bounds on ATT and $ATUT$ by reducing the upper bound for δ^1 and increasing the lower bound for δ^0.

3.4.2. *Bounds on ATT, ATUT and ATE under conditional monotone treatment selection.* Another type of additional restriction introduced by Manski and Pepper (2000) is the monotone treatment selection (MTS). It states that $\mathbb{E}[Y_d|D=1] \geq \mathbb{E}[Y_d|D=0]$. Notice that $\{D=1\} = \{a,c\}$ and $\{D=0\} = \{n,c\}$. Instead of assuming MTS over the sets $\{a,c\}$ and $\{n,c\}$, I assume it only over the sets a and n.

Assumption 5 (CMTS)**.** $\mathbb{E}[Y_d|T=a] \geq \mathbb{E}[Y_d|T=n] \ \forall \ d \in \{0,1\}$. □

This assumption has also been considered in Chen et al. (2012), which they called a mean dominance assumption. In the leading example 1, this assumption is equivalent to saying that the average earnings level of people with the highest gross returns is no less than that of people with the lowest gross returns.

Assumptions CI, MON and CMTS imply

$$\begin{aligned} \mathbb{E}[Y_1|T=n] &\leq \mathbb{E}[Y_1|T=a] = \mathbb{E}[Y|D=1,Z=0], \\ \mathbb{E}[Y_0|T=a] &\geq \mathbb{E}[Y_0|T=n] = \mathbb{E}[Y|D=0,Z=1]. \end{aligned}$$

The following proposition holds.

Proposition 4. *Under CI, MON, REL, TR and CMTS, tighter bounds on ATT and ATUT are obtained from nontrivial bounds on the parameters δ^1 and δ^0 in Corollary 2: $\delta^1 \in [\mathbb{E}[Y|D=0,Z=1], y^u]$ and $\delta^0 \in \left[y^\ell, \mathbb{E}[Y|D=1,Z=0]\right]$.* □

The CMTS assumption tightens the bounds on ATT and $ATUT$ by increasing the lower bound for δ^1 and reducing the upper bound for δ^0.

Remark 1. *The CMTR and CMTS assumptions together imply tighter bounds on δ^1 and δ^0: $\delta^1, \delta^0 \in [\mathbb{E}[Y|D=0,Z=1], \mathbb{E}[Y|D=1,Z=0]]$.* □

4. Empirical illustration: Returns to college

4.1. **Data.** In this application, I use Card's (1995) data set. The data are drawn from the NLSYM. The Young Men's cohort includes 5,225 men who were ages 14-24 when first interviewed in 1966, with follow-up surveys through 1981. Due to sample attrition, Card (1995) used labor market information from the 1976 interview. The 1976 sample represents 71 percent of the original sample and has the advantage that all respondents were directly asked their educational attainment during the interview.

The outcome variable Y is the log hourly wage (lwage) while the treatment variable D is the indicator that the individual has a four-year college degree (college). The presence of a four-year college in the county of residence (nearc4) is the instrument Z and the family structure: presence of both parents at home at age 14 (momdad14) is the instrument W. The latter instrument has been used by Ginther (2000). The author has argued that this variable lowers the cost of schooling because two-parent families usually have higher incomes than single parent families. However, when I test the validity of the LATE assumptions for this instrument using Mourifié and Wan (2017), I find rejection of the assumptions. Kitagawa (2015) and Mourifié and Wan (2017) have rejected the LATE assumptions for the college proximity instrument. In this sense, both college proximity and presence of both parents at home are invalid instruments. The descriptive statistics are summarized in Table 1 below.

Table 1. Summary Statistics

	Total	
Observations	3,010	
lwage	6.2618	(0.4438)
college	0.2714	(0.4448)
nearc4	0.6821	(0.4658)
momdad14	0.7894	(0.4078)
age	28.1196	(3.1370)
black	0.2336	(0.4232)

Average and standard deviation (in parentheses)

College proximity reduces the marginal cost of schooling (see Example 1) and therefore increases the likelihood of getting a college degree. Likewise, the presence of both parents at home increases

the taste for schooling and thus the likelihood of going to college. Below, I run two regressions (linear probability and logit) of the college degree variable on the instruments college proximity and the presence of both parents at home to check these implications. It turns out that both instruments are positively associated with the probability of obtaining a college degree (see Table 2). The association seems stronger for the presence of both parents at home.

TABLE 2. College degree, college proximity and both parents at home

college	Linear probability	Logit
nearc4	0.0685***	0.3670***
	(0.0172)	(0.0924)
momdad14	0.1700***	1.0298***
	(0.0196)	(0.1239)
n	3010	3010

Standard errors (in parentheses); *** stands for 1% significant.

4.2. **Empirical bounds for the LATE.** For simplicity, I fix $w_0^0 = 0 = w_0^1$, and $w_1^0 = 1 = w_1^1$. I implement the confidence set for the parameters θ^1 and θ^0 using the clrbound command of CKLR. Afterwards, I construct confidence sets for the expectations $\mathbb{E}[Y_1|T=c]$, $\mathbb{E}[Y_0|T=c]$ and then for the LATE using bounds on the densities f_{1c} and f_{0c}, which are generated by the *clrbound* command. I assume that the sample $\{(Y_i, D_i, Z_i, W_i)\}_{i=1}^n$ is i.i.d., and I use the estimators $\hat{c}_0 = 1/\sum_{i=1}^n D_i(1-Z_i)$, $\hat{c}_1 = 1/\sum_{i=1}^n D_i Z_i \mathbb{1}\{W_i = w_1^1\}$, and $\hat{c}_2 = 1/\sum_{i=1}^n D_i Z_i \mathbb{1}\{W_i = w_0^1\}$ in place of $c_0 = 1/\mathbb{E}[D(1-Z)]$, $c_1 = 1/\mathbb{E}[DZ\mathbb{1}\{W = w_1^1\}]$, and $c_2 = 1/\mathbb{E}[DZ\mathbb{1}\{W = w_0^1\}]$ to make the CLR inferential procedure feasible. The validity of this plug-in approach within the CLR method has been shown by Mourifié and Wan (2017).

Implementation. For each candidate θ^1 in $[-M,-1]$ or $[1,M]$, where M is arbitrarily large, check if it belongs to the identified set Θ^1 using the clrbound command to test the moment inequality (3.7) in Corollary 1. I know that θ^1 has the same sign as η^1, which is identified. Hence, the sign of the empirical analog of η^1 helps identify the sign of θ^1. The set Θ^1 is convex. From the clrbound command, I obtain the estimate $\hat{m}_0^1(y;\theta^1)$ of $\mathbb{E}[m_0^1(\theta^1, D, Z, W)|Y=y]$, its standard error $\hat{s}_0^1(y;\theta^1)$ and the critical value $k_{0.95}$. From there, I get the estimate of the density f_{1c}:

$$\hat{f}_{1c}^{0.95}(y;\theta^1) = \left[\hat{m}_0^1(y;\theta^1) + k_{0.95}\hat{s}_0^1(y;\theta^1)\right]\hat{f}(y),$$

where $\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^n K(\frac{y-Y_i}{h})$, $K(u) = \frac{3}{4\sqrt{5}}(1-\frac{1}{5}u^2)\mathbb{1}\{|u| \le \sqrt{5}\}$, $h = n^{-1/5}\left[0.9\min(\sigma_Y, \frac{Q_3-Q_1}{1.349})\right]$, σ_Y, Q_1, Q_3 are empirical standard deviation, first and third quartiles of Y, respectively.

Table 3 shows that the 2SLS estimate is 2.27 log points (i.e., 871.53% average increase in hourly wages) and lies outside the confidence set of the LATE, meaning that it cannot be interpreted as a causal effect of college degree on wages. This is also evidence for rejection of the assumption that

college proximity is independent of the type. As can be seen, the effect of college degree on the log wage is positive for the compliers and ranges from 0.32 to 0.58. This means that the effect of college degree on wages varies between 37.8% and 79.0% for people who obtained the degree only because they lived in a county that had a four-year college. Results for the case where the treatment group

TABLE 3. Confidence sets for parameters

Parameters	Estimates	95% conf. LB	95% conf. UB
θ^1		1	2.3
θ^0		-8.4	-1
$\mathbb{E}[Y_1\|T=c]$		6.4016	6.4248
$\mathbb{E}[Y_0\|T=c]$		5.8425	6.0813
$LATE$		0.3204	0.5824
$2SLS$	2.2737***	1.1463	3.4012
	(0.5750)		
OLS	0.2282***	0.1935	0.2630
	(0.0177)		

Standard errors (in parentheses); *** stands for 1% significant; conf.: confidence; LB: lower bound; UB: upper bound.

is the indicator that the individual has some college education (at least 13 years of education) while the control group is the indicator that she is a high school graduate (12 years of education) are shown in Appendix B.2.2.

Note that the 95% confidence set for the LATE based on the instrument presence of both parents at home is $[0.39, 0.65]$ (i.e., 47.7–91.6% increase) and the 2SLS estimate is 0.89 (with standard error 0.14). In this case, the 2SLS estimate still lies outside the confidence bounds for the LATE, meaning that it does not have a causal interpretation either. The two confidence sets for the two LATEs overlap, suggesting that without further information we cannot reject the hypothesis that the effect of college degree on earnings is homogeneous across individuals.

Adding controls. I control for age and race, and it turns out that the effect does not seem different for black and non-black people (see Table 4). However, the minimum effect of college degree on wages is higher for black compliers (0.21 log points vs 0.10). I compute the LATE for four groups of age and race: (black=1, age $\leq$ 28), (black=1, age $\leq$ 28), (black=0, age $>$ 28) and (black=1, age $>$ 28). The bounds cross for the group (black=1, age $\geq$ 28), suggesting that the identifying assumptions are rejected for this group. Thus, adding covariates may invalidate the identifying assumptions. The bounds for the group (black=1, age $\leq$ 28) are so tight that the upper bound is less than the lower bound of the group (black=0, age $\leq$ 28). This suggests that the return is higher for nonblack compliers in the subpopulation of individuals aged less than 28.

TABLE 4. Confidence sets for LATE

Parameters	95% conf. LB	95% conf. UB
black=0	0.0955	0.3465
black=1	0.2130	0.3957
black=0, age ≤ 28	0.0715	0.2240
black=1, age ≤ 28	0.0586	0.0589
black=0, age > 28	0.1371	0.3089
black=1, age > 28	empty	

conf.: confidence; LB: lower bound; UB: upper bound.

4.3. **Point-identified estimates under tail restrictions.** Under the tail restrictions that the distribution of Y_0 for the never-takers left tail dominates that for the compliers, while the distribution of Y_1 for the always-takers right tail dominates that for the compliers, I point-identify the distributions of Y_0 and Y_1 for the compliers according to Theorem 2. I use estimation results in Jochmans, Henry and Salanié (2017, JHS) to estimate returns to college degree on later earnings for the marginal individuals, i.e., the compliers. See Appendix A.8 for more details. The estimated distributions F_{0c_n} and F_{1c_n} are depicted in Figure 4.

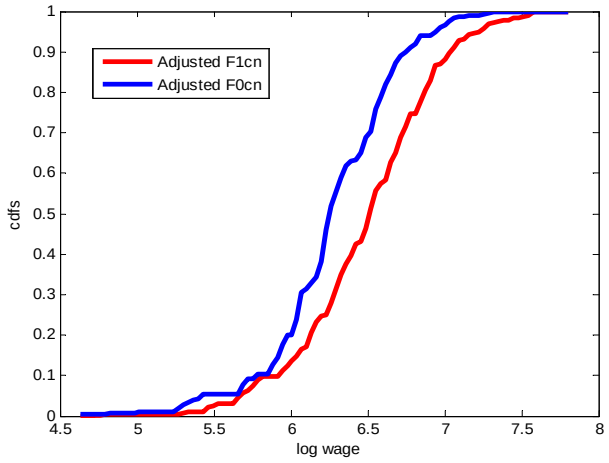

FIGURE 4. Estimates of the distributions F_{1c} and F_{0c}.

From Figure 4, it can be seen that the distribution of Y_1 first order stochastically dominates that of Y_0 for the compliers, implying that the LATE, as well as the LQTE are positive. The estimate of the LATE is 0.22 log points, meaning that college degree increases the average hourly wage for the compliers by 24.6%. However, the fact that the LATE estimate (0.22) under Assumption TR lies outside the LATE confidence bounds ([0.32, 0.58]) suggests that this assumption is rejected.

5. Summary and discussion

In this paper, I develop a new identification strategy when the LATE independence assumption is violated in the sense that the instrument is correlated with the compliance status. I show that when the instrument is independent of the potential outcomes conditional on the type, an additional (invalid) instrument allows to partially identify the potential outcomes distributions for the compliers. Under some testable tail restrictions, these distributions are point-identified. I also show how one can extend my results to settings where the LATE monotonicity assumption does not hold. Moreover, I extend the model to account for the sample selection problem that arises in the return to schooling literature.

Finally, I apply the results on data from the National Longitudinal Survey of Young Men to estimate the returns to college education for the compliers. I find that getting a college degree has a positive effect on the log hourly wage. The evidence suggests that college degree increases the hourly wage of the compliers by 38–79%. The 2SLS estimate is an 872% increase, suggesting that the college proximity instrument is invalid.

I explain in Appendix C.2 how to extend my approach to continuous instruments. However, the marginal treatment effect would be of greater interest in this case than the LATE. This would rely on continuous mixture partial identification result.

