Policy Evaluation with Multiple Instrumental Variables*

Magne Mogstad† Alexander Torgovitsky‡ Christopher R. Walters§

August 26, 2021

Abstract

Marginal treatment effect methods are widely used for causal inference and policy evaluation with instrumental variables. However, they fundamentally rely on the well-known monotonicity (threshold-crossing) condition on treatment choice behavior. This condition cannot hold with multiple instruments unless treatment choice is effectively homogeneous. We develop a new marginal treatment effect framework under a weaker, partial monotonicity condition. The partial monotonicity condition is implied by standard choice theory and allows for rich unobserved heterogeneity even in the presence of multiple instruments. The new framework can be viewed as having multiple different choice models for the same observed treatment variable, all of which must be consistent with the data and with each other. Using this framework, we develop a methodology for partial identification of clearly stated, policy-relevant target parameters while allowing for a wide variety of nonparametric shape restrictions and parametric functional form assumptions. We show how the methodology can be used to combine multiple instruments together to yield more informative empirical conclusions than one would obtain by using each instrument separately. The methodology provides a blueprint for extracting and aggregating information from multiple controlled or natural experiments while still allowing for rich unobserved heterogeneity in both treatment effects and choice behavior.

JEL classification: C01; C14; C21; C26; C51
Keywords: marginal treatment effects, instrumental variables, partial identification, local average treatment effect, heterogeneous treatment effects, policy evaluation

*We thank Deniz Dutz, Jim Heckman, Vishal Kamat, Ed Vytlacil, and three anonymous referees for helpful comments. Omkar Katta provided outstanding research assistance.

†Department of Economics, University of Chicago; Statistics Norway; NBER.

‡Department of Economics, University of Chicago. Research supported in part by National Science Foundation grant SES-1846832.

§Department of Economics, University of California, Berkeley; NBER.

1 Introduction

Heckman and Vytlacil (1999) introduced the marginal treatment effect (MTE) as a unifying concept for program and policy evaluation.[1] Since then, MTE methods have become a fundamental tool for empirical work, and have been applied in a variety of different settings including the returns to schooling (Moffitt, 2008; Carneiro, Heckman, and Vytlacil, 2011; Carneiro, Lokshin, and Umapathi, 2016; Nybom, 2017) and its impacts on wage inequality (Carneiro and Lee, 2009), discrimination (Arnold, Dobbie, and Yang, 2018; Arnold, Dobbie, and Hull, 2020), the effects of foster care (Doyle Jr., 2007), the impacts of welfare (Moffitt, 2019) and disability insurance (Maestas, Mullen, and Strand, 2013; French and Song, 2014; Autor, Kostøl, Mogstad, and Setzler, 2019) programs on labor supply, the performance of charter schools (Walters, 2018), health care (Kowalski, 2018; Depalo, 2020), the effects of early childhood programs (Kline and Walters, 2016; Cornelissen, Dustmann, Raute, and Schönberg, 2018; Felfe and Lalive, 2018), the efficacy of preventative health products (Mogstad, Santos, and Torgovitsky, 2017), the quantity–quality theory of fertility (Brinch, Mogstad, and Wiswall, 2017), and the effects of incarceration (Bhuller, Dahl, Løken, and Mogstad, 2020), among many others. Mogstad and Torgovitsky (2018) provide a recent review of the MTE methodology and its connection to other instrumental variable (IV) approaches.

A key assumption underlying the MTE methodology is that shifts in the instrument have a uniform effect on the treatment choices of all individuals. This is the well-known "monotonicity" condition introduced by Imbens and Angrist (1994), which Vytlacil (2002) showed is equivalent to the type of separable threshold-crossing selection equation that had been extensively used in prior econometric work (e.g. Heckman, 1974, 1976, 1979). Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) observed that the content of the monotonicity condition makes it more appropriately described as a *uniformity* condition, since it restricts unobserved heterogeneity in how treatment choice can respond to the instruments.

Building on their intuition, we showed elsewhere (Mogstad, Torgovitsky, and Walters, 2020) that the Imbens-Angrist monotonicity (IAM) condition cannot hold when there are multiple distinct instruments unless there is no unobserved heterogeneity in treatment choice behavior.[2] Yet, as we also documented in that paper, empirical re-

[1]Heckman and Vytlacil (1999) initially referred to the MTE as the *local instrumental variable* (LIV) before later drawing a distinction between the MTE, as an unobservable parameter, and the LIV as an estimand that can potentially identify the MTE (Heckman and Vytlacil, 2001c,a). The ideas behind the MTE also appear in earlier work by Björklund and Moffitt (1987), Heckman (1997), and Heckman and Smith (1998).

[2]This amplifies previous observations, such as by Vytlacil (2002, p. 355) and Heckman et al. (2006, p. 399), who both note that IAM can fail in a random coefficients specification.

searchers frequently combine multiple distinct instruments together using the two-stage least squares estimator, presumably motivated by the efficiency gains that arise in the classical linear IV model under homogeneous treatment effects. The two-stage least squares estimator does not necessarily answer a well–posed counterfactual even when IAM is satisfied. When it is not satisfied, as with multiple instruments, it can even give negative weight to some complier groups (Mogstad et al., 2020). MTE methods can be used to conduct inference on specific parameters that answer clear counterfactuals, but again, the MTE methodology is premised on IAM, which generically does not hold with multiple instruments.

In this paper, we provide a solution to this problem by developing the MTE methodology under a strictly weaker version of the IAM condition called partial monotonicity (PM). The PM condition with multiple instruments is that IAM is satisfied for each instrument separately, holding all of the other instruments fixed. The condition is satisfied if each instrument by itself makes every individual weakly more likely to choose treatment. While PM restricts the sign of the effects of each instrument on treatment choices, it still allows for rich unobserved heterogeneity in the relative magnitudes of these effects, unlike IAM (Mogstad et al., 2020).

We show that PM with multiple instruments gives rise to multiple threshold-crossing selection equations, one for each instrument. Each selection equation is separable in a single scalar unobservable that is derived from the marginal potential choices induced by a single instrument. This unobservable is independent of the instrument from which it was derived after conditioning on all of the other instruments as control variables. This sets up a complex and unique structure that can be viewed as having multiple different models for the same observed treatment variable, all of which must be consistent with the data and with each other if the model is correctly specified.