Appendix A. Proofs of the main results

A.1. **Proof of Lemma 1.**

Proof. From $(Z,W) \perp\!\!\!\perp U|V$, we have $U \perp\!\!\!\perp (Z,W)|V,W$. Then for any $A \in \mathcal{B}_{\mathcal{Y}}$, $z \in \{0,1\}$, and $w \in \mathcal{W}$,

$$\begin{aligned}
\mathbb{P}(Y_d \in A|T=c, Z=z, W=w) &= \mathbb{P}(g(d,U) \in A|h(0,W,V)=0, h(1,W,V)=1, Z=z, W=w),\\
&= \mathbb{P}(g(d,U) \in A|(V,W) \in h^{-1}(0,0) \cap h^{-1}(1,1), Z=z, W=w),\\
&= \mathbb{P}(g(d,U) \in A|(V,W) \in h^{-1}(0,0) \cap h^{-1}(1,1)),\\
&= \mathbb{P}(Y_d \in A|T=c),
\end{aligned}$$

where $h^{-1}(z,d) = \{(v,w) : h(z,w,v) = d\}$, and the third equality holds because $U \perp\!\!\!\perp (Z,W)|V,W$. The reasoning is similar for the other types a, df, n. □

Notation 3. *Define*

$$r^1(y) \equiv \frac{f(y|1,1,w_1^1)}{f(y|1,1,w_0^1)}, \qquad \underline{r}^1 \equiv \inf_{y\in\mathcal{Y}} r^1(y), \qquad \overline{r}^1 \equiv \sup_{y\in\mathcal{Y}} r^1(y).$$

The quantity $r^1(y)$ is the likelihood ratio evaluated at y of the distribution of Y conditional on $(D=1, Z=1, W=w_1^1)$ and $(D=1, Z=1, W=w_0^1)$, respectively. We have $\int (r^1(y)-1) f(y|1,1,w_0^1)dy = 0$,

which implies that $\underline{r}^1 < 1$ *and* $\overline{r}^1 > 1$.

$$f_*^1 \equiv -\tfrac{1}{\overline{r}^1-1}, \quad f^{*1} \equiv \tfrac{1}{1-\underline{r}^1}, \quad \overline{\Lambda}^1 \equiv \sup_{w\in\mathcal{W}} \Lambda^1(w), \quad \underline{\Lambda}^1 \equiv \inf_{w\in\mathcal{W}} \Lambda^1(w),$$

$$\theta_\ell^1 = \begin{cases} f_*^1+\eta^1 \text{ if } \overline{\Lambda}^1 \le -\eta^1 \le f^{*1} \\ \overline{\Lambda}^1+\eta^1 \text{ if } f_*^1 \le -\eta^1 \le \underline{\Lambda}^1 \\ +\infty \text{ otherwise} \end{cases} \quad \text{and} \quad \theta_u^1 = \begin{cases} \underline{\Lambda}^1+\eta^1 \text{ if } \overline{\Lambda}^1 \le -\eta^1 \le f^{*1} \\ f^{*1}+\eta^1 \text{ if } f_*^1 \le -\eta^1 \le \underline{\Lambda}^1 \\ -\infty \text{ otherwise} \end{cases}$$

□

A.2. **Proof of Theorem 1.**

Proof. I have to show that the constraints on the distributions F_{1a}, F_{0n}, F_{1c} and F_{0c} are valid and sharp under CI, MON and REL.

Validity: As shown in the main text, $F_{1a}(y) = F(y|1,0)$ and $F_{0n}(y) = F(y|0,1)$ under CI and MON. Consider the mixture model (3.1)

$$F(y|1,1,w) = \alpha^1(w)F_{1c}(y) + \big(1-\alpha^1(w)\big)F_{1a}(y),$$

where $0 \le \alpha^1(w) \le 1$.

Under RAN, Equation (3.4) holds with the following constraints on (θ^1, η^1) according to Theorem 1 of HKS:

$$f_*^1 \le \min\big(\theta^1-\eta^1, -\eta^1\big) \le \underline{\Lambda}^1, \text{ and } \overline{\Lambda}^1 \le \max\big(\theta^1-\eta^1, -\eta^1\big) \le f^{*1}.$$

Now, adding the constraint $F_{1a}(y) = F(y|1,0)$, the parameter η^1 is identified as follows:

$$\eta^1 = \frac{F(y_1^1|1,1,w_0^1) - F(y_1^1|1,0)}{F(y_1^1|1,1,w_1^1) - F(y_1^1|1,1,w_0^1)},$$

and we have the following constraints on the remaining parameter θ^1

$$f_*^1+\eta^1 \le \min\big(\theta^1, 0\big) \le \underline{\Lambda}^1+\eta^1, \text{ and } \overline{\Lambda}^1+\eta^1 \le \max\big(\theta^1, 0\big) \le f^{*1}+\eta^1. \tag{A.1}$$

Because $\underline{\Lambda}^1 \le 0$ and $\overline{\Lambda}^1 \ge 1$, the following holds: If $\overline{\Lambda}^1 \le -\eta^1 \le f^{*1}$, then $\theta^1 < 0$ and $f_*^1+\eta^1 \le \theta^1 \le \underline{\Lambda}^1+\eta^1$; If $f_*^1 \le -\eta^1 \le \underline{\Lambda}^1$, then $\theta^1 > 0$ and $\overline{\Lambda}^1+\eta^1 \le \theta^1 \le f^{*1}+\eta^1$. Indeed, if $\overline{\Lambda}^1 \le -\eta^1 \le f^{*1}$, then $\overline{\Lambda}^1+\eta^1 \le 0$, which implies that $\underline{\Lambda}^1+\eta^1 < 0$, and therefore $\theta^1 < 0$. Similarly, if $f_*^1 \le -\eta^1 \le \underline{\Lambda}^1$, then $\underline{\Lambda}^1+\eta^1 \ge 0$, which implies that $\overline{\Lambda}^1+\eta^1 > 0$, and thus $\theta^1 > 0$.

Hence, $\theta_\ell^1 \le \theta^1 \le \theta_u^1$, where θ_ℓ^1 and θ_u^1 are defined in Notation 3 above. Analogously for θ^0, we have $\theta_\ell^0 \le \theta^0 \le \theta_u^0$, where θ_ℓ^0 and θ_u^0 are defined in a similar way as θ_ℓ^1 and θ_u^1, respectively.

If RAN does not hold, then we have trivial bounds on θ^1, i.e., $\theta^1 \in (-\infty,-1] \cup [1,+\infty)$, in which case F_{1c} is point-identified: $F_{1c}(y) = F(y|1,0)$. Similar results hold for θ^0 and F_{0c}.

Note that the constraints on the distribution F_{dc} $(d = 0, 1)$ do not depend on the choices of w_0^d and w_1^d. Under RAN, refer to the proof of Theorem 1 in HKS for details. If RAN does not hold, this is straightforward.

Sharpness: It remains to show that for any pair $(\theta^0, \theta^1) \in \Theta^0 \times \Theta^1$, there exists a joint distribution of (Y_0, Y_1, T, Z, W) that generates the joint distribution of the data (Y, D, Z, W) through the potential outcome model (2.2) and satisfies assumptions CI, MON and REL. Assumption REL is satisfied as long as Θ^d is well-defined.

Define $\alpha^1(w)$, $\alpha^0(w)$, $F_{1c}(y)$, $F_{0c}(y)$, $F_{1a}(y)$ and $F_{0n}(y)$ as above. It is clear that $\alpha^1(w)$ and $\alpha^0(w)$ lie within [0,1], and $F_{1c}(y)$, $F_{0c}(y)$, $F_{1a}(y)$ and $F_{0n}(y)$ are cdfs. Denote $F_d(y, t|z, w) \equiv \mathbb{P}(Y_d \leq y, T = t|Z = z, W = w)$, $d = 0, 1$, $p(t|z, w) \equiv \mathbb{P}(T = t|Z = z, W = w)$. Define the conditional probabilities of the types

$$\begin{aligned}
p(c|1, w) &\equiv \alpha^1(w)\mathbb{P}(D = 1|Z = 1, W = w), \\
p(c|0, w) &\equiv \alpha^0(w)\mathbb{P}(D = 0|Z = 0, W = w), \\
p(a|0, w) &\equiv \mathbb{P}(D = 1|Z = 0, W = w), \\
p(a|1, w) &\equiv (1 - \alpha^1(w))\mathbb{P}(D = 1|Z = 1, W = w), \\
p(n|1, w) &\equiv \mathbb{P}(D = 0|Z = 1, W = w), \\
p(n|0, w) &\equiv (1 - \alpha^0(w))\mathbb{P}(D = 0|Z = 0, W = w),
\end{aligned}$$

and the joint distributions of (Y_0, Y_1, T) conditional on $(Z = z, W = w)$

$$\begin{aligned}
\mathbb{P}(Y_0 \leq y_0, Y_1 \leq y_1, T = c|Z = z, W = w) &\equiv F_{0c}(y_0)F_{1c}(y_1)p(c|z, w), \\
\mathbb{P}(Y_0 \leq y_0, Y_1 \leq y_1, T = a|Z = z, W = w) &\equiv F_{0a}(y_0)F_{1a}(y_1)p(a|z, w), \\
\mathbb{P}(Y_0 \leq y_0, Y_1 \leq y_1, T = n|Z = z, W = w) &\equiv F_{0n}(y_0)F_{1n}(y_1)p(n|z, w),
\end{aligned}$$

where F_{0a} and F_{1n} are abitrary cdfs defined on the measurable space $(\mathcal{Y}, \mathcal{B}_\mathcal{Y})$. It is straightforward that the above joint distribution satisfies CI and MON. Assumption MON is satified by construction. For CI, we have for instance

$$\begin{aligned}
\mathbb{P}(Y_1 \leq y_1|T = c, Z = 1, W = w) &= \frac{\mathbb{P}(Y_1 \leq y_1, T = c|Z = z, W = w)}{\mathbb{P}(T = c|Z = z, W = w)}, \\
&= \frac{\lim_{y_0 \uparrow \infty} \mathbb{P}(Y_0 \leq y_0, Y_1 \leq y_1, T = c|Z = z, W = w)}{\lim_{y_0 \uparrow \infty} \lim_{y_1 \uparrow \infty} \mathbb{P}(Y_0 \leq y_0, Y_1 \leq y_1, T = c|Z = z, W = w)}, \\
&= \frac{F_{1c}(y_1)p(c|1, w)}{p(c|1, w)}, \\
&= F_{1c}(y_1).
\end{aligned}$$

Finally, this joint distribution of (Y_0, Y_1, T) conditional on $(Z = z, W = w)$ induces the joint distribution of (Y, D) conditional on $(Z = z, W = w)$.

$$\begin{aligned}
\mathbb{P}(Y \le y, D = 1|Z = 1, W = w) &= \mathbb{P}(Y_1 \le y, D_1 = 1|Z = 1, W = w), \\
&= \mathbb{P}(Y_1 \le y, T = c|Z = 1, W = w) + \mathbb{P}(Y_1 \le y, T = a|Z = 1, W = w), \\
&= F_{1c}(y)p(c|1, w) + F_{1a}(y)p(a|1, w), \\
&= \left[\alpha^1(w)F_{1c}(y) + (1 - \alpha^1(w))F_{1a}(y)\right]\mathbb{P}(D = 1|Z = 1, W = w).
\end{aligned}$$

This reasoning is similar for $\mathbb{P}(Y \le y, D = 1|Z = 0, W = w)$, $\mathbb{P}(Y \le y, D = 0|Z = 1, W = w)$, and $\mathbb{P}(Y \le y, D = 0|Z = 0, W = w)$. This completes the proof. □

A.3. **Proof of Corollary 1.**

Proof. Recall that the bounds on θ^1 have been derived using only the constraints that the cdf $F_{1c}(y)$ is nondecreasing as in (3.5), and the weight function $\alpha^1(w)$ lies between 0 and 1 as in (3.6). Under RAN, the parameter η^1 and the function Λ^1 are point-identified:

$$\eta^1 = \frac{F(y_1^1|1,1,w_0^1) - F(y_1^1|1,0)}{F(y_1^1|1,1,w_1^1) - F(y_1^1|1,1,w_0^1)}, \quad \text{and} \quad \Lambda^1(w) = \frac{F(y_1^1|1,1,w) - F(y_1^1|1,1,w_0^1)}{F(y_1^1|1,1,w_1^1) - F(y_1^1|1,1,w_0^1)}.$$

I can then equivalently rewrite condition (3.6) as follows:

$$\begin{cases}
sign(\theta^1)\left(\theta^1 - \frac{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_0^1]}\right) \ge 0 \\
sign(\theta^1)\frac{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_0^1]} \ge 0
\end{cases}$$

where $sign(\theta^1) = 1\{\theta^1 > 0\} - 1\{\theta^1 < 0\}$. Therefore, the identified set for θ^1 is fully characterized by the following inequality:

$$\inf_{(y,w)\in\mathcal{Y}\times\mathcal{W}} \beta^1(y, w) \ge 0, \tag{A.2}$$

where

$$\beta^1(y, w) = \begin{bmatrix}
f(y|1, 0) + \theta^1\left[f(y|1, 1, w_1^1) - f(y|1, 1, w_0^1)\right] \\
sign(\theta^1)\left(\theta^1 - \frac{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_0^1]}\right) \\
sign(\theta^1)\frac{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=0]}{\mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_1^1] - \mathbb{E}[\mathbb{1}\{Y \le y_1^1\}|D=1,Z=1,W=w_0^1]}
\end{bmatrix}.$$

Let $f(y)$ denote the density (probability mass) function of Y. Using Bayes' rule, we have:

$$f(y|d, z, w) = \frac{\mathbb{P}(D = d, Z = z, W = w|Y = y)f(y)}{\mathbb{P}(D = d, Z = z, W = w)},$$

for all $d, z \in \{0, 1\}$ and $w \in \mathcal{W}$. Therefore, for all $y \in \mathcal{Y}$ such that $f(y) > 0$, condition (3.5) is equivalent to:

$$\frac{\mathbb{P}(D = 1, Z = 0|Y = y)f(y)}{\mathbb{P}(D = 1, Z = 0)} + \theta^1\left[\frac{\mathbb{P}(D = 1, Z = 1, W = w_1^1|Y = y)f(y)}{\mathbb{P}(D = 1, Z = 1, W = w_1^1)} - \frac{\mathbb{P}(D = 1, Z = 1, W = w_0^1|Y = y)f(y)}{\mathbb{P}(D = 1, Z = 1, W = w_0^1)}\right] \ge 0,$$

which in turn is equivalent to:

$$c_0\mathbb{E}[D(1-Z)|Y=y]+\theta^1\left[c_1\mathbb{E}[DZ\mathbb{1}\left\{W=w_1^1\right\}|Y=y]-c_2\mathbb{E}[DZ\mathbb{1}\left\{W=w_0^1\right\}|Y=y]\right]\geq 0,$$

where $c_0 = 1/\mathbb{E}[D(1-Z)]$, $c_1 = 1/\mathbb{E}[DZ\mathbb{1}\left\{W=w_1^1\right\}]$, and $c_2 = 1/\mathbb{E}[DZ\mathbb{1}\left\{W=w_0^1\right\}]$. This last inequality can be rewritten as:

$$\mathbb{E}\left[c_0D(1-Z)+\theta^1\left(c_1DZ\mathbb{1}\left\{W=w_1^1\right\}-c_2DZ\mathbb{1}\left\{W=w_0^1\right\}\right)|Y=y\right]\geq 0.$$

□

A.4. **Proof of Proposition 1.**

Proof. We have $\mathbb{E}[Y_1-Y_0|T=c]=\mathbb{E}[Y_1|T=c]-\mathbb{E}[Y_0|T=c]=\mathbb{E}[F_{1c}]-\mathbb{E}[F_{0c}]$. From Theorem 1, there exists $(\theta^1,\theta^0)\in\Theta^1\times\Theta^0$ such that $F_{1c}=F_{1c}^{\theta^1}$ and $F_{0c}=F_{0c}^{\theta^0}$. Therefore, the bounds hold. It remains to show their sharpness. We have

$$\begin{aligned}\mathbb{E}[F_{1c}] &= \mathbb{E}[Y|1,0]+\theta^1\left(\mathbb{E}[Y|1,1,w_1^1]-\mathbb{E}[Y|1,1,w_0^1]\right),\\ \mathbb{E}[F_{0c}] &= \mathbb{E}[Y|0,1]+\theta^0\left(\mathbb{E}[Y|0,0,w_1^0]-\mathbb{E}[Y|0,0,w_0^0]\right),\end{aligned}$$

and

$$\inf_{(\theta^1,\theta^0)\in\Theta^1\times\Theta^0}\left\{\mathbb{E}\left[F_{1c}^{\theta^1}\right]-\mathbb{E}\left[F_{0c}^{\theta^0}\right]\right\}=\inf_{\theta^1\in\Theta^1}\mathbb{E}[F_{1c}]-\sup_{\theta^0\in\Theta^0}\mathbb{E}[F_{0c}].$$

If RAN does not hold, the bounds on F_{dc} $(d=0,1)$ do not depend on θ^d as F_{dc} is point-identified. In this case, sharpness is straightforward.

Under RAN, since the expectation $\mathbb{E}[F_{dc}]$ is continuous in θ^d and Θ^d is compact for all $d\in\{0,1\}$, the infimum and the supremum are attainable, say at θ^{1*} and θ^{0*}, respectively. Define $F_{1c}^*(y)=F_{1c}^{\theta^{1*}}(y)$, $F_{0c}^*(y)=F_{0c}^{\theta^{0*}}(y)$. Set $F_{0a}^*=F_{1a}^*=F(y|1,0)$, $F_{1n}^*=F_{0n}^*=F(y|0,1)$, and $F^*_{(Y_1,Y_0)|T=t,Z=z,W=w}(y_1,y_2)=F_{1t}^*(y_1)\times F_{0t}^*(y_2)$ for all $t\in\{a,c,n\}$. Then $F^*_{(Y_1,Y_0)|T,Z,W}$ achieves the lower bound. Similar reasoning works for the upper bound. Therefore, the bounds on the LATE are sharp.

In the same way, since $F_{dc}(y)$ is continuous in θ^d and Θ^d is compact, so is the quantile function F_{dc}^{-1} for all $d\in\{0,1\}$ if the outcome Y is continuous and has compact support (see Lemma 2 below). By similar reasoning, the bounds on the LQTE are attained. This completes the proof. □

This result is probably attained in the literature, but since I could not find a reference, I give a proof for completeness.

Lemma 2. *Let $F(y;\theta)$ be a cumulative distribution function of a real-valued random variable Y_θ with compact support $\mathcal{Y}$. Assume that $F(y;\theta)$ is continuous in y for all θ and continuous in θ for all y. Then, the generalized inverse $F^{-1}(p;\theta)$ defined for every $p\in(0,1)$ by*

$$F^{-1}(p;\theta)=\inf\{y\in\mathcal{Y}:F(y;\theta)\geq p\}$$

is also continuous in θ for all p. □

Proof of Lemma 2.

Proof. Notice that $F^{-1}(p;\theta)$ is the unique solution of the following optimization problem:

$$\min f(\theta, y) \equiv y \ s.t. \ y \in \Gamma(\theta) = \{y \in \mathcal{Y} : F(y;\theta) \geq p\}$$

I use the Theorem of the Maximum (Theorem 3.6 of Stokey and Lucas p.62). Define $h(\theta) = \min_{y\in\Gamma(\theta)} y = -\max_{y\in\Gamma(\theta)} -y$ and $G(\theta) = \{y \in \Gamma(\theta) : f(\theta, y) = h(\theta)\}$ The function f is continuous. I am going to show that the correspondence Γ is compact-valued and continuous.

Compactness: $\Gamma(\theta) \subset \mathcal{Y}$ compact. Then, $\Gamma(\theta)$ is bounded. Now, take $y_n \in \Gamma(\theta)$ s.t. $y_n \to y$. Let us show that $y \in \Gamma(\theta)$.

$$\begin{aligned}
y_n \in \Gamma(\theta) \quad &\Rightarrow \quad F(y_n;\theta) \geq p \\
&\Rightarrow \quad \lim_{n\to\infty} F(y_n;\theta) \geq p \\
&\Rightarrow \quad F(\lim_{n\to\infty} y_n;\theta) \geq p \text{ by continuity of } F(y;\theta) \text{ in } y \\
&\Rightarrow \quad F(y;\theta) \geq p \\
&\Rightarrow \quad y \in \Gamma(\theta)
\end{aligned}$$

Then, $\Gamma(\theta)$ is closed. Thus, $\Gamma(\theta)$ is compact.

Continuity: I show that $\Gamma(\theta)$ is lower hemicontinuous (l.h.c.) and upper hemicontinuous (u.h.c.). I use the definitions in Stokey and Lucas p.56. $\Gamma(\theta)$ is nonempty for all θ.

l.h.c.: Take $y \in \Gamma(\theta)$. Then, $F(y;\theta) \geq p$. Let θ_n be a sequence s.t. $\theta_n \to \theta$. By continuity of $F(y;\theta)$ in θ, $F(y;\theta_n) \to F(y;\theta)$. That is, $\forall\ \epsilon > 0,\ \exists\ n_\epsilon : \forall\ n > n_\epsilon,\ |F(y;\theta_n) - F(y;\theta)| < \epsilon$, which implies that $F(y;\theta_n) > F(y;\theta) - \epsilon \geq p - \epsilon$. Hence, for $\epsilon \to 0,\ \exists\ n_0 :\ \forall\ n > n_0,\ F(y;\theta_n) \geq p$. Then, considering the sequence $\{y_n = y\}_{n=n_0}^{\infty}$, we have $y_n \to y$ and $y_n \in \Gamma(\theta_n)$. This shows that $\Gamma(\theta)$ is l.h.c..

u.h.c.: Take $\theta_n \to \theta$ and $y_n \in \Gamma(\theta_n)$. I am going to show that there exists a subsequence $\{y_k\} \to y \in \Gamma(\theta)$. We have $F(y_n;\theta_n) \geq p\ \forall\ n$. Moreover, $y_n \in \mathcal{Y}$, which is compact. Then, by Bolzano-Weierstrass theorem, there exists a subsequence $\{y_k\}$ s.t. $y_k \to y$. Now, it remains to show that $y \in \Gamma(\theta)$. Indeed, we have the following implications:

$$\begin{aligned}
F(y_k;\theta_n) \geq p\ \forall\ n,\ \forall\ k \quad &\Rightarrow \quad \lim_{n\to\infty} F(y_k;\theta_n) \geq p \\
&\Rightarrow \quad F(y_k; \lim_{n\to\infty} \theta_n) = F(y_k;\theta) \geq p \text{ by continuity in } z \\
&\Rightarrow \quad \lim_{k\to\infty} F(y_k;\theta) \geq p \\
&\Rightarrow \quad F(\lim_{k\to\infty} y_k;\theta) = F(y;\theta) \geq p \text{ by continuity in } y \\
&\Rightarrow \quad y \in \Gamma(\theta).
\end{aligned}$$

Then, $\Gamma(\theta)$ is u.h.c..

Therefore, by the Theorem of Maximum, the function $h(\theta)$ is continuous and the correspondence $G(\theta)$ is nonempty, compact-valued and u.h.c.. Since $F^{-1}(p;\theta)$ is the unique solution of the problem, $F^{-1}(p;\theta)$ is continuous in θ. □

A.5. **Proof of Proposition 2.**

Proof. The necessity of the bounds is shown as follows. We have

$$\begin{aligned} F_{Y_0}(y|D=1,W=w) &= \mathbb{P}(Z=0|D=1,W=w)F_{Y_0}(y|D=1,Z=0,W=w) \\ &+ \mathbb{P}(Z=1|D=1,W=w)F_{Y_0}(y|D=1,Z=1,W=w), \end{aligned}$$

and

$$\begin{aligned} F_{Y_0}(y|D=1,Z=0,W=w) &= \mathbb{P}(T=a|D=1,Z=0,W=w)F_{Y_0}(y|T=a,Z=0,W=w) \\ &+ \mathbb{P}(T=df|D=1,Z=0,W=w)F_{Y_0}(y|T=df,Z=0,W=w) \\ &= F_{Y_0}(y|T=a,Z=0,W=w) \\ &= F_{Y_0}(y|T=a), \end{aligned}$$

where the first equality follows from the law of iterated expectations (LIE), the second holds under MON, and the last holds under CI.

We also have

$$\begin{aligned} F_{Y_0}(y|D=1,Z=1,W=w) &= \mathbb{P}(T=a|D=1,Z=1,W=w)F_{Y_0}(y|T=a,Z=1,W=w) \\ &+ \mathbb{P}(T=c|D=1,Z=1,W=w)F_{Y_0}(y|T=c,Z=1,W=w) \\ &= \mathbb{P}(T=a|D=1,Z=1,W=w)F_{Y_0}(y|T=a) \\ &+ \mathbb{P}(T=c|D=1,Z=1,W=w)F_{Y_0}(y|T=c), \end{aligned}$$

where the first equality follows from the LIE, and the second holds under CI. This second equality can be rewritten as follows:

$$F_{Y_0}(y|D=1,Z=1,W=w) = (1-\alpha^1(w))F_{Y_0}(y|T=a) + \alpha^1(w)F_{Y_0}(y|T=c).$$

Therefore,

$$\begin{aligned} F_{Y_0}(y|D=1,W=w) &= \left[\mathbb{P}(Z=0|D=1,W=w) + \mathbb{P}(Z=1|D=1,W=w)(1-\alpha^1(w))\right]F_{Y_0}(y|T=a) \\ &+ \mathbb{P}(Z=1|D=1,W=w)\alpha^1(w)F_{Y_0}(y|T=c). \end{aligned}$$

The worst case bounds on $F_{Y_0}(y|T=a)$ ([0,1]), the sharp bounds on $\alpha^1(w)$ and the sharp bounds on $F_{Y_0}(y|T=c)$ imply nontrivial bounds on $F(Y_0|D=1,W=w)$.