We exploit this structure by expanding the framework of Mogstad, Santos, and Torgovitsky (2018). The flexibility of that framework allows us to incorporate multiple different selection equations by including additional constraints that ensure they are all consistent with one another. We call these constraints mutual consistency, since they enforce the requirement that the multiple models of treatment choice do not contradict each other on the collection of unobserved instrument–invariant parameters for which they all generate values. As we show through analytic and numerical examples, the mutual consistency condition enables information from different instruments to be aggregated even while allowing for rich unobserved heterogeneity in both choices and treatment effects. This provides a blueprint for thinking about how to combine exogenous variation from multiple controlled or natural experiments.

Other authors have considered modified MTE frameworks for settings in which the IAM condition is unattractive. Carneiro, Hansen, and Heckman (2003), Heckman et al. (2006), Heckman and Vytlacil (2007b) and Cunha, Heckman, and Navarro (2007) considered multivalued ordered treatments. Heckman et al. (2006), Heckman, Urzua, and Vytlacil (2008), Kline and Walters (2016), Heckman and Pinto (2018), Lee and Salanié (2018), Mountjoy (2019), and Pinto (2019) analyzed settings with a discrete, unordered treatment. Lee and Salanié (2018) also considered a double hurdle model for a binary treatment, which ends up being somewhat related to the multiple selection equations that arise under PM. Gautier and Hoderlein (2015) and Gautier (2020) consider threshold-crossing selection equations with a random coefficient structure.

The organization of the paper is as follows. In Section 2 we introduce the model, discuss the differences between IAM and PM, and show how PM leads to multiple selection equations. In Section 3, we develop the MTE methodology under PM with a particular emphasis on the concept of mutual consistency that arises as a consequence of the multiple selection equations. In Section 4, we provide analytic and numerical examples to provide further intuition as to how mutual consistency works to aggregate information across different instruments. Section 5 contains some brief concluding remarks.

2 Model

2.1 Potential Outcomes and Treatments

For each individual we observe an outcome Y, their binary treatment status, $D \in \{0,1\}$, an L–vector of instruments, Z, with support $\mathcal{Z} \subseteq \mathbb{R}^L$, and a vector of covariates, X. Conditioning on X allows for observed heterogeneity, whereas our focus here is on unobserved heterogeneity. To keep the notation more concise, we will suppress X throughout the paper, but all assumptions and analysis can be understood to hold conditional–on–X.

For each $d \in \{0,1\}$ and $z \in \mathcal{Z}$, let $Y(d,z)$ denote an individual's latent potential outcome that they would have realized had their treatment and instrument been externally set to d and z. Similarly, for each $z \in \mathcal{Z}$, let $D(z)$ denote their latent potential treatment choice if the instrument were z. The observed and potential variables are related through

$$Y = \sum_{d \in \{0,1\}} \sum_{z \in \mathcal{Z}} \mathbb{1}[D = d, Z = z] Y(d,z) \quad \text{and} \quad D = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z = z] D(z), \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function that is 1 if $\cdot$ is true and 0 otherwise.

We maintain the following standard conditions throughout the paper:

Assumptions E.

E.1 $Y(d,z) = Y(d,z') \equiv Y(d)$ *for all* $d \in \{0,1\}$ *and* $z, z' \in \mathcal{Z}$.

E.2 $\mathbb{E}\left[Y(d)|Z, \{D(z)\}_{z\in\mathcal{Z}}\right] = \mathbb{E}\left[Y(d)|\{D(z)\}_{z\in\mathcal{Z}}\right]$ *and* $\mathbb{E}[Y(d)^2] < \infty$ *for* $d \in \{0,1\}$.

E.3 $\{D(z)\}_{z\in\mathcal{Z}} \perp\!\!\!\perp Z$, *where* $\perp\!\!\!\perp$ *denotes statistical independence.*

Assumption E.1 is the traditional exclusion restriction that the instruments have no direct causal effect on outcomes. Given this assumption, we can write the first part of (1) more simply as

$$Y = DY(1) + (1-D)Y(0). \tag{2}$$

Assumptions E.2 and E.3 are exogeneity conditions on the instruments. These are usually stated together as one stronger condition: $(Y(0), Y(1), \{D(z)\}_{z\in\mathcal{Z}} \perp\!\!\!\perp Z$, see e.g. Imbens and Angrist (1994, Condition 1) or Vytlacil (2002, L-1(i)). We have weakened full independence to mean independence because our methodological framework will focus on quantities that can be expressed as a mean effect of D on Y. This is common in the MTE literature, see e.g. Heckman and Vytlacil (2007b) or Mogstad et al. (2018).[3] The moment condition in Assumption E.2 merely serves to ensure the existence of the relevant conditional means.

2.2 The Imbens-Angrist Monotonicity Condition

Imbens and Angrist (1994) introduced the following assumption about the potential treatment states. They described it as "monotonicity."

Assumption IAM. (Imbens and Angrist Monotonicity) *For all* $z, z' \in \mathcal{Z}$ *either* $D(z) \geq D(z')$ *almost surely, or else* $D(z) \leq D(z')$ *almost surely.*

Assumption IAM requires a shift from one instrument value z to another value z' to either act as an incentive to take treatment for *all* individuals, or as a disincentive for *all* individuals. It does not allow some individuals to respond positively and others negatively. There is no presumption in Assumption IAM that Z is a scalar as opposed to a vector, or indeed, some more exotic random object. The possibility that Z is a vector was explicitly entertained by Imbens and Angrist (1994, pg. 470).

[3] Marx (2020) considers the additional information in the stronger, full independence assumption.

Vytlacil (2002) showed that Assumptions E.2, E.3, and IAM together were equivalent to assuming that $D(z)$ obeys a threshold-crossing model

$$D(z) = \mathbb{1}[V \leq \eta(z)], \tag{3}$$

for some unknown function η and some continuously distributed unobservable V such that $\mathbb{E}[Y(d)|Z, V] = \mathbb{E}[Y(d)|V]$ and $V \perp\!\!\!\perp Z$. Intuitively, the potential choices $\{D(z)\}_{z\in\mathcal{Z}}$ can be viewed as discretizations of some underlying latent proneness to take treatment, V. Individuals with smaller values of V are more likely to take treatment, while those with larger values of V are less likely. As Vytlacil (2002) discusses, this one-dimensional ordering by V is a different—but equivalent—way of viewing Assumption IAM. The primary benefit is in providing a tidy, unidimensional measure of unobserved heterogeneity.