Similarly,

$$\begin{aligned} F_{Y_1}(y|D=0,W=w) &= \left[\mathbb{P}(Z=1|D=0,W=w)+\mathbb{P}(Z=0|D=0,W=w)\alpha^0(w)\right]F_{Y_1}(y|T=n) \\ &+ \mathbb{P}(Z=0|D=0,W=w)(1-\alpha^0(w))F_{Y_1}(y|T=c). \end{aligned}$$

Take $(\theta^1,\theta^0)\in\Theta^1\times\Theta^0$, $\psi^1\in\Psi$, and $\psi^0\in\Psi$. Define

$$\begin{aligned} &F_{0a}^{\theta^0}=\psi^0\in\Psi, \text{ and } F_{1n}^{\theta^1}=\psi^1\in\Psi, \\ &\tilde{F}_{(Y_1,Y_0)|T=t,Z=z,W=w}=F_{1t}^{\theta^1}\times F_{0t}^{\theta^0} \text{ for all } t\in\{a,c,n\},\ z\in\{0,1\},\ w\in\mathcal{W}. \end{aligned}$$

The conditional joint distribution $\tilde{F}_{(Y_1,Y_0)|T,Z,W}$ achieves each element in the identified sets of $F_{Y_0}(y|D=1,W=w)$ and $F_{Y_1}(y|D=0,W=w)$. □

A.6. **Proof of Corollary 2.**

Proof. We have

$$\begin{aligned} ATT(w) &= \mathbb{E}[Y_1-Y_0|D=1,W=w], \\ &= \mathbb{E}[Y_1|D=1,W=w]-\mathbb{E}[Y_0|D=1,W=w], \\ &= \mathbb{E}[Y|D=1,W=w]-\mathbb{E}\left[F(Y_0|D=1,W=w)\right], \\ &= \mathbb{E}[Y|D=1,W=w]-\left[\mathbb{P}(Z=0|D=1,W=w)+\mathbb{P}(Z=1|D=1,W=w)(1-\alpha^1(w))\right]\delta \\ &- \mathbb{P}(Z=1|D=1,W=w)\alpha^1(w)\mu_{0c}^{\theta^0}. \end{aligned}$$

Similarly,

$$\begin{aligned} ATUT(w) &= \mathbb{E}[Y_1-Y_0|D=0,W=w], \\ &= \mathbb{E}[Y_1|D=0,W=w]-\mathbb{E}[Y_0|D=0,W=w], \\ &= \mathbb{E}\left[F(Y_1|D=0,W=w)\right]-\mathbb{E}[Y|D=0,W=w], \\ &= \left[\mathbb{P}(Z=1|D=0,W=w)+\mathbb{P}(Z=0|D=0,W=w)\alpha^0(w)\right]\delta \\ &+ \mathbb{P}(Z=0|D=0,W=w)(1-\alpha^0(w))\mu_{1c}^{\theta^1}-\mathbb{E}[Y|D=0,W=w]. \end{aligned}$$

The rest of the proof is straightforward as Θ^0 and Θ^1 are sharp. □

A.7. **Proof of Theorem 2.**

Proof. From Theorem 1, we have

$$\begin{aligned} F_{1a}(y) &= F(y|1,0), \\ F_{1c}(y) &= F(y|1,1,w_0^1)+\left(\theta^1-\eta^1\right)\left[F(y|1,1,w_1^1)-F(y|1,1,w_0^1)\right]. \end{aligned}$$

Then

$$\lim_{y\uparrow y^u}\frac{1-F_{1c}(y)}{1-F_{1a}(y)} = \lim_{y\uparrow y^u}\frac{1-F(y|1,1,w_0^1)}{1-F(y|1,0)}+\left(\theta^1-\eta^1\right)\left[\lim_{y\uparrow y^u}\frac{1-F(y|1,1,w_1^1)}{1-F(y|1,0)}-\lim_{y\uparrow y^u}\frac{1-F(y|1,1,w_0^1)}{1-F(y|1,0)}\right].$$

Therefore,

$$\lim_{y\uparrow y^u}\frac{1-F_{1c}(y)}{1-F_{1a}(y)}=0\Rightarrow\theta^1-\eta^1=\frac{1}{1-\pi^1(w_1^1,w_0^1)},$$

where

$$\pi^1(w_1^1,w_0^1) \quad = \quad \lim_{y\uparrow y^u}\frac{1-F(y|1,1,w_1^1)}{1-F(y|1,1,w_0^1)}.$$

The reasoning is similar for $d=0$. This completes the proof. □

Now, I show that Theorem 2 still holds under only CI, REL and TR without MON.

Proof. Under CI, Equation (3.1) holds and we have

$$1-F(y|1,1,w)=\alpha^1(w)\left[1-F_{1c}(y)\right]+\left(1-\alpha^1(w)\right)\left[1-F_{1a}(y)\right].$$

Under REL, at least one of the weights $\alpha^1(w_1^1)$ and $\alpha^1(w_0^1)$ is different from 1. Assume without loss of generality that $\alpha^1(w_0^1)\neq 1$. Then

$$\lim_{y\uparrow y^u}\frac{1-F(y|1,1,w_1^1)}{1-F(y|1,1,w_0^1)}=\lim_{y\uparrow y^u}\frac{\alpha^1(w_1^1)\frac{1-F_{1c}(y)}{1-F_{1a}(y)}+1-\alpha^1(w_1^1)}{\alpha^1(w_0^1)\frac{1-F_{1c}(y)}{1-F_{1a}(y)}+1-\alpha^1(w_0^1)}=\frac{1-\alpha^1(w_1^1)}{1-\alpha^1(w_0^1)}\equiv\pi^1(w_1^1,w_0^1),$$

where the second equality holds under TR.

Under REL, we have $\frac{1}{1-\pi^1(w_1^1,w_0^1)}=\frac{1-\alpha^1(w_0^1)}{\alpha^1(w_1^1)-\alpha^1(w_0^1)}$. Then

$$\begin{aligned}\tfrac{1}{1-\pi^1(w_1^1,w_0^1)}\left[F(y|1,1,w_1^1)-F(y|1,0,w_0^1)\right] &= \tfrac{\alpha^1(w_1^1)-\alpha^1(w_0^1)}{1-\pi^1(w_1^1,w_0^1)}\left[F_{1c}(y)-F_{1a}(y)\right],\\ &= \left(1-\alpha^1(w_0^1)\right)\left[F_{1c}(y)-F_{1a}(y)\right],\\ &= F_{1c}(y)-F(y|1,1,w_0^1),\end{aligned}$$

where the first equality follows from Equation (3.3), the second from the above equality, and the last holds from (3.1). Thus,

$$F_{1c}(y)=F(y|1,1,w_0^1)+\frac{1}{1-\pi^1(w_1^1,w_0^1)}\left[F(y|1,1,w_1^1)-F(y|1,0,w_0^1)\right].$$

The reasoning is similar for F_{0c}. This completes the proof. □

A.8. **Estimation under TR (JHS).** The following derivation is directly from JHS and is given for completeness. Recall that

$$\begin{aligned}F_{0c}(y) &= F(y|0,0,w_0^0)+\frac{1}{1-\zeta^0(w_1^0,w_0^1)}\left[F(y|0,0,w_1^0)-F(y|0,0,w_0^0)\right],\\ F_{1c}(y) &= F(y|1,1,w_0^1)+\frac{1}{1-\pi^1(w_1^1,w_0^1)}\left[F(y|1,1,w_1^1)-F(y|1,1,w_0^1)\right],\end{aligned}$$

where

$$\zeta^0(w_1^0,w_0^0) \quad = \quad \lim_{y\downarrow y^\ell}\tfrac{F(y|0,0,w_1^0)}{F(y|0,0,w_0^0)}, \quad \text{and} \quad \pi^1(w_1^1,w_0^1) \quad = \quad \lim_{y\uparrow y^u}\tfrac{1-F(y|1,1,w_1^1)}{1-F(y|1,1,w_0^1)}.$$

Define

$$F_n(y|d,z,w) \equiv n_{dzw}^{-1} \sum_{i=1}^{n} \mathbb{1}\{Y_i \leq y, D = d, Z = z, W = w\},$$

where $n_{dzw} \equiv \sum_{i=1}^{n} \mathbb{1}\{D = d, Z = z, W = w\}$ for $d, z, w \in \{0,1\}$.

From JHS, a consistent estimator[1] for $\zeta^0(1,0)$ is

$$\zeta_n^0(1,0) = \frac{F_n(l_{n_{000}}|0,0,w_1^0)}{F_n(l_{n_{000}}|0,0,w_0^0)},$$

where $l_{n_{000}}$ denotes the $(\iota_{n_{000}}+1)th$ order statistics of Y conditional on $(D=0, Z=0, W=0)$, $\iota_{n_{000}}$ being chosen such that $l_{n_{000}} \downarrow -\infty$ as $n \uparrow \infty$. Similarly, a consistent estimator for $\pi^1(1,0)$ is

$$\pi_n^1(1,0) = \frac{1 - F_n(r_{n_{110}}|1,1,w_1^1)}{1 - F_n(r_{n_{110}}|1,1,w_0^1)},$$

where $r_{n_{110}}$ denotes the $(n_{110} - \kappa_{n_{110}})th$ order statistics of Y conditional on $(D=1, Z=1, W=0)$, $\kappa_{n_{110}}$ being chosen such that $r_{n_{110}} \uparrow \infty$ as $n \uparrow \infty$.

Therefore, we have the following estimators for F_{0c} and F_{1c}:

$$\begin{aligned} F_{0c_n}(y) &= F_n(y|0,0,w_0^0) + \frac{1}{1-\zeta_n^0(1,0)}\left[F_n(y|0,0,w_1^0) - F_n(y|0,0,w_0^0)\right], \\ F_{1c_n}(y) &= F_n(y|1,1,w_0^1) + \frac{1}{1-\pi_n^1(1,0)}\left[F_n(y|1,1,w_1^1) - F_n(y|1,1,w_0^1)\right]. \end{aligned}$$

Asymptotic normality results for $\zeta_n^0(1,0)$, $\pi_n^1(1,0)$, $F_{dc_n}(y)$, $d \in \{0,1\}$, are given in JHS. As in the simulation experiments of JHS, I use $\iota_{n_{000}} = C_0(n_{000}\ln\ln n_{000})^{0.6}$ and $\kappa_{n_{110}} = C_1(n_{110}\ln\ln n_{110})^{0.6}$ for reasonable choices of the constants C_0 and C_1. In the application, I choose $C_0 = 1$ and $C_1 = 0.5$.

A.9. **Numerical illustration.** In what follows, I use the following notation: $F_d(y|t,z,w) \equiv \mathbb{P}(Y_d \leq y|T=t, Z=z, W=w)$, $p(t|z,w) \equiv \mathbb{P}(T=t|Z=z, W=w)$, $t \in \{a,c,n\}$, $p(z,w) \equiv \mathbb{P}(Z=z, W=w)$, $z,w \in \{0,1\}$. I consider DGP: $\mathcal{Y} = [0,\infty)$,

$$\begin{array}{lll} p(a|1,1) = 0.2, & p(c|1,1) = 0.5, & p(n|1,1) = 0.3, \\ p(a|1,0) = 0.1, & p(c|1,0) = 0.4, & p(n|1,0) = 0.5, \\ p(a|0,1) = 0.4, & p(c|0,1) = 0.2, & p(n|0,1) = 0.4, \\ p(a|0,0) = 0.35, & p(c|0,0) = 0.45, & p(n|0,0) = 0.20. \end{array}$$

$$p(1,1) = 0.3,\ p(1,0) = 0.2,\ p(0,1) = 0.2, \text{ and } p(0,0) = 0.3.$$

$$\begin{array}{c} F_1(y|a,z,w) = \frac{y}{y+1},\ F_1(y|c,z,w) = \frac{y^2}{y^2+1},\ F_1(y|n,z,w) = \frac{y^3}{y^3+1}, \\ F_0(y|a,z,w) = \frac{y^2}{y^2+1},\ F_0(y|c,z,w) = \frac{y^3}{y^3+1},\ F_0(y|n,z,w) = \frac{y^{3/2}}{y^{3/2}+1}. \end{array}$$

[1] under some assumptions that I omit here. See JHS for more details.

It can be shown that the type is confounded. For instance, $\mathbb{P}(T = c|Z = 1) = 0.46 \neq 0.35 = \mathbb{P}(T = c|Z = 0)$. It is easy to see that CI holds. After computing the identfied set numerically, I find $\Theta^1 = [-11.79, -9.33]$, and $\Theta^0 = [-2.79, -1.93]$. We have an uncountable number of distributions F_{1c} and F_{0c}. I discretize the identified sets Θ^1 and Θ^0. Figure 5 displays distributions of the potential outcomes Y_0 and Y_1 for four values of the parameters θ^0 and θ^1, including their bounds.

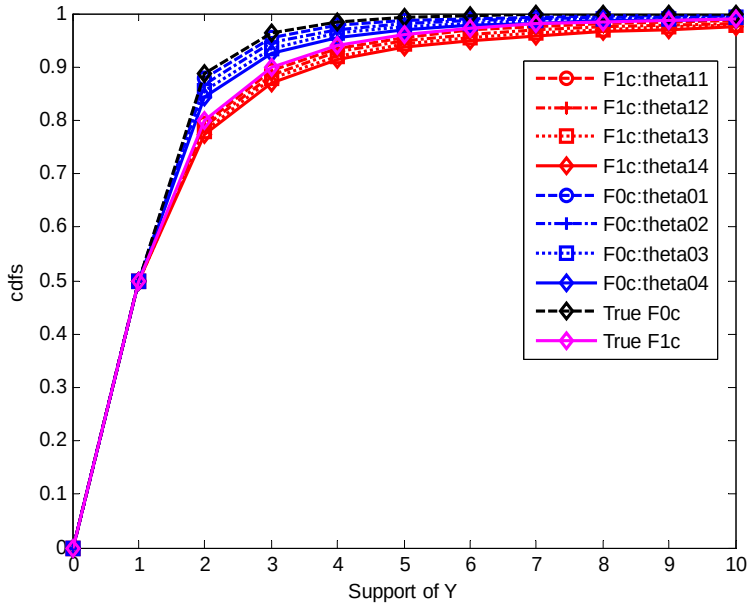

FIGURE 5. Bounds on the distributions F_{1c} and F_{0c}.

A.10. **LATE and IV estimand.** By definition, the IV estimand is

$$\alpha_{IV} = \frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\mathbb{E}[D|Z = 1] - \mathbb{E}[D|Z = 0]}.$$

When the type is confounded, the IV estimand becomes:

$$\alpha_{IV} = \frac{p(c|1)\mathbb{E}[Y_1|T = c] - p(c|0)\mathbb{E}[Y_0|T = c]}{p(c|1) + p(a|1) - p(a|0)} + \frac{[p(a|1) - p(a|0)]\mathbb{E}[Y_1|T = a]}{p(c|1) + p(a|1) - p(a|0)} + \frac{[p(n|1) - p(n|0)]\mathbb{E}[Y_0|T = n]}{p(c|1) + p(a|1) - p(a|0)},$$

where $p(t|z)$ denotes $\mathbb{P}(T = t|Z = z)$, $t \in \{a, c, d, n\}$ and $z \in \{0, 1\}$. As can be seen, the IV estimand does not have a clear interpretation in this case. If the treatment effect is the same accross all individuals, i.e. $\alpha = Y_1 - Y_0$ is constant, the IV estimand is

$$\alpha_{IV} = \alpha + \frac{\mathbb{E}[Y_0|Z = 1] - \mathbb{E}[Y_0|Z = 0]}{p(c|1) + p(a|1) - p(a|0)}.$$

We observe that the IV estimand does not identify a causal effect unless $\mathbb{E}[Y_0|Z = 1] = \mathbb{E}[Y_0|Z = 0]$, which is unlikely to hold when the type is confounded.

Appendix B. Additional empirical results

B.1. **Relationship between IQ and college proximity.** Table 5 shows that ability as measured by IQ appears to be affected by the college proximity instrument in the NLSYM data. Cameron and Taber (2004) as well as Carneiro and Heckman (2002) have also shown that distance to college in the NLSY79 data is correlated with a measure of ability (Armed Forces Qualifcation Test (AFQT)).

Table 5. Relationship between IQ and College proximity

Dependent variable IQ	(1)	(2)
Near 4-year college	2.60	1.73
	(0.75)	(0.73)
Other Controls	No	Yes
n	2,061	1,619

Notes: The controls are parental education, age, race, and family structure.

B.2. **Results with college attendance (at least 13 years of education).**

B.2.1. *Some college vs. no college.* In this subsection, I consider college attendance as treatment variable and no college attendance as control group. The results are summarized in Table 6.

Table 6. Confidence sets for parameters

Parameters	Estimates	95% conf. LB	95% conf. UB
θ^1		1	9
θ^0		-8.6	-1
$\mathbb{E}[Y_1\|T=c]$		6.3437	6.4597
$\mathbb{E}[Y_0\|T=c]$		5.8791	6.0763
$LATE$		0.2674	0.5806
$2SLS$	1.2787***	0.8418	1.7155
	(0.2228)		

Standard errors (in parentheses); *** stands for 1% significant; conf.: confidence; LB: lower bound; UB: upper bound.

B.2.2. *Some college (at least 13 years) vs. high school (12 years).* In this subsection, I consider some college as treatment group and high school as control group. The results are summarized in Table 7.

TABLE 7. Confidence sets for parameters

Parameters	Estimates	95% conf. LB	95% conf. UB
θ^1		1	11
θ^0		-4.7	-1.5
$\mathbb{E}[Y_1\|T=c]$		6.3455	6.4713
$\mathbb{E}[Y_0\|T=c]$		6.0270	6.1171
$LATE$		0.2284	0.4443
$2SLS$	1.3702***	0.5952	2.1451
	(0.3952)		

Standard errors (in parentheses); *** stands for 1% significant;
conf.: confidence; LB: lower bound; UB: upper bound.

B.3. **Possible comparison with Card's (1995) results.** Card (1995) considered a linear constant return to years of schooling model where the return is the same for each individual and each additional year of education. Using college proximity as instrument, he found an estimate of 0.132 log points increase for each additional year of schooling. In this paper, I consider a different framework where the return could be nonlinear and heterogenous across individuals. Even though the two settings are different, I try an extrapolation of Card's (1995) results in Table 8 to allow for a possible comparison with my results. I consider three cases: college degree vs. no college degree, some college vs. no college, and some college vs. high school. I display the bounds on LATE that I get (in percent) using my methodology and an aproximate of Card's (1995) point-estimate in each case. For instance, to obtain a Card's estimate in the first case (college degree vs. no college degree), I compute the average difference in years of schooling for the two groups, which I multiply by Card's coefficient (0.132) to get the average return in log points. While Card's estimate for the case college degree vs. no college degree is outside my bounds, the estimates for the two other cases lie within my bounds.

TABLE 8. Comparison

Parameters	Estimates	Returns	my LB	my UB
Card (1995) coefficient for each year	0.132	14.1%		
mean(educ $\geq$ 16)-mean(educ < 16)	4.705	86.1%	38%	79%
mean(educ $\geq$ 13)-mean(educ < 13)	4.251	75.3%	31%	79%
mean(educ $\geq$ 13)-12	3.366	55.9%	26%	56%

Return for 4.705 years of schooling: exp(4.705*0.132)-1=86.1%;
LB: lower bound; UB: upper bound.

Appendix C. Extensions

C.1. **Relaxing monotonicity.** As before, denote $\alpha_1^0(w) \equiv \mathbb{P}(T = df|D = 0, Z = 1, W = w)$, $\alpha_0^0(w) \equiv \mathbb{P}(T = c|D = 0, Z = 0, W = w)$, $\alpha_1^1(w) \equiv \mathbb{P}(T = c|D = 1, Z = 1, W = w)$, and $\alpha_0^1(w) \equiv \mathbb{P}(T = df|D = 1, Z = 0, W = w)$. We have:

$$\begin{aligned}
F(y|0,1,w) &= \alpha_1^0(w)F_{0df}(y) + \left(1-\alpha_1^0(w)\right)F_{0n}(y),\\
F(y|0,0,w) &= \alpha_0^0(w)F_{0c}(y) + \left(1-\alpha_0^0(w)\right)F_{0n}(y),\\
F(y|1,1,w) &= \alpha_1^1(w)F_{1c}(y) + \left(1-\alpha_1^1(w)\right)F_{1a}(y),\\
F(y|1,0,w) &= \alpha_0^1(w)F_{1df}(y) + \left(1-\alpha_0^1(w)\right)F_{1a}(y).
\end{aligned}$$

Assumption 6 (RELM). *There exist w_{dz}^0 and w_{dz}^1 in the support $\mathcal{W}$ and y_{dz}^1 in the support $\mathcal{Y}$ such that $F(y_{dz}^1|d,z,w_{dz}^0) \neq F(y_{dz}^1|d,z,w_{dz}^1)$ for all $d,z \in \{0,1\}$.* □

Define

$$\Lambda_z^d(w) \equiv \frac{F(y_{dz}^1|d,z,w) - F(y_{dz}^1|d,z,w_{dz}^0)}{F(y_{dz}^1|d,z,w_{dz}^1) - F(y_{dz}^1|d,z,w_{dz}^0)},\ \phi_z^d \equiv \alpha_z^d(w_{dz}^0),\ \psi_z^d \equiv \alpha_z^d(w_{dz}^1) - \alpha_z^d(w_{dz}^0),$$

$$\overline{\Lambda}_z^d \equiv \sup_w \Lambda_z^d(w),\ \underline{\Lambda}_z^d \equiv \inf_w \Lambda_z^d(w),$$

$$r_z^d(y) \equiv \frac{f(y|d,z,w_{dz}^1)}{f(y|d,z,w_{dz}^0)},\ \underline{r}_z^d \equiv \inf_{y\in\mathcal{Y}} r_z^d(y),\ \overline{r}_z^d \equiv \sup_{y\in\mathcal{Y}} r_z^d(y),$$

$$f_{*z}^{\ d} \equiv -\frac{1}{\overline{r}_z^d - 1},\ \text{and}\ f^{*d}_{\ z} \equiv \frac{1}{1-\underline{r}_z^d}\ \text{for all}\ d,z \in \{0,1\}.$$

The following proposition holds.

Proposition 5. *Under CI and RELM, the following holds for all $y \in \mathcal{Y}$:*

$$\begin{aligned}
F_{0c}(y) &= F(y|0,0,w_{00}^0) + \left(\theta_0^0 - \eta_0^0\rho_1^0\right)\left[F(y|0,0,w_{00}^1) - F(y|0,0,w_{00}^0)\right],\\
F_{1c}(y) &= F(y|1,1,w_{11}^0) + \left(\theta_1^1 - \eta_1^1\rho_0^1\right)\left[F(y|1,1,w_{11}^1) - F(y|1,1,w_{11}^0)\right],
\end{aligned}$$

$$\begin{aligned}
F_{0n}(y) &= F(y|0,0,w_{00}^0) - \eta_0^0\rho_1^0\left[F(y|0,0,w_{00}^1) - F(y|0,0,w_{00}^0)\right],\\
F_{1a}(y) &= F(y|1,1,w_{11}^0) - \eta_1^1\rho_0^1\left[F(y|1,1,w_{11}^1) - F(y|1,1,w_{11}^0)\right],
\end{aligned}$$

$$\begin{aligned}
F_{0df}(y) &= F(y|0,1,w_{01}^0) + \left(\theta_1^0 - \rho_1^0\right)\left[F(y|0,1,w_{01}^1) - F(y|0,1,w_{01}^0)\right],\\
F_{1df}(y) &= F(y|1,0,w_{10}^0) + \left(\theta_0^1 - \rho_0^1\right)\left[F(y|1,0,w_{10}^1) - F(y|1,0,w_{10}^0)\right],
\end{aligned}$$

where

$$\begin{array}{rcccl}
f^1_{*1}+\eta^1_1\rho^1_0 & \leq & \min\left(\theta^1_1,0\right) & \leq & \underline{\Lambda}^1_1+\eta^1_1\rho^1_0,\\
\overline{\Lambda}^1_1+\eta^1_1\rho^1_0 & \leq & \max\left(\theta^1_1,0\right) & \leq & f^{*1}_1+\eta^1_1\rho^1_0,\\
f^0_{*0}+\eta^0_0\rho^0_1 & \leq & \min\left(\theta^0_0,0\right) & \leq & \underline{\Lambda}^0_0+\eta^0_0\rho^0_1,\\
\overline{\Lambda}^0_0+\eta^0_0\rho^0_1 & \leq & \max\left(\theta^0_0,0\right) & \leq & f^{*0}_0+\eta^0_0\rho^0_1,\\
f^0_{*1}+\rho^0_1 & \leq & \min\left(\theta^0_1,0\right) & \leq & \underline{\Lambda}^0_1+\rho^0_1,\\
\overline{\Lambda}^0_1+\rho^0_1 & \leq & \max\left(\theta^0_1,0\right) & \leq & f^{*0}_1+\rho^0_1,\\
f^1_{*0}+\rho^1_0 & \leq & \min\left(\theta^1_0,0\right) & \leq & \underline{\Lambda}^1_0+\rho^1_0,\\
\overline{\Lambda}^1_0+\rho^1_0 & \leq & \max\left(\theta^1_0,0\right) & \leq & f^{*1}_0+\rho^1_0,
\end{array}$$

and

$$\begin{aligned}
\eta^0_0 &= \frac{F(y^1_{00}|0,0,w^0_{00})}{F(y^1_{00}|0,1,w^0_{01})} * \frac{F(y^1_{01}|0,1,w^1_{01})-F(y^1_{01}|0,1,w^0_{01})}{F(y^1_{00}|0,0,w^1_{00})-F(y^1_{00}|0,0,w^0_{00})},\\
\eta^1_1 &= \frac{F(y^1_{11}|1,1,w^0_{11})}{F(y^1_{10}|1,0,w^0_{10})} * \frac{F(y^1_{10}|1,0,w^1_{10})-F(y^1_{10}|1,0,w^0_{10})}{F(y^1_{11}|1,1,w^1_{00})-F(y^1_{11}|1,1,w^0_{11})}.
\end{aligned}$$

□

Proof. Suppose Assumption CI and RELM holds. Then Theorem 1 by HKS applies:

$$\begin{aligned}
\alpha^1_1(w) &= \phi^1_1+\psi^1_1\Lambda^1_1(w),\\
F_{1a}(y) &= F(y|1,1,w^0_{11})-\frac{\phi^1_1}{\psi^1_1}\left[F(y|1,1,w^1_{11})-F(y|1,1,w^0_{11})\right],\\
F_{1c}(y) &= F(y|1,1,w^0_{11})+\frac{1-\phi^1_1}{\psi^1_1}\left[F(y|1,1,w^1_{11})-F(y|1,1,w^0_{11})\right],
\end{aligned}$$

where

$$\begin{array}{rcccl}
f^1_{*1} & \leq & \min\left((1-\phi^1_1)/\psi^1_1,-\phi^1_1/\psi^1_1\right) & \leq & \underline{\Lambda}^1_1,\\
\overline{\Lambda}^1_1 & \leq & \max\left((1-\phi^1_1)/\psi^1_1,-\phi^1_1/\psi^1_1\right) & \leq & f^{*1}_1.
\end{array}$$

Similarly, we have

$$\begin{aligned}
\alpha^0_0(w) &= \phi^0_0+\psi^0_0\Lambda^0_0(w),\\
F_{0n}(y) &= F(y|0,0,w^0_{00})-\frac{\phi^0_0}{\psi^0_0}\left[F(y|0,0,w^1_{00})-F(y|0,0,w^0_{00})\right],\\
F_{0c}(y) &= F(y|0,0,w^0_{00})+\frac{1-\phi^0_0}{\psi^0_0}\left[F(y|0,0,w^1_{00})-F(y|0,0,w^0_{00})\right],
\end{aligned}$$

where

$$\begin{array}{rcccl}
f^0_{*0} & \leq & \min\left((1-\phi^0_0)/\psi^0_0,-\phi^0_0/\psi^0_0\right) & \leq & \underline{\Lambda}^0_0,\\
\overline{\Lambda}^0_0 & \leq & \max\left((1-\phi^0_0)/\psi^0_0,-\phi^0_0/\psi^0_0\right) & \leq & f^{*0}_0;
\end{array}$$

$$\begin{aligned}
\alpha_1^0(w) &= \phi_1^0 + \psi_1^0 \Lambda_1^0(w), \\
F_{0n}(y) &= F(y|0,1,w_{01}^0) - \frac{\phi_1^0}{\psi_1^0}\left[F(y|0,1,w_{01}^1) - F(y|0,1,w_{01}^0)\right], \\
F_{0df}(y) &= F(y|0,1,w_{01}^0) + \frac{1-\phi_1^0}{\psi_1^0}\left[F(y|0,1,w_{01}^1) - F(y|0,1,w_{01}^0)\right],
\end{aligned}$$

where

$$\begin{array}{rcccl}
f_{*1}^0 & \leq & \min\left((1-\phi_1^0)/\psi_1^0, -\phi_1^0/\psi_1^0\right) & \leq & \underline{\Lambda}_1^0, \\
\overline{\Lambda}_1^0 & \leq & \max\left((1-\phi_1^0)/\psi_1^0, -\phi_1^0/\psi_1^0\right) & \leq & f_1^{*0};
\end{array}$$

$$\begin{aligned}
\alpha_0^1(w) &= \phi_0^1 + \psi_0^1 \Lambda_0^1(w), \\
F_{1a}(y) &= F(y|1,0,w_{10}^0) - \frac{\phi_0^1}{\psi_0^1}\left[F(y|1,0,w_{10}^1) - F(y|1,0,w_{10}^0)\right], \\
F_{1df}(y) &= F(y|1,0,w_{10}^0) + \frac{1-\phi_0^1}{\psi_0^1}\left[F(y|1,0,w_{10}^1) - F(y|1,0,w_{10}^0)\right],
\end{aligned}$$

where

$$\begin{array}{rcccl}
f_{*0}^1 & \leq & \min\left((1-\phi_0^1)/\psi_0^1, -\phi_0^1/\psi_0^1\right) & \leq & \underline{\Lambda}_0^1, \\
\overline{\Lambda}_0^1 & \leq & \max\left((1-\phi_0^1)/\psi_0^1, -\phi_0^1/\psi_0^1\right) & \leq & f_0^{*1}.
\end{array}$$