It is important to emphasize, however, that the interpretation of V is inextricable from the definition of the instrument, Z. Indeed, Vytlacil's (2002) equivalence argument constructs V directly from the potential choices, $\{D(z)\}_{z\in\mathcal{Z}}$. Different instruments thus yield different unobservables, V. An individual who responds to one instrument may not respond to another. If so, their V for one instrument would be different than their V for another instrument. This distinction becomes especially salient when Z is a vector comprised of multiple different economic variables.

Selection equations like (3) have long been used in econometrics, typically with additional parametric assumptions on η and/or the distribution of V (e.g. Heckman, 1974, 1976; Heckman, Tobias, and Vytlacil, 2001). Vytlacil's (2002) result shows that these traditional econometric models can be viewed as parameterized special cases of the potential outcomes model with Assumptions E and IAM. This observation forms the cornerstone of the MTE literature. It implies that the potential outcomes model under Assumptions E and IAM—the standard model of many authors since Imbens and Angrist (1994)—can be simply viewed as a nonparametric descendent of a lineage of fully parametric selection models. Choosing to analyze or implement these models nonparametrically or parametrically is a research decision that comes with an attendant trade-off between the strength of the assumptions and the strength of the conclusions. These two poles and a broad range of research decisions in between can be unified using the MTE (Mogstad and Torgovitsky, 2018).

2.3 Implications of IAM with Multiple Instruments

In Mogstad et al. (2020), we showed that Assumption IAM/(3) is an extremely strong condition if Z is a vector comprising multiple distinct economic variables. It effectively rules out any meaningful unobserved heterogeneity in treatment choice behavior. Our results amplify the observation by Heckman and Vytlacil (2005, pp. 715-716) that Assumption IAM requires uniformity across individuals, not monotonicity in the instrument. The assumption that all individuals respond in a uniform direction can be reasonable if the instrument is something like a price. However, it is a strong assumption if the instrument consists of multiple types of incentives or disincentives.

For example, suppose that $D(z) \in \{0, 1\}$ is the decision to attend college, and that $z = (z_1, z_2)$, where $z_1 \in \{0, 1\}$ is a tuition subsidy and $z_2 \in \{0, 1\}$ is proximity to college. These two instruments have been widely used in the empirical literature (e.g. Kane and Rouse, 1993; Card, 1995). If Assumption IAM is satisfied, then it is not possible that some individuals respond to tuition subsidies but not to distance, while other individuals respond to distance but not to tuition subsidies. If this were the case, then we would have

$$\mathbb{P}[\underbrace{1 = D(1,0) > D(0,1) = 0}_{\text{respond to tuition, not distance}}] > 0 \quad \text{and} \quad \mathbb{P}[\underbrace{1 = D(0,1) > D(1,0) = 0}_{\text{respond to distance, not tuition}}] > 0, \tag{4}$$

which contradicts Assumption IAM, since it implies both

$$\mathbb{P}[D(0,1) \geq D(1,0)] < 1 \quad \text{and} \quad \mathbb{P}[D(1,0) \geq D(0,1)] < 1.$$

Alternatively, suppose that we start with a threshold-crossing model such as

$$D(z) = \mathbb{1}[B_0 + B_1 z_1 + z_2 \geq 0], \tag{5}$$

where B_0 and B_1 are both unobservable random variables with $(B_0, B_1) \perp\!\!\!\perp Z$. If we view the index in (5) as the net indirect utility from attending college, then B_1 can be interpreted as the marginal rate of substitution between tuition and proximity. This model cannot generally be re-written in the form of (3) with a single unobservable V unless $\text{Var}(B_1) = 0$. That is, unless the marginal rate of substitution is homogeneous across individuals.

Proceeding anyway with the assumption that there is no unobserved heterogeneity in selection seems unpalatable. Unobserved heterogeneity is routinely found to be important in empirical work (Heckman, 2001), and it is an emphasis of the modern literature on causal inference (Imbens, 2014). Indeed, the very motivation for Assumption IAM was to make sense of linear IV estimators in the presence of unobserved treatment effect heterogeneity (Imbens and Angrist, 1994). Allowing for unrestricted unobserved heterogeneity in outcomes but assuming away all unobserved heterogeneity in treatment choice behavior would be an even more extreme form of what Heckman and Vytlacil (2005, pg. 671) called a "fundamental asymmetry" in IV models that maintain Assumption IAM.

2.4 Partial Monotonicity

To allow for unobserved heterogeneity in treatment choices, we replace Assumption IAM with a weaker assumption called *partial monotonicity* (Mogstad et al., 2020).[4] To state the condition, we divide vectors $z \in \mathcal{Z} \subseteq \mathbb{R}^L$ into their ℓth component, z_ℓ, and all other $(L-1)$ components, $z_{-\ell}$. We write $z = (z_\ell, z_{-\ell})$ to emphasize the separation of the ℓth component.

Assumption PM. (Partial Monotonicity) *Take any $\ell = 1, \ldots, L$, and let $(z_\ell, z_{-\ell})$ and $(z'_\ell, z_{-\ell})$ be any two points in $\mathcal{Z}$. Then either $D(z_\ell, z_{-\ell}) \geq D(z'_\ell, z_{-\ell})$ almost surely, or else $D(z_\ell, z_{-\ell}) \leq D(z'_\ell, z_{-\ell})$ almost surely.*

It is immediate that Assumption IAM implies Assumption PM, and that the two assumptions are equivalent when there is only a single instrument ($L = 1$). When $L > 1$, Assumption PM is strictly weaker than Assumption IAM; see Mogstad et al. (2020) for more detail. A simple sufficient condition for Assumption PM is that $D(z)$ is an increasing function of z with respect to the usual vector partial order on $\mathcal{Z}$ (Mogstad et al., 2020). Such a condition still allows for unobserved heterogeneity in the responses to these incentives, unlike Assumption IAM.

For example, with the two binary college attendance instruments, Assumption PM is satisfied if all individuals are more likely to attend college when it is closer and/or subsidized. That is, if

$$D(1, z_2) \geq D(0, z_2) \text{ for } z_2 = 0, 1 \quad \text{and} \quad D(z_1, 1) \geq D(z_1, 0) \text{ for } z_1 = 0, 1.$$

This does not involve a comparison between $D(0,1)$ and $D(1,0)$, and thus allows for (4), so that some individuals may respond to distance but not to subsidies, and vice-versa. In terms of the random coefficient threshold-crossing model (5), Assumption PM allows for $\text{Var}(B_1) > 0$, as long as B_1 is non-negative (or non-positive) with probability 1. Using the net utility interpretation of this equation, Assumption PM allows for unobserved heterogeneity in the magnitude of the marginal rate of substitution, just not in the sign.

[4] Mountjoy (2019) used a similar assumption in a setting with multiple unordered treatments.

2.5 Selection Equations under Partial Monotonicity

Assumption PM can be used to derive selection equations similar to (3). Consider first the marginal potential treatments, defined for each ℓ as

$$D_\ell(z_\ell) \equiv \sum_{z_{-\ell}\in\mathcal{Z}_{-\ell}} \mathbb{1}[Z_{-\ell} = z_{-\ell}]D(z_\ell, z_{-\ell}) \equiv D(z_\ell, Z_{-\ell}), \tag{6}$$

where $\mathcal{Z}_\ell$ denotes the support of $Z_\ell \in \mathbb{R}$, and $\mathcal{Z}_{-\ell}$ denotes the support of $Z_{-\ell} \in \mathbb{R}^{L-1}$. For example, if there are two binary instruments, so that $\mathcal{Z} = \{0,1\}^2$, then there are two sets of marginal potential treatments, and (6) can be written for $\ell = 1, 2$ as

$$\begin{aligned} & D_1(z_1) = Z_2 D(z_1, 1) + (1 - Z_2)D(z_1, 0) \\ \text{and} \qquad & D_2(z_2) = Z_1 D(1, z_2) + (1 - Z_1)D(0, z_2). \end{aligned} \tag{7}$$

That is, $D_1(z_1)$ is the treatment choice an individual would have made had Z_1 been set to z_1 while Z_2 remained at its observed realization, whereas $D_2(z_2)$ is the treatment choice they would have made if Z_2 were set to z_2 with Z_1 unchanged.

Conditional on $Z_{-\ell}$, each marginal potential treatment is equal to a single joint potential treatment:

$$\mathbb{P}[D_\ell(z_\ell) = D(z_\ell, z_{-\ell})|Z_{-\ell} = z_{-\ell}] = 1. \tag{8}$$

As a consequence, Assumption PM implies that each collection of marginal potential treatments $\{D_\ell(z_\ell)\}_{z_\ell\in\mathcal{Z}_\ell}$ satisfies Assumption IAM *conditional on* any realization of $Z_{-\ell}$, since

$$\mathbb{P}[D_\ell(z_\ell) \geq D_\ell(z'_\ell)|Z_{-\ell} = z_{-\ell}] = \mathbb{P}[D(z_\ell, z_{-\ell}) \geq D(z'_\ell, z_{-\ell})|Z_{-\ell} = z_{-\ell}] \in \{0, 1\} \tag{9}$$

for any $z_\ell, z'_\ell \in \mathcal{Z}_\ell$ and any $z_{-\ell} \in \mathcal{Z}_{-\ell}$. With two binary instruments, this means that either

$$\mathbb{P}[D_1(1) \geq D_1(0)|Z_2 = z_2] = 1 \quad \text{or} \quad \mathbb{P}[D_1(0) \geq D_1(1)|Z_2 = z_2] = 1$$

for both $z_2 \in \{0, 1\}$, as well as the analogous condition with the roles of the two instruments flipped.

Notice that Assumption E.3 *does not* imply that $\{D_\ell(z_\ell)\}_{z_\ell\in\mathcal{Z}_\ell} \perp\!\!\!\perp Z$. That is, the marginal potential treatments *are not* independent of the entire instrument vector. This is because $D_\ell(z_\ell)$ directly depends on $Z_{-\ell}$, which is itself a subvector of Z; see (6) and the special case (7). However, Assumption E.3 does imply the weaker condition

$$\{D_\ell(z_\ell)\}_{z_\ell \in \mathcal{Z}_\ell} \perp\!\!\!\perp Z_\ell | Z_{-\ell} \quad \text{for every } \ell = 1, \ldots, L, \tag{10}$$

so that each set of marginal potential treatments is independent of the single instrument from which it was derived, *conditional* on the other instruments. Using only a subset of the instruments without conditioning on the others will generally violate Assumption E.3 whether or not IAM is satisfied (see Section 3.6 of Heckman, 2010 or Section 4.5 of Mogstad et al., 2020).

Combining (9) with (10) means that Vytlacil's (2002) equivalence result can still be applied to construct L different threshold-crossing equations of the form (3). Specifically, the result implies that for *each* $\ell = 1, \ldots, L$,

$$D_\ell(z_\ell) = \mathbb{1}\left[V_\ell \leq \eta_\ell(z_\ell, Z_{-\ell})\right], \tag{11}$$

for some unknown function η_ℓ and some continuously distributed unobservable V_ℓ such that $V_\ell \perp\!\!\!\perp Z_\ell | Z_{-\ell}$ and

$$\mathbb{E}[Y(d)|Z, V_\ell] = \mathbb{E}[Y(d)|Z_{-\ell}, V_\ell] \quad \text{for } d = 0, 1. \tag{12}$$

This is the same conclusion as in (3), but now there is one selection equation for each component Z_ℓ of the L–dimensional vector of instruments, Z, and each selection equation is conditional on a set of "controls" consisting of all other instruments, $Z_{-\ell}$.