We can summarize the identified set as follows:

$$\begin{aligned}
F_{0c}(y) &= F(y|0,0,w_{00}^0) + \frac{1-\phi_0^0}{\psi_0^0}\left[F(y|0,0,w_{00}^1) - F(y|0,0,w_{00}^0)\right], \\
F_{1c}(y) &= F(y|1,1,w_{11}^0) + \frac{1-\phi_1^1}{\psi_1^1}\left[F(y|1,1,w_{11}^1) - F(y|1,1,w_{11}^0)\right],
\end{aligned}$$

$$\begin{aligned}
F_{0n}(y) &= F(y|0,0,w_{00}^0) - \frac{\phi_0^0}{\psi_0^0}\left[F(y|0,0,w_{00}^1) - F(y|0,0,w_{00}^0)\right], \\
F_{1a}(y) &= F(y|1,1,w_{11}^0) - \frac{\phi_1^1}{\psi_1^1}\left[F(y|1,1,w_{11}^1) - F(y|1,1,w_{11}^0)\right],
\end{aligned}$$

$$\begin{aligned}
F_{0df}(y) &= F(y|0,1,w_{01}^0) + \frac{1-\phi_1^0}{\psi_1^0}\left[F(y|0,1,w_{01}^1) - F(y|0,1,w_{01}^0)\right], \\
F_{1df}(y) &= F(y|1,0,w_{10}^0) + \frac{1-\phi_0^1}{\psi_0^1}\left[F(y|1,0,w_{10}^1) - F(y|1,0,w_{10}^0)\right],
\end{aligned}$$

where

$$\begin{array}{rcccl} f^1_{*1} & \leq & \min\left((1-\phi^1_1)/\psi^1_1, -\phi^1_1/\psi^1_1\right) & \leq & \underline{\Lambda}^1_1, \\ \overline{\Lambda}^1_1 & \leq & \max\left((1-\phi^1_1)/\psi^1_1, -\phi^1_1/\psi^1_1\right) & \leq & f^{*1}_1, \\ f^0_{*0} & \leq & \min\left((1-\phi^0_0)/\psi^0_0, -\phi^0_0/\psi^0_0\right) & \leq & \underline{\Lambda}^0_0, \\ \overline{\Lambda}^0_0 & \leq & \max\left((1-\phi^0_0)/\psi^0_0, -\phi^0_0/\psi^0_0\right) & \leq & f^{*0}_0, \\ f^0_{*1} & \leq & \min\left((1-\phi^0_1)/\psi^0_1, -\phi^0_1/\psi^0_1\right) & \leq & \underline{\Lambda}^0_1, \\ \overline{\Lambda}^0_1 & \leq & \max\left((1-\phi^0_1)/\psi^0_1, -\phi^0_1/\psi^0_1\right) & \leq & f^{*0}_1, \\ f^1_{*0} & \leq & \min\left((1-\phi^1_0)/\psi^1_0, -\phi^1_0/\psi^1_0\right) & \leq & \underline{\Lambda}^1_0, \\ \overline{\Lambda}^1_0 & \leq & \max\left((1-\phi^1_0)/\psi^1_0, -\phi^1_0/\psi^1_0\right) & \leq & f^{*1}_0, \end{array}$$

and

$$\begin{aligned} \frac{\phi^0_0}{\psi^0_0}\Big/\frac{\phi^0_1}{\psi^0_1} &= \frac{F(y^1_{00}|0,0,w^0_{00})}{F(y^1_{00}|0,1,w^0_{01})} * \frac{F(y^1_{01}|0,1,w^1_{01}) - F(y^1_{01}|0,1,w^0_{01})}{F(y^1_{00}|0,0,w^1_{00}) - F(y^1_{00}|0,0,w^0_{00})}, \\ \frac{\phi^1_1}{\psi^1_1}\Big/\frac{\phi^1_0}{\psi^1_0} &= \frac{F(y^1_{11}|1,1,w^0_{11})}{F(y^1_{10}|1,0,w^0_{10})} * \frac{F(y^1_{10}|1,0,w^1_{10}) - F(y^1_{10}|1,0,w^0_{10})}{F(y^1_{11}|1,1,w^1_{00}) - F(y^1_{11}|1,1,w^0_{11})}. \end{aligned}$$

Set $\theta^j_i = \frac{1}{\psi^j_i}$, and $\rho^j_i = \frac{\phi^j_i}{\psi^j_i}$ for $i, j = 0, 1$. Then the result follows. This completes the proof. □

C.2. **Continuous instruments.** Suppose that I have a continuous instrument X. Define $Z = \mathbb{1}\{X > x_0\}$ and $W = (X\mathbb{1}\{X > x_0\}, X\mathbb{1}\{X \leq x_0\})$ for some $x_0 \in \mathcal{X} = Supp(X)$. Then, I can apply the methodology described in this paper to derive bounds on the potential outcome distributions for compliers defined based on Z. From there, I get bounds on the ATE: $LB(x_0) \leq ATE \leq UB(x_0)$. Thus, tighter bounds on ATE can be obtained by varying x_0 over the support of X:

$$\sup_{x_0 \in \mathcal{X}} LB(x_0) \leq ATE \leq \inf_{x_0 \in \mathcal{X}} UB(x_0).$$

C.3. **Accounting for sample selection.** Now, I extend the model to account for sample selection, for instance self-selection into employment. Then, we have the following model:

$$\begin{array}{rcl} Y & = & SY^*, \\ Y^* & = & Y^*_1 D + Y^*_0 (1-D), \\ S & = & S_1 D + S_0 (1-D), \\ D & = & D_1 Z + D_0 (1-Z), \end{array} \tag{C.1}$$

where S is the selection variable (e.g. employment status); S_0 and S_1 the potential selection when the treatment D is 0 and 1, respectively; Y^* is the outcome variable (e.g. wage-offer), Y^*_1 and Y^*_0 the potential outcomes. Y is observed, Y^* is not.

Define the type variable as $T^* = (D_0, D_1, S_0, S_1)$. Then, we have 16 types: $\{a, c, df, n\} \times \{EE, EN, NE, NN\}$, where EE, EN, NE and NN are defined in Table 9.

I use the following assumptions for identification.

TABLE 9. Employment status subgroups

subgroups	S_0	S_1	Notion
EE	1	1	Always-employed
EN	1	0	Employed-nonemployed
NE	0	1	Nonemployed-employed
NN	0	0	Never-employed

Assumption 7. *The vector (Z, W) is independent of Y_d^* given the type T^*, i.e., $(Z, W) \perp\!\!\!\perp Y_d^* | T^*$, $d = 0, 1$.* □

Assumption 8 (Monotonicity of S in D)**.** *$S_1 \geq S_0$ a.s. (i.e., there are no EN).* □

I derive bounds on the local average treatment effect and the local quantile treatment effects for the always-employed compliers: $\mathbb{E}[Y_1^* - Y_0^* | T^* = cEE]$ and $F_{1\,cEE}^{*-1}(\alpha) - F_{0\,cEE}^{*-1}(\alpha)$, respectively. Denote $F_{dt}^* \equiv F(Y_d^* | T = t)$, $d \in \{0, 1\}$, $t \in \{a, c, df, n\} \times \{EE, EN, NE, NN\}$. Under Assumptions 2 (no defiers), 7 (conditional independence), and 8 (no employed-nonemployed), $T^* \in \{a, c, n\} \times \{EE, NE, NN\}$ and we have for all $w \in \mathcal{W}$:

$$F_Y(y|S = 1, D = 0, Z = 1, W = w) = F_{Y_0^*}(y|T^* = nEE),$$

$$\begin{aligned} F_Y(y|S = 1, D = 0, Z = 0, W = w) &= \mathbb{P}(D_1 = 0|S_0 = 1, D_0 = 0, Z = 0, W = w) F_{Y_0^*}(y|T^* = nEE) \\ &+ \mathbb{P}(D_1 = 1|S_0 = 1, D_0 = 0, Z = 0, W = w) F_{Y_0^*}(y|T^* = cEE), \end{aligned}$$

$$\begin{aligned} F_Y(y|S = 1, D = 1, Z = 0, W = w) &= \mathbb{P}(S_0 = 0|S_1 = 1, D_0 = 1, Z = 0, W = w) F_{Y_1^*}(y|T^* = aNE) \\ &+ \mathbb{P}(S_0 = 1|S_1 = 1, D_0 = 1, Z = 0, W = w) F_{Y_1^*}(y|T^* = aEE), \end{aligned}$$

and

$$\begin{aligned} F_Y(y|S = 1, D = 1, Z = 1, W = w) &= \mathbb{P}(S_0 = 0, D_0 = 0|S_1 = 1, D_1 = 1, Z = 1, W = w) F_{Y_1^*}(y|T^* = cNE) \\ &+ \mathbb{P}(S_0 = 0, D_0 = 1|S_1 = 1, D_1 = 1, Z = 1, W = w) F_{Y_1^*}(y|T^* = aNE) \\ &+ \mathbb{P}(S_0 = 1, D_0 = 0|S_1 = 1, D_1 = 1, Z = 1, W = w) F_{Y_1^*}(y|T^* = cEE) \\ &+ \mathbb{P}(S_0 = 1, D_0 = 1|S_1 = 1, D_1 = 1, Z = 1, W = w) F_{Y_1^*}(y|T^* = aEE). \end{aligned}$$

Identification. I rewrite the model as follows:

$$
\begin{aligned}
F(y|S=1,D=0,Z=1,W=w) &= F^*_{0nEE}(y), && \text{(C.2)}\\
F(y|S=1,D=0,Z=0,W=w) &= \tau(w)F^*_{0nEE}(y) + (1-\tau(w))\,F^*_{0cEE}(y), && \text{(C.3)}\\
F(y|S=1,D=1,Z=0,W=w) &= \gamma(w)F^*_{1aNE}(y) + (1-\gamma(w))\,F^*_{1aEE}(y), && \text{(C.4)}\\
F(y|S=1,D=1,Z=1,W=w) &= \lambda_0(w)F^*_{1cNE}(y) + \lambda_1(w)F^*_{1aNE}(y)\\
&\quad + \lambda_2(w)F^*_{1cEE}(y) + \lambda_3(w)F^*_{1aEE}(y). && \text{(C.5)}
\end{aligned}
$$

Equation (C.2) shows that the distribution of the potential outcome of the untreated that are never-takers and always-employed (F^*_{0nEE}) is point-identified. On the other hand, the distributions F^*_{0cEE}, F^*_{1aNE}, F^*_{1aEE}, F^*_{1cNE} and F^*_{1cEE} are only partially identified. I use Theorem 1 in the main text to derive sharp bounds on F^*_{0nEE} and F^*_{0cEE} combining Equations (C.2) and (C.3). Afterwards, I apply Theorem 1 of HKS on the mixture model of Equation (C.4) to get bounds on F^*_{1aNE} and F^*_{1aEE}, while I apply their Theorem 2 on Equation (C.5) to derive bounds on F^*_{1cNE}, F^*_{1aNE}, F^*_{1cEE} and F^*_{1aEE}. I then use the fact F^*_{1aNE} and F^*_{1aEE} must satisfy the constraints imposed by (C.4) and (C.5) to get sharp bounds on the distributions F^*_{1cNE}, F^*_{1aNE}, F^*_{1cEE} and F^*_{1aEE}. Hence, I get sharp bounds on the average treatment effect for the always-employed compliers $\mathbb{E}[Y^*_1 - Y^*_0|T = cEE]$.

Let me start with Equation (C.5). For any (y, w, w_0),

$$
\begin{aligned}
F(y|S=1,D=1,Z=1,W=w) &- F(y|S=1,D=1,Z=1,W=w_0)\\
&= (\lambda_1(w)-\lambda_1(w_0))\,(F^*_{1aNE}(y) - F^*_{1cNE}(y))\\
&+ (\lambda_2(w)-\lambda_2(w_0))\,(F^*_{1cEE}(y) - F^*_{1cNE}(y))\\
&+ (\lambda_3(w)-\lambda_3(w_0))\,(F^*_{1aEE}(y) - F^*_{1cNE}(y)).
\end{aligned}
$$

Denote

$$
\boldsymbol{\psi}(w) \equiv \begin{bmatrix} \lambda_1(w)-\lambda_1(w_0) \\ \lambda_2(w)-\lambda_2(w_0) \\ \lambda_3(w)-\lambda_3(w_0) \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\delta}(y) \equiv \begin{bmatrix} F^*_{1aNE}(y) - F^*_{1cNE}(y) \\ F^*_{1cEE}(y) - F^*_{1cNE}(y) \\ F^*_{1aEE}(y) - F^*_{1cNE}(y) \end{bmatrix}.
$$

Then

$$
F(y|S=1,D=1,Z=1,W=w) - F(y|S=1,D=1,Z=1,W=w_0) = \boldsymbol{\psi}(w)^t\boldsymbol{\delta}(y),
$$

where $\boldsymbol{\psi}(w)^t$ denote the transpose of $\boldsymbol{\psi}(w)$. The following assumption is needed for identification.