As in the usual analysis under Assumption IAM, we will normalize (8) so that the distribution of the unobservable is uniform. This is possible because the distribution function of V_ℓ conditional on $Z_{-\ell}$—call it F_ℓ—is strictly increasing on its support, so that (11) can be written as

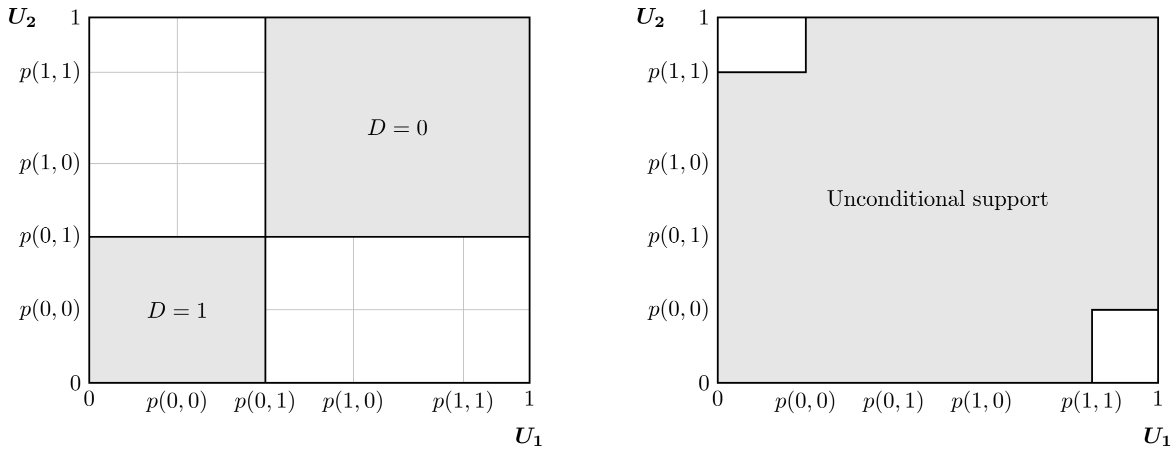
$$D_\ell(z_\ell) = \mathbb{1}\left[F_\ell(V_\ell|Z_{-\ell}) \leq F_\ell(\eta(z_\ell, Z_{-\ell})|Z_{-\ell})\right] = \mathbb{1}[U_\ell \leq p(z_\ell, Z_{-\ell})], \tag{13}$$

where $U_\ell \equiv F_\ell(V_\ell|Z_{-\ell})$ is uniformly distributed over $[0, 1]$ for each ℓ, conditional on any value of $Z_{-\ell}$. The second equality in (13) follows because

$$\begin{aligned} p(z_\ell, Z_{-\ell}) &= \mathbb{P}[D_\ell(z_\ell) = 1 | Z_\ell = z_\ell, Z_{-\ell}] \\ &= \mathbb{P}[V_\ell \leq \eta_\ell(z_\ell, Z_{-\ell})|Z_{-\ell}] = F(\eta(z_\ell, Z_{-\ell})|Z_{-\ell}), \end{aligned}$$

as a consequence of V_ℓ being independent of Z_ℓ, conditional on $Z_{-\ell}$.

As in the standard threshold-crossing model, (3), U_ℓ can be interpreted as a latent



(a) *Support of (U_1, U_2) given $Z = (0, 1)$.*

(b) *Unconditional support of (U_1, U_2).*

Figure 1: *The joint support of (U_1, U_2) when $L = 2$ and $\mathcal{Z} = \{0, 1\}^2$. While both U_1 and U_2 have uniform marginal distributions over $[0, 1]$ by construction, their joint support will be a proper subset of the unit square conditional on Z. If Z has finite support, as in this example, then the unconditional support of U will also be a proper subset of the unit square.*

proneness to take the treatment, with smaller values corresponding to higher proneness. However, now there is a different U_ℓ for each instrument, so that this proneness is measured against the incentive (or disincentive) created by the ℓth instrument. This reflects the point raised earlier that the interpretation of the latent variable V_ℓ (or U_ℓ) is *derived from* the instrument, and so cannot be interpreted in isolation from the instrument. It is therefore possible for individuals to have a high value of U_1 and a low value of U_2 or vice versa, since these variables measure proneness to take treatment along different preference dimensions.

However, these possibilities have limits. This is because each selection model provides a different representation of the same observed treatment status through (13). As a consequence, the components of $(U_1, \ldots, U_L)$ must be statistically dependent, even though their marginal distributions are uniform. That is, the distribution of $(U_1, \ldots, U_L)$—which is a copula given the normalizations—cannot be the product copula. Not only that, but $(U_1, \ldots, U_L)$ will also generally be dependent with the entire vector Z, since each U_ℓ is only independent of Z_ℓ given $Z_{-\ell}$, but is not generally independent of $Z_{-\ell}$, as observed in (6) and (7).

To visualize these properties, return to the case with two binary instruments and consider the joint distribution of (U_1, U_2) conditional on $Z = (z_1, z_2)$. In order for (13) to hold for both $\ell = 1$ and $\ell = 2$, one must have that

$$\mathbb{P}\left[\mathbb{1}\left[U_1 \le p(z_1, z_2)\right] = \mathbb{1}\left[U_2 \le p(z_1, z_2)\right] \Big| Z = (z_1, z_2)\right] = 1,$$

for any realization of (z_1, z_2). That is, either U_1 and U_2 are both smaller than $p(z_1, z_2)$, or else U_1 and U_2 are both larger than $p(z_1, z_2)$. This region of support is depicted in Figure 1a. Taking the union of this set across all four values of (z_1, z_2) gives the unconditional support of (U_1, U_2), which is depicted in Figure 1b. Two subsets of the unit square necessarily have zero mass: It is not possible to have $U_1 \le p(0, 0)$ and $U_2 > p(1, 1)$ together, nor is it possible to have $U_1 > p(1, 1)$ and $U_2 \le p(0, 0)$ together. The reason is that under (13), either pair of realizations would entail always choosing both $D = 1$ and $D = 0$ for any realization of Z.

Instead of deriving (13) from Assumption PM, one can also derive it directly from a *non*separable threshold-crossing equation with multiple unobservables. For example, suppose that $L = 2$ with potential outcomes determined by the random coefficients specification of indirect utility in (5). From (5), the two pre-normalized selection equations (11) can be derived as

$$D_1(z_1) = \mathbb{1}\Big[\overbrace{-\frac{(B_0 + Z_2)}{B_1}}^{\equiv V_1} \le \overbrace{z_1}^{\equiv \eta_1(z)}\Big] \quad \text{and} \quad D_2(z_2) = \mathbb{1}\Big[\overbrace{-(B_0 + B_1 Z_1)}^{\equiv V_2} \le \overbrace{z_2}^{\equiv \eta_2(z)}\Big].$$

Notice in particular that even though (B_0, B_1) is independent of (Z_1, Z_2), this will not be the case for (V_1, V_2). Instead, V_1 is dependent with Z_2, and in general only independent with Z_1 after conditioning on Z_2. Similarly, V_2 is dependent with Z_1 with independence between V_2 and Z_2 only guaranteed after conditioning on Z_1. In addition, V_1 and V_2 are clearly dependent, since they are both functions of B_0 and B_1.

If $\text{Var}(B_1) = 0$, so that $B_1 = b_1$ is constant, then we can write

$$D_1(z_1) = \mathbb{1}\left[-B_0 \le b_1 z_1 + Z_2\right] \quad \text{and} \quad D_2(z_2) = \mathbb{1}\left[-B_0 \le b_1 Z_1 + z_2\right],$$

so that both equations could be rationalized by a single threshold-crossing equation with a single unobservable,

$$D(z_1, z_2) = \mathbb{1}\Big[\overbrace{-B_0}^{\equiv V} \le \overbrace{b_1 z_1 + z_2}^{\equiv \eta(z_1, z_2)}\Big].$$

That is, (5) could be written in form (3), and Assumption IAM would be satisfied. Without the assumption that $\text{Var}(B_1) = 0$—that is, homogeneity in the marginal rate of substitution—such a reformulation is not generally possible.

3 Methodology

In this section we develop the MTE methodology under Assumption PM. We begin in Section 3.1 by defining different instrument–specific MTEs as the fundamental unit of analysis in the model. In Section 3.2, we describe the class of target parameters that we focus on. In Section 3.3, we review the identification analysis developed by (Mogstad et al., 2018, "MST" hereafter), which showed how to flexibly move between point and partial identification under Assumption IAM. In Section 3.4, we then adapt this methodology to Assumption PM by using the concept of mutual consistency, which links together the various instrument–specific MTEs into a greater whole. In Section 3.5 we discuss some implications for testing whether Assumptions IAM and PM hold.

3.1 Marginal Treatment Response Functions

For each instrument, Z_ℓ, and its unobservable, U_ℓ, we define the marginal treatment response (MTR) function

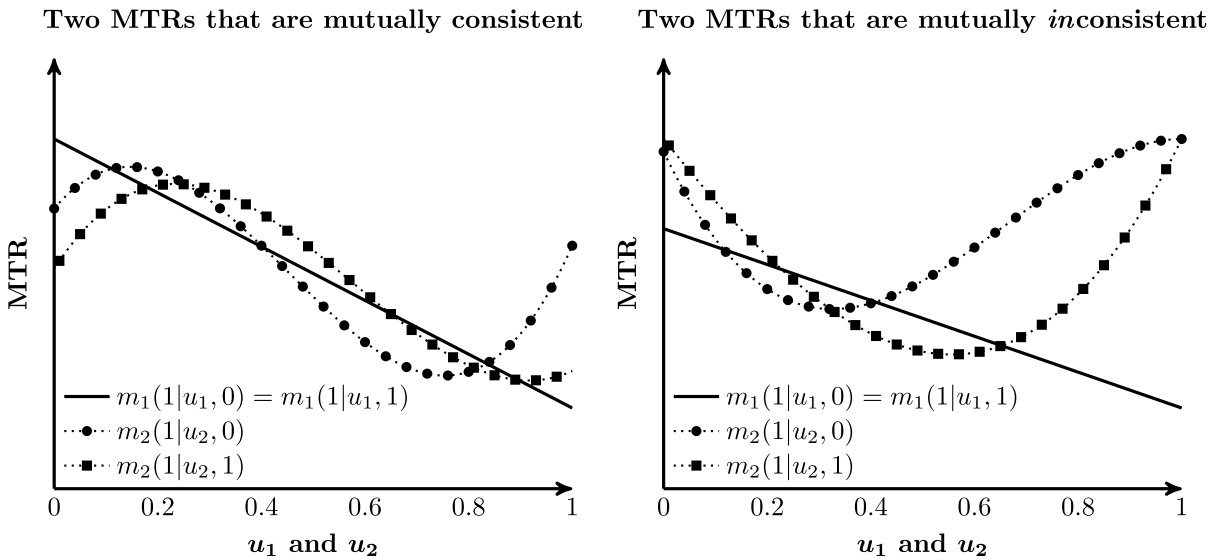
$$m_\ell(d|u_\ell, z_{-\ell}) \equiv \mathbb{E}\left[Y(d)|U_\ell = u_\ell, Z_{-\ell} = z_{-\ell}\right] \quad \text{for } d = 0, 1. \tag{14}$$

The MTR function describes variation in potential outcomes as a function of the propensity to take treatment along the ℓth margin, U_ℓ, again conditioning on all other instruments, $Z_{-\ell} = z_{-\ell}$. Each MTR function, m_ℓ, generates a marginal treatment effect (MTE) function (Heckman and Vytlacil, 1999, 2001c, 2005, 2007a,b) formed as $m_\ell(1|u_\ell, z_{-\ell}) - m_\ell(0|u_\ell, z_{-\ell})$. We let $m \equiv (m_1, \ldots, m_L)$, and assume that m belongs to a known parameter space $\mathcal{M} \subseteq \mathcal{M}_1 \times \cdots \times \mathcal{M}_L$ that encodes prior information (assumptions) that the researcher wants to impose about the MTR pairs.[5]

Each MTR and its corresponding MTE is defined in terms of a different margin of selection, U_ℓ, which is itself defined by the ℓth instrument component. Since the MTRs are instrument-specific, they are not directly comparable. However, a key point for our discussion ahead is that each MTR still describes the entire population, just organized along a different dimension of choice behavior. Thus, while the MTRs for different ℓ will typically be different, they cannot be arbitrarily different.

For example, consider again the case with two binary instruments. Figure 2a plots $m_1(1|\cdot, z_2)$ assuming (for simplicity) that this function does not vary with z_2, and so can be represented by the single solid line. It also plots $m_2(1|\cdot, z_1)$ for both $z_1 = 0$ and $z_1 =$

[5] We assume throughout that each $\mathcal{M}_\ell$ is contained in a vector space.



(a) *Both m_1 and m_2 imply the same value of $\mathbb{E}[Y(1)]$. These MTR pairs are mutually consistent.*

(b) *The value of $\mathbb{E}[Y(1)]$ implied by m_1 is different (smaller) from that implied by m_2. These MTR pairs are mutually inconsistent.*

Figure 2: *MTRs along different margins of selection (different U_ℓ) are not directly comparable. Nevertheless, they are not completely unrelated, since both MTRs provide a description of the entire population.*

1. While m_1 is not directly comparable to m_2, both functions are a conditional mean for the same random variable, $Y(1)$. To be *mutually consistent* then, one requirement is that

$$\mathbb{E}\left[m_1(1|U_1, Z_2)\right] = \mathbb{E}[Y(1)] = \mathbb{E}\left[m_2(1|U_2, Z_1)\right], \tag{15}$$

so that both m_1 and m_2 generate the same mean for $Y(1)$. In Figure 2a this is the case, since the integrals of both dotted curves are the same as the integral of the solid line.