Assumption 9 (Relevance)**.** *There exist (w_0, w_1, w_2, w_3) in the support $\mathcal{W}$ of W such that the 3×3 matrix $\boldsymbol{\Psi}$ with jth column $\boldsymbol{\psi}(w_j)$, $j = 1, 2, 3$, is invertible.* □

Assumption 9 is the same as Assumption 5 in HKS. It requires that the support $\mathcal{W}$ of the instrument W have four distincts points. Denote $F(y|s, d, z, w) \equiv F(y|S = s, D = d, Z = z, W = w)$

for all $s, d, z = 0, 1$, $w \in \mathcal{W}$, and

$$\boldsymbol{h_u}(y) \equiv \begin{bmatrix} F(y|1,1,1,w_1) - F(y|1,1,1,w_0) \\ F(y|1,1,1,w_2) - F(y|1,1,1,w_0) \\ F(y|1,1,1,w_3) - F(y|1,1,1,w_0) \end{bmatrix}.$$

Therefore, $\boldsymbol{h_u}(y) = \boldsymbol{\Psi}^t \boldsymbol{\delta}(y)$, and $\boldsymbol{\delta}(y) = \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y)$. Denote $\boldsymbol{\phi} \equiv [\lambda_1(w_0)\ \lambda_2(w_0)\ \lambda_3(w_0)]^t$. From Equation (C.5), I have $F(y|1,1,1,w_0) = F^*_{cNE}(y) + \boldsymbol{\phi}^t \boldsymbol{\delta}(y)$. And I know that

$$\begin{bmatrix} F^*_{1aNE}(y) \\ F^*_{1cEE}(y) \\ F^*_{1aEE}(y) \end{bmatrix} = \begin{bmatrix} F^*_{1cNE}(y) \\ F^*_{1cNE}(y) \\ F^*_{1cNE}(y) \end{bmatrix} + \boldsymbol{\delta}(y).$$

Therefore,

$$\begin{aligned} F^*_{1cNE}(y) &= F(y|1,1,1,w_0) + (\boldsymbol{e}_0 - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y), \\ F^*_{1aNE}(y) &= F(y|1,1,1,w_0) + (\boldsymbol{e}_1 - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y), \\ F^*_{1cEE}(y) &= F(y|1,1,1,w_0) + (\boldsymbol{e}_2 - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y), \\ F^*_{1aEE}(y) &= F(y|1,1,1,w_0) + (\boldsymbol{e}_3 - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y) \end{aligned} \tag{C.6}$$

where $\boldsymbol{e}_0 \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, $\boldsymbol{e}_1 \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\boldsymbol{e}_2 \equiv \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\boldsymbol{e}_3 \equiv \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Now, assume that the following holds.

Assumption 10 (Rank)**.** *There exist (y_1, y_2, y_3) in the support $\mathcal{Y}$ of Y such that the 3×3 matrix $\boldsymbol{\Delta}$ with jth column $\boldsymbol{\delta}(y_j)$, $j = 1, 2, 3$, is invertible.* □

Assumption 10 is the same as Assumption 6 in HKS. It requires that the support $\mathcal{Y}$ of the outcome Y have four distincts points, which is in general true when the ouctome is wage, earnings. However, it excludes binary outcomes. Lemma 1 in HKS states that Assumptions 9 and 10 are jointly testable since they are equivalent to Assumption 11 below.

Assumption 11. *There exist (w_0, w_1, w_2, w_3) in the support $\mathcal{W}$ and (y_1, y_2, y_3) in the support $\mathcal{Y}$ such that the 3×3 matrix $\boldsymbol{H}$ with jth column $\boldsymbol{h_u}(y_j)$, $j = 1, 2, 3$, is invertible.* □

Note that the matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Delta}$ are both unobservable quantities while the matrix $\boldsymbol{H}$ is an observable one. Since Assumption 11 relates directly to the data, the researcher can check whether it holds or not. If it does, then Assumptions 9 and 10 hold and we can identifiy the mixture weights $\lambda_1(w)$, $\lambda_2(w)$, and $\lambda_3(w)$ as follows:

$$F(y|1,1,1,w) - F(y|1,1,1,w_0) = \boldsymbol{\psi}(w)^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(y).$$

Denote $\boldsymbol{h_l}(w) \equiv \begin{bmatrix} F(y_1|1,1,1,w) - F(y_1|1,1,1,w_0) \\ F(y_2|1,1,1,w) - F(y_2|1,1,1,w_0) \\ F(y_3|1,1,1,w) - F(y_3|1,1,1,w_0) \end{bmatrix}$, $\boldsymbol{\lambda}(w) \equiv \begin{bmatrix} \lambda_1(w) \\ \lambda_2(w) \\ \lambda_3(w) \end{bmatrix}$. Therefore,

$$\boldsymbol{h_l}(w) = \boldsymbol{\psi}(w)^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{H}, \tag{C.7}$$

and

$$\boldsymbol{\lambda}(w) = \boldsymbol{\phi} + \boldsymbol{\psi}(w) = \boldsymbol{\phi} + \boldsymbol{\Psi}\boldsymbol{\Lambda}(w), \tag{C.8}$$

where the first equality of (C.8) holds by definition and the second from (C.7), with $\boldsymbol{\Lambda}(w) \equiv \left(\boldsymbol{H}^t\right)^{-1} \boldsymbol{h_l}(w)$.

Equations (C.6) and (C.8) combined with sharp bounds for $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ yield the identified set for the distributions $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, $F^*_{1aEE}(y)$ and the vector of mixture weights $\boldsymbol{\lambda}(w)$. As in HKS, bounds for $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ are obtained by imposing probability constraints on $\boldsymbol{\lambda}(w)$, and monotonicity constraints on the distributions $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, $F^*_{1aEE}(y)$.

Probability constraints: $0 \leq \boldsymbol{\lambda}(w)$ and $\mathbf{1}^t\boldsymbol{\lambda}(w) \leq 1$.

Monotonicity constraints: $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, $F^*_{1aEE}(y)$ should be nondecreasing, right-continuous, have left and right limits 0 and 1. The last two properties are satisfied from Equations (C.6). That is why I only need the monotonicity constraints. For convenience, suppose now that the following assumption holds.

Assumption 12. *The outcome variable Y is continuously distributed conditional on (S, D, Z, W).* □

Denote $f(y|s,d,z,w)$ the density of Y conditional on $(S = s, D = d, Z = z, W = w)$, and $\boldsymbol{h_l}'(y)$ the derivative of $\boldsymbol{h_l}(y)$. The density of $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, $F^*_{1aEE}(y)$ should be positive, i.e.,

$$\text{for all } j = 0,1,2,3,\ f(y|1,1,1,w_0) + (\boldsymbol{e}_j - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_l}'(y) \geq 0 \text{ for all } y \in \mathcal{Y}.$$

These restrictions are linear in $\boldsymbol{\Omega}_j \equiv (\boldsymbol{e}_j - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1}$. Therefore, they only need to be checked at the extreme points of the range of the function $\boldsymbol{\Pi}(y) \equiv -\boldsymbol{h_l}'(y)/f(y|1,1,1,w_0)$ defined on the support of Y given $(S = 1, D = 1, Z = 1, W = w_0)$.

Under Assumptions 7, 11, and 12, Theorem 2 in HKS applies and the identification regions for $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, $F^*_{1aEE}(y)$ and $\boldsymbol{\lambda}(w)$ are the twelve parameter family defined by Equations C.6 and (C.8) along with the following constraints on $(\boldsymbol{\phi}, \boldsymbol{\Psi})$:

- The linear constraints $\boldsymbol{\phi} + \boldsymbol{\Psi}\boldsymbol{e} > 0$ and $\mathbf{1}^t(\boldsymbol{\phi} + \boldsymbol{\Psi}\boldsymbol{e}) < 1$ for all extreme points $\boldsymbol{e}$ of the convex hull of the range of the identified function $w \mapsto \boldsymbol{\Lambda}(w) = (\boldsymbol{H}^t)^{-1}\boldsymbol{h_l}(w)$.

- For all extreme points $\boldsymbol{\pi}$ of the convex hull of the range of the known function $y \mapsto \mathbf{\Pi}(y) \equiv -\boldsymbol{h_u}'(y)/f(y|1,1,1,w_0)$,

$$\boldsymbol{\pi}^t \mathbf{\Psi}^{-1}(\boldsymbol{e}_j - \boldsymbol{\phi}) \leq 1 \text{ for all } j = 0,1,2,3. \tag{C.9}$$

I now consider Equation (C.4):

$$F(Y|S=1,D=1,Z=0,W=w) = \gamma(w)F^*_{1aNE} + (1-\gamma(w))\,F^*_{1aEE}.$$

As before, the following assumption is needed for identification.

Assumption 13. *There exist $\tilde{w}_0$ and $\tilde{w}_1$ in the support $\mathcal{W}$ and $\tilde{y}_1$ in the support $\mathcal{Y}$ such that $F(\tilde{y}_1|1,1,0,\tilde{w}_0) \neq F(\tilde{y}_1|1,1,0,\tilde{w}_1)$.* □

Define

$$\Lambda_1(w) \equiv \frac{F(\tilde{y}_1|1,1,0,w) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_0) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)},\ \tilde{\phi}_1 \equiv \gamma(\tilde{w}_0),\ \tilde{\psi}_1 \equiv \gamma(\tilde{w}_1) - \gamma_1(\tilde{w}_0),$$

$$\overline{\Lambda}_1 \equiv \sup_w \Lambda_1(w),\ \underline{\Lambda}_1 \equiv \inf_w \Lambda_1(w),$$

$$r_1(y) \equiv \frac{f(y|1,1,0,\tilde{w}_1)}{f(y|1,1,0,\tilde{w}_0)},\ \underline{r}_1 \equiv \inf_{y\in\mathcal{Y}} r_1(y),\ \overline{r}_1 \equiv \sup_{y\in\mathcal{Y}} r_1(y),$$

$$f_{*1} \equiv -\frac{1}{\overline{r}_1 - 1},\ \text{and}\ f^*_1 \equiv \frac{1}{1-\underline{r}_1}.$$

Then I can apply Theorem 1 by HKS:

$$\gamma(w) = \tilde{\phi}_1 + \tilde{\psi}_1\Lambda_1(w), \tag{C.10}$$

$$F^*_{1aEE}(y) = F(y|1,1,0,\tilde{w}_0) - \frac{\tilde{\phi}_1}{\tilde{\psi}_1}\left[F(y|1,1,0,\tilde{w}_1) - F(y|1,1,0,\tilde{w}_0)\right], \tag{C.11}$$

$$F^*_{1aNE}(y) = F(y|1,1,0,\tilde{w}_0) + \frac{1-\tilde{\phi}_1}{\tilde{\psi}_1}\left[F(y|1,1,0,\tilde{w}_1) - F(y|1,1,0,\tilde{w}_0)\right], \tag{C.12}$$

where

$$f_{*1} \leq \min\left((1-\tilde{\phi}_1)/\tilde{\psi}_1, -\tilde{\phi}_1/\tilde{\psi}_1\right) \leq \underline{\Lambda}_1, \tag{C.13}$$

and

$$\overline{\Lambda}_1 \leq \max\left((1-\tilde{\phi}_1)/\tilde{\psi}_1, -\tilde{\phi}_1/\tilde{\psi}_1\right) \leq f^*_1. \tag{C.14}$$

Combining Equations (C.6) and (C.11), I get

$$-\frac{\tilde{\phi}_1}{\tilde{\psi}_1} = \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_3 - \boldsymbol{\phi})^t\left(\mathbf{\Psi}^t\right)^{-1}\boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}.$$

Similarly, combining (C.6) and (C.12), I obtain

$$\frac{1-\tilde{\phi}_1}{\tilde{\psi}_1} = \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_1-\boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}.$$

Therefore, the constraints (C.13) and (C.14) become

$$\begin{aligned} f_{*1} \quad \leq \quad \min\Bigg[& \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_1-\boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}, \\ & \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_3-\boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)} \Bigg] \leq \underline{\Lambda}_1, \end{aligned} \tag{C.15}$$

and

$$\begin{aligned} \overline{\Lambda}_1 \quad \leq \quad \max\Bigg[& \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_1-\boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}, \\ & \frac{F(\tilde{y}_1|1,1,1,w_0) + (\boldsymbol{e}_3-\boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \boldsymbol{h_u}(\tilde{y}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)}{F(\tilde{y}_1|1,1,0,\tilde{w}_1) - F(\tilde{y}_1|1,1,0,\tilde{w}_0)} \Bigg] \leq f_1^*, \end{aligned} \tag{C.16}$$

respectively.

Thus, the following proposition holds.

Proposition 6. *Under Assumptions 2, 7, 8, 11, 12, and 13, the identified set for $F^*_{1cNE}(y)$, $F^*_{1aNE}(y)$, $F^*_{1cEE}(y)$, and $F^*_{1aEE}(y)$ is given by equations (C.6), along with the constraints (C.15) and (C.16) on $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ such that:*

- *For all extreme points $\boldsymbol{e}$ of the convex hull of the range of the identified function $w \mapsto \boldsymbol{\Lambda}(w) = (\boldsymbol{H}^t)^{-1}\boldsymbol{h_r}(w)$, $\boldsymbol{\phi} + \boldsymbol{\Psi}\boldsymbol{e} > 0$ and $\mathbf{1}^t(\boldsymbol{\phi} + \boldsymbol{\Psi}\boldsymbol{e}) < 1$.*
- *For all extreme points $\boldsymbol{\pi}$ of the convex hull of the range of the known function $y \mapsto \boldsymbol{\Pi}(y) \equiv -\boldsymbol{h_u}'(y)/f(y|1,1,1,w_0)$, $\boldsymbol{\pi}^t\boldsymbol{\Psi}^{-1}(\boldsymbol{e}_j - \boldsymbol{\phi}) \leq 1$ for all $j = 0,1,2,3$.*

□

Finally, I derive bounds on F^*_{0cEE} using the two equations (C.2) and (C.3) as I do in Subsection 3.1. Indeed, the model is as follows:

$$\begin{aligned} F(Y|S=1,D=0,Z=0,W=w) &= \tau(w)F^*_{0nEE} + (1-\tau(w))\,F^*_{0cEE}, \\ F(Y|S=1,D=0,Z=1,W=w) &= F^*_{0nEE}. \end{aligned}$$

As before, similar to Assumption 13, the following is needed for identification.