In contrast, m_2 in Figure 2b is not mutually consistent with m_1. The areas under $m_2(1|\cdot, 0)$ and $m_2(1|\cdot, 1)$ are clearly greater than the area under $m_1(1|\cdot, 0) = m_1(1|\cdot, 1)$. It cannot be the case that both m_1 and m_2 describe the conditional mean of $Y(1)$, since these two MTRs would imply different values of $\mathbb{E}[Y(1)]$ through (15). In the following, we will develop a method that *requires* mutual consistency, so that pairs like those in Figure 2b are excluded from consideration. As we show later, excluding such pairs allows information gained about one instrument's MTR be used to restrict the MTR of another instrument.

3.2 The Target Parameter

We assume that the researcher has a well-posed empirical or policy question that can be informed by a specific target parameter. This could be a traditional target parameter, such as an average treatment effect for the population, or for the treated or untreated subpopulations. It could also be a policy-relevant treatment effect (Heckman and Vytlacil, 1999; Carneiro, Heckman, and Vytlacil, 2010) that captures the effect of changing choice sets. The focus on constructing an estimator for a specific target parameter answering a specific counterfactual distinguishes the MTE literature from work that attempts to reverse engineer interpretations for commonly-used estimators, such as two-stage least squares (e.g. Angrist and Pischke, 2009; Mogstad et al., 2020).

We require the target parameter, $\beta^\star$, to be a linear function of the L MTR pairs, having the form

$$\beta^\star(m) = \sum_{\ell=1}^{L} \beta^\star_\ell(m_\ell) \equiv \sum_{\ell=1}^{L} \sum_{d\in\{0,1\}} \mathbb{E}\left[\int_0^1 m_\ell(d|u_\ell, Z_{-\ell})\omega^\star_\ell(d|u_\ell, Z_\ell, Z_{-\ell})\, du_\ell\right], \quad (16)$$

where $\omega^\star_\ell$ are weighting functions specified by the researcher. The weighting functions are assumed to be known given knowledge of the joint distribution of (Y, D, Z). Heckman and Vytlacil (2005, 2007b), MST, and Mogstad and Torgovitsky (2018) provided catalogues of weighting functions for common target parameters. When $L = 1$, (16) reduces to the form used for the target parameter by MST.

When $L > 1$, there might be several ways to express the same target parameter. For example, if $\beta^\star$ is the population average treatment effect (ATE), $\mathbb{E}[Y(1) - Y(0)]$, then one could take $\omega^\star_\ell(1|u_\ell, z) = 1$ and $\omega^\star_\ell(0|u_\ell, z) = -1$ for any ℓ, while setting all other weight functions to 0. This is another manifestation of the mutual consistency issue illustrated in Figure 2. When using multiple instruments, we will *impose* mutual consistency, so that the implied value of the ATE is the same for any ℓ. Thus, as a practical matter, any choice of ℓ will yield the same inference on an instrument-invariant parameter such as the ATE, the average effect of the treatment on the treated (ATT), or the average effect of the treatment on the untreated (ATU).

Other interesting target parameters might be instrument-specific. For example, the class of policy-relevant treatment effects (PRTEs) introduced by Heckman and Vytlacil (2001a, 2005) includes parameters that measure the impact of changing the incentive associated with a given instrument. A special case of a PRTE is an extrapolated local average treatment effect (LATE), such as

$$\text{LATE}_1(+\delta\%) \equiv \mathbb{E}\left[Y(1) - Y(0) \,\middle|\, p(0, Z_2) < U_1 \leq \left(1 + \frac{\delta}{100}\right) \times p(1, Z_2)\right], \tag{17}$$

which is the LATE that would result if the Z_1 instrument were changed sufficiently to cause a $\delta\%$ increase in participation under $Z_1 = 1$. This target parameter can be used to gauge the sensitivity of point identified IV estimates to the definition of the complier group. See Heckman and Vytlacil (2005), Carneiro et al. (2010), and MST for further discussion and additional examples of PRTEs.

When the definition of the target parameter depends on the instrument, as in (17), the weights will also need to depend on the instrument, and the mutual consistency issue will not immediately arise. Nevertheless, there will still be benefits to requiring instrument-invariant parameters to be mutually consistent across different MTRs. As we demonstrate ahead, this requirement will allow information to flow between different instruments, so that their exogenous variation can be aggregated. The surprising implication is that even if the target parameter is instrument-specific, inference on that target parameter can still benefit from combining multiple instruments.

3.3 Using Each Instrument Separately

In this section we briefly review the MST methodology for inference on $\beta^\star$ under Assumption IAM. In the multiple instrument setting this can be equivalently viewed as using one instrument at a time, conditioning on the rest as covariates. In the next section we then augment the methodology to combine multiple instruments together using the concept of mutual consistency.

Suppose that $(Y(0), Y(1), D)$ were generated by (13) for any ℓ, with MTR function m_ℓ. Then Proposition 1 of MST shows that for any (measurable) known or identified function s,

$$\mathbb{E}[s(D, Z)Y] = \sum_{d \in \{0,1\}} \mathbb{E}\left[\int_0^1 m_\ell(d|u_\ell, Z_{-\ell})\omega^s(d|u_\ell, Z)\, du_\ell\right] \tag{18}$$

$$\text{where} \quad \omega^s(0|u, Z) \equiv s(0, Z)\mathbb{1}[u > p(Z)] \quad \text{and} \quad \omega^s(1|u, Z) \equiv s(1, Z)\mathbb{1}[u \leq p(Z)].$$

MST refer to a choice of s as an IV–like specification, and show that by choosing s appropriately, one can reproduce any linear IV estimand on the left-hand side of (18). Given a collection $\mathcal{S}$ of IV–like specifications, we say that an MTR m_ℓ is consistent with the observed data under $\mathcal{S}$ if it satisfies (18) for every $s \in \mathcal{S}$. We denote the set of such pairs by

$$\mathcal{M}_\ell^{\text{obs}} \equiv \{m_\ell : m_\ell \text{ satisfies (18) for each } s \in \mathcal{S}\}.$$