Assumption 14. *There exist $\overline{w}_0$ and $\overline{w}_1$ in the support $\mathcal{W}$ and $\overline{y}_1$ in the support $\mathcal{Y}$ such that $F(\overline{y}_1|1,0,0,\overline{w}_0) \neq F(\overline{y}_1|1,0,0,\overline{w}_1)$.* □

Define

$$\Lambda_0(w) \equiv \frac{F(\overline{y}_1|1,0,0,w) - F(\overline{y}_1|1,0,0,\overline{w}_0)}{F(\overline{y}_1|1,0,0,\overline{w}_0) - F(\overline{y}_1|1,0,0,\overline{w}_0)},\ \overline{\phi}_0 \equiv \gamma(\overline{w}_0),\ \overline{\psi}_0 \equiv \gamma(\overline{w}_1) - \gamma_1(\overline{w}_0),$$

$$\overline{\Lambda}_0 \equiv \sup_w \Lambda_0(w),\ \underline{\Lambda}_0 \equiv \inf_w \Lambda_0(w),$$

$$r_0(y) \equiv \frac{f(y|1,0,0,\overline{w}_1)}{f(y|1,0,0,\overline{w}_0)},\ \underline{r}_0 \equiv \inf_{y\in\mathcal{Y}} r_0(y),\ \overline{r}_0 \equiv \sup_{y\in\mathcal{Y}} r_0(y),$$

$$f_{*0} \equiv -\frac{1}{\overline{r}_0 - 1},\ \text{and}\ f_0^* \equiv \frac{1}{1-\underline{r}_0}.$$

Therefore, I use Theorem 1 in the main text to obtain the following bounds:

$$\begin{aligned}
\tau(w) &= \frac{1}{\theta_0}\left(\eta_0 + \Lambda_0(w)\right), \\
F^*_{0nEE}(y) &= F(y|1,0,1), \\
F^*_{0cEE}(y) &= F(y|1,0,0,\overline{w}_0) + (\theta_0 - \eta_0)\left[F(y|1,0,0,\overline{w}_1) - F(y|1,0,0,\overline{w}_0)\right],
\end{aligned}$$

where

$$\eta_0 = \frac{F(\overline{y}_1|1,0,0,\overline{w}_0) - F(\overline{y}_1|1,0,1)}{F(\overline{y}_0|1,0,0,\overline{w}_1) - F(\overline{y}_1|1,0,0,\overline{w}_0)},$$

$$f_{*0} + \eta_0 \leq \min(\theta_0, 0) \leq \underline{\Lambda}_0 + \eta_0, \tag{C.17}$$

and

$$\overline{\Lambda}_0 + \eta_0 \leq \max(\theta_0, 0) \leq f_0^* + \eta_0. \tag{C.18}$$

Denote $\mathbf{\Theta_0}$ the set of all parameters θ_0 that satisfy constraints (C.17) and (C.18), $\mathbf{\Omega}$ the set of all $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ that satisfy all constraints in Proposition 6, and $F^{\theta}_{dT^*}$ the distribution of Y^*_d conditional on the type T^* that is associated with the parameter θ.

Proposition 7. *Under Assumptions 2, 7, 8, 11, 12, 13, and 14, we have the following bounds for the average and quantile treatment effects for the always-employed compliers:*

$$\inf_{\omega\in\mathbf{\Omega}} \mathbb{E}\left[F^{\omega}_{1cEE}\right] - \sup_{\theta\in\mathbf{\Theta_0}} \mathbb{E}\left[F^{\theta}_{0cEE}\right] \leq \mathbb{E}\left[Y^*_1 - Y^*_0 | T^* = cEE\right] \leq \sup_{\omega\in\mathbf{\Omega}} \mathbb{E}\left[F^{\omega}_{1cEE}\right] - \inf_{\theta\in\mathbf{\Theta_0}} \mathbb{E}\left[F^{\theta}_{0cEE}\right],$$

and

$$\inf_{\omega\in\mathbf{\Omega}} \left(F^{\omega}_{1cEE}\right)^{-1}(\alpha) - \sup_{\theta\in\mathbf{\Theta_0}} \left(F^{\theta}_{0cEE}\right)^{-1}(\alpha) \leq \left(F^{*-1}_{1cEE} - F^{*-1}_{0cEE}\right)(\alpha) \leq \sup_{\omega\in\mathbf{\Omega}} \left(F^{\omega}_{1cEE}\right)^{-1}(\alpha) - \inf_{\theta\in\mathbf{\Theta_0}} \left(F^{\theta}_{0cEE}\right)^{-1}(\alpha),$$

for all $\alpha \in (0,1)$.

These bounds are sharp. □

C.4. **Allowing for misclassified treatment.** In this subsection, I allow for measurement errors on the treatment. Let D^* be the true (unobserved) treatment. I consider the same specification as in Ura (2015):

$$\begin{cases} D^* = D_1^* Z + D_0^*(1-Z) \\ Y = Y_1 D^* + Y_0(1-D^*) \\ D = D_1 D^* + D_0(1-D^*) \end{cases} \tag{C.19}$$

Assume that $(Z,W) \perp\!\!\!\perp (Y_{d^*}, D_{d^*})|(D_0^*, D_1^*)$ and $D_1^* \geq D_0^*$ almost surely. Denote $F(y,d|Z=z,W=w) \equiv \mathbb{P}(Y \leq y, D=d|Z=z,W=w)$ and $F_{d^*t^*}(y,d) \equiv \mathbb{P}(Y_{d^*} \leq y, D_{d^*}=d|T^*=t^*)$, where $T^* \equiv (D_0^*, D_1^*) \in \{a,c,d,n\}$ is the type variable. One can show the following:

$$\begin{aligned}
F(y,1|Z=1,W=w) &= \lambda_{10}^1(w)F_{1a}(y,1) + \lambda_{11}^1(w)F_{1c}(y,1) + (1-\lambda_{10}^1(w)-\lambda_{11}^1(w))F_{0n}(y,1), \\
F(y,0|Z=1,W=w) &= \lambda_{10}^0(w)F_{1a}(y,0) + \lambda_{11}^0(w)F_{1c}(y,0) + (1-\lambda_{10}^1(w)-\lambda_{11}^0(w))F_{0n}(y,0), \\
F(y,1|Z=0,W=w) &= \lambda_{00}^1(w)F_{1a}(y,1) + \lambda_{01}^1(w)F_{0c}(y,1) + (1-\lambda_{00}^1(w)-\lambda_{01}^1(w))F_{0n}(y,1), \\
F(y,0|Z=0,W=w) &= \lambda_{00}^0(w)F_{1a}(y,0) + \lambda_{01}^0(w)F_{0c}(y,0) + (1-\lambda_{00}^0(w)-\lambda_{01}^0(w))F_{0n}(y,0),
\end{aligned}$$

where $0 \leq \lambda_{lk}^j \leq 1$ and $0 \leq 1-\lambda_{l0}^j(w)-\lambda_{l1}^j(w) \leq 1$ for $j,l,k \in \{0,1\}$. The same technique that I use throughout the paper can also be used to derive sharp bounds on the LATE defined here as $\mathbb{E}[Y_1 - Y_0|D_0^*=0, D_1^*=1]$.

References

Altonji, J., T. Elder, and C. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113 (1), 151–184.

Altonji, J. G., T. Conley, T. Elder, and C. Taber (2011): "Methods for using selection on observed variables to address selection on unobserved variables.," *Yale University Department of Economics Working Paper*.

Andrews, D. W. K., W. Kim, and X. Shi (2016): "Stata Commands for Testing Conditional Moment Inequalities/Equalities," *Unpublished manuscript*.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81, 609–666.

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(434), 444–455.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2008): "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan-Ganz Catheterization," *American Economic Review: Papers and Proceedings*, 98(2), 351–356.

Cameron, S. V., and J. J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106(2), 262–333.

CAMERON, S. V., and C. TABER (2004): "Estimation of Educational Borrowing Constraints Using Returns to Schooling," *Journal of Political Economy*, 112(1), 132–182.

CARD, D. (1995): *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. in Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp, ed, by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press.

CARD, D. (2001): "Estimating the Return to Schooling: Progress on some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160.

CARNEIRO, P., and J. J. HECKMAN (2002): "The Evidence on Credit Constraints in Post-Secondary Schooling," *Economic Journal*, 112(482), 705–734.

CARNEIRO, P., J. J. HECKMAN, and E. VYTLACIL (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78(1), 377–394.

CARNEIRO, P., J. J. HECKMAN, and E. VYTLACIL (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

CARNEIRO, P., and S. LEE (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149, 191–208.

CHEN, L.-Y., and J. SZROETER (2014): "Testing Multiple Inequality Hypotheses: A Smoothed Indicator Approach," *Journal of Econometrics*, 178, 678–693.

CHEN, X., and C. A. FLORES (2015): "Bounds on Treatment Effects in the Presence of Sample Selection and Noncompliance: The Wage Effects of Job Corps," *Journal of Business and Economic Statistics*, 33(4), 523–540.

CHEN, X., C. A. FLORES, and A. FLORES-LAGUNES (2012): "Bounds on Population Average Treatment Effects with an Instrumental Variable," *Unpublished manuscript*.

CHERNOZHUKOV, V., W. KIM, S. LEE, and A. M. ROSEN (2015): "Implementing Intersection Bounds in Stata," *Stata Journal*, 15(1), 21–44.

CHERNOZHUKOV, V., S. LEE, and A. M. ROSEN (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

CHIBURIS, R. C. (2010): "Semiparametric Bounds on Treatment Effects," *Journal of Econometrics*, 159, 267–275.

CONLEY, T. G., C. B. HANSEN, and P. E. ROSSI (2012): "Plausibly Exogenous," *The Review of Economics and Statistics*, 94(1), 260–272.

DE CHAISEMARTIN, C. (2017): "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 8, 367–396.

DITRAGLIA, F. J., and C. GARCIA-JIMENO (2015): "On Mis-measured Binary Regressors: New Results and Some Comments on the Literature," *Unpublished manuscript*.

FRANGAKIS, C. E., and D. B. RUBIN (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

FRICKE, H., M. FROLICH, M. HUBER, and M. LECHNER (2015): "Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identifcation of Causal Effects by Instruments," *Unpublished manuscript*.

GINTHER, D. K. (2000): "Alternative Estimates of the Effect of Schooling on Earnings," *The Review of Economics and Statistics*, 82(1), 103–116.

HECKMAN, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

HECKMAN, J. J., D. SCHMIERER, and S. URZUA (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158, 177–203.

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

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

HECKMAN, J. J., and E. VYTLACIL (2001): "Local Instrumental Variables," *in Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46.

HECKMAN, J. J., and E. VYTLACIL (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

HENRY, M., Y. KITAMURA, and B. SALANIÉ (2014): "Partial Identification of Finite Mixtures Econometric Models," *Quantitative Economics*, 5, 123–144.

HIRANO, K., G. W. IMBENS, D. B. RUBIN, and X.-H. ZHOU (2000): "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1(1), 69–88.

HOTZ, V. J., C. MULLIN, and S. SANDERS (1997): "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing.," *Review of Economic Studies*, 64(4), 575–603.

IMBENS, G. W. (2014): "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29(3), 323–358.

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

IMBENS, G. W., and D. B. RUBIN (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

JOCHMANS, K., M. HENRY, and B. SALANIÉ (2017): "Inference on Two Component Mixtures under Tail Restrictions," *Econometric Theory*, 33, 610–635.

KÉDAGNI, D., and I. MOURIFIÉ (2015): "Generalized Instrumental Inequalities: Testing the IV Independence Assumption," *Unpublished manuscript*.

KÉDAGNI, D., and I. MOURIFIÉ (2016): "Testing the IV Zero-Covariance Assumption and Identification with an Invalid Instrument," *Unpublished manuscript*.

KITAGAWA, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063.

KOLESÁR, M., R. CHETTY, J. FRIEDMAN, E. GLAESER, and G. W. IMBENS (2015): "Identification and Inference with Many Invalid Instruments," *Journal of Business and Economic Statistics*, 33(4), 474–484.

LEE, D. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76(3), 1071–1102.

LEWBEL, A. (2007): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 75(2), 537–551.

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

MANSKI, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

MANSKI, C. F., and J. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

MANSKI, C. F., and J. PEPPER (2009): "More on Monotone Instrumental Variables," *Econometrics Journal*, 12, S200–S216.

MOURIFIÉ, I., and Y. WAN (2017): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99(2), 305–313.

NEVO, A., and A. M. ROSEN (2012): "Identification with Imperfect Instruments," *The Review of Economics and Statistics*, 94(3), 659–671.

SHAIKH, A. M., and E. VYTLACIL (2005): "Threshold Crossing Models and Bounds on Treatment Effects: A Nonparametric Analysis," *National Bureau of Economic Research, Technical working paper*, 307.

SHAIKH, A. M., and E. VYTLACIL (2011): "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables," *Econometrica*, 79(3), 949–955.

SHI, X., and M. SHUM (2015): "Simple Two-Stage Inference for a Class of Partially Identified Models," *Econometric Theory*, 31, 493–520.

STOKEY, N. L., and J. R. E. LUCAS (1989): "Recursive Methods in Economic Dynamics, with E. C. Prescott," *Havard University Press*.

URA, T. (2015): "Heterogenous Treatment Effects with Mismeasured Endogenous Treatment," *Unpublished manuscript*.