The identified set for the ℓth MTR pair is defined as

$$\mathcal{M}_\ell^{\text{id}} \equiv \mathcal{M}_\ell \cap \mathcal{M}_\ell^{\text{obs}}.$$

That is, $\mathcal{M}_\ell^{\text{id}}$ is the collection of MTR pairs for the ℓth instrument that satisfy the researcher's prior assumptions ($m_\ell \in \mathcal{M}_\ell$) and are consistent with the observed data for the choice of IV–like estimands in $\mathcal{S}$ ($m_\ell \in \mathcal{M}_\ell^{\text{obs}}$). The identified set for the ℓth component of the target parameter in (16) is the projection of $\mathcal{M}_\ell^{\text{id}}$ under $\beta_\ell^\star$, or

$$\mathcal{B}_\ell^{\text{id}} \equiv \left\{\beta_\ell^\star(m_\ell) : m_\ell \in \mathcal{M}_\ell^{\text{id}}\right\}.$$

If $\mathcal{M}_\ell$ is a convex set, then $\mathcal{B}_\ell^{\text{id}}$ is an interval, $[\underline{\beta}_\ell^\star, \overline{\beta}_\ell^\star]$, with endpoints given by

$$\underline{\beta}_\ell^\star \equiv \inf_{m_\ell \in \mathcal{M}_\ell^{\text{id}}} \beta_\ell^\star(m_\ell) \qquad \text{and} \qquad \overline{\beta}_\ell^\star \equiv \sup_{m_\ell \in \mathcal{M}_\ell^{\text{id}}} \beta_\ell^\star(m_\ell),$$

see Proposition 2 in MST.

To compute the endpoints of this interval, MST assume that $\mathcal{M}_\ell$ has a linear–in–parameters form, so that each $m_\ell \in \mathcal{M}_\ell$ can be written as

$$m_\ell(d|u_\ell, z_{-\ell}) = \sum_{k=1}^{K_\ell} \theta_{\ell k} b_{\ell k}(d|u_\ell, z_{-\ell}) \quad \text{for some } \theta_\ell \equiv (\theta_{\ell 1}, \ldots, \theta_{\ell K_\ell}) \in \Theta_\ell \subseteq \mathbb{R}^{K_\ell}, \tag{19}$$

where $b_{\ell k}$ are known basis functions. If Θ_ℓ can be specified as a set of linear equalities and inequalities, then this assumption makes $\underline{\beta}_\ell^\star$ (and $\overline{\beta}_\ell^\star$) the optimal value of a finite–dimensional linear program with θ_ℓ as the variables of optimization:

$$\underline{\beta}_\ell^\star = \min_{\theta_\ell \in \Theta_\ell} \sum_{k=1}^{K_\ell} \theta_{\ell k} \Gamma_{\ell k}^\star \quad \text{s.t.} \quad \sum_{k=1}^{K_\ell} \theta_{\ell k} \Gamma_{\ell k}^s = \mathbb{E}[s(D,Z)Y] \quad \text{for all } s \in \mathcal{S}, \tag{20}$$

where

$$\Gamma_{\ell k}^\star \equiv \sum_{d \in \{0,1\}} \mathbb{E}\left[\int_0^1 b_{\ell k}(d|u_\ell, Z_{-\ell}) \omega_\ell^\star(d|u_\ell, Z_\ell, Z_{-\ell})\, du_\ell\right],$$

$$\text{and} \quad \Gamma_{\ell k}^s \equiv \sum_{d \in \{0,1\}} \mathbb{E}\left[\int_0^1 b_{\ell k}(d|u_\ell, Z_{-\ell}) \omega_\ell^s(d|u_\ell, Z_\ell, Z_{-\ell})\, du_\ell\right],$$

are both identified quantities that can be directly estimated from the observed data. MST, Mogstad and Torgovitsky (2018), and Shea and Torgovitsky (2019) discuss different ways of specifying Θ_ℓ to incorporate nonparametric or parametric specifications, with or without additional shape constraints.

The identified set $\mathcal{B}_\ell^{\text{id}}$ could be either a non-degenerate interval, or a point, depending on how restrictively the MTRs are specified and on the richness of the set of IV–like estimands, $\mathcal{S}$. MST propose an estimator that can be used in either case. Shea and Torgovitsky (2019) observe that a linear GMM estimator can be used in point identified cases with no shape constraints.

3.4 Combining Instruments through Mutual Consistency

Condition (18) can be applied for each ℓ to restrict each of the L MTR functions in isolation. We connect them by requiring mutual consistency in the unobservable quantities they imply. For example, in Figure 2 we noted that every m_ℓ implies a value for $\mathbb{E}[Y(1)]$ given by

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\int_0^1 m_\ell(1|u_\ell, Z_{-\ell})\, du\right]. \tag{21}$$

We will restrict attention to choices of m for which the right-hand side of (21) is invariant to $\ell = 1, \ldots, L$. This restricts our attention to MTRs like those in Figure 2a, while ruling out inconsistent pairs like those in Figure 2b. The result will be tighter inference on each $\beta_\ell^\star$, as well as on the overall target parameter, $\beta^\star$.

We formalize the property of mutual consistency in a similar fashion to the data consistency condition, (18).[6] A straightforward modification of Proposition 1 in MST shows that if $(Y(0), Y(1), D)$ were generated by (13) with MTR function m_ℓ, then

$$\begin{aligned} \mathbb{E}[s(D,Z)Y(d)] &= \mathbb{E}\left[\int_0^1 m_\ell(d|u_\ell, Z_{-\ell})\bar{\omega}^s(u_\ell, Z)du_\ell\right] \\ \text{where} \quad \bar{\omega}^s(u,Z) &\equiv \omega^s(0|u,Z) + \omega^s(1|u,Z). \end{aligned} \tag{22}$$

This equation is like (18) with the important difference that it is in terms of the latent potential outcomes, $Y(d)$. In contrast to (18), where the left-hand side quantity was a direct function of the observed data, in (22) the left-hand side is in general not point identified. Nevertheless, the left-hand side of (22) does not vary with ℓ, so the right-

[6] In a previous draft, we referred to "mutual consistency" as "logical consistency." Torgovitsky (2019, Section S6.2) proposed using the mutual/logical consistency idea to combine overlapping dynamic potential outcomes models of state dependence.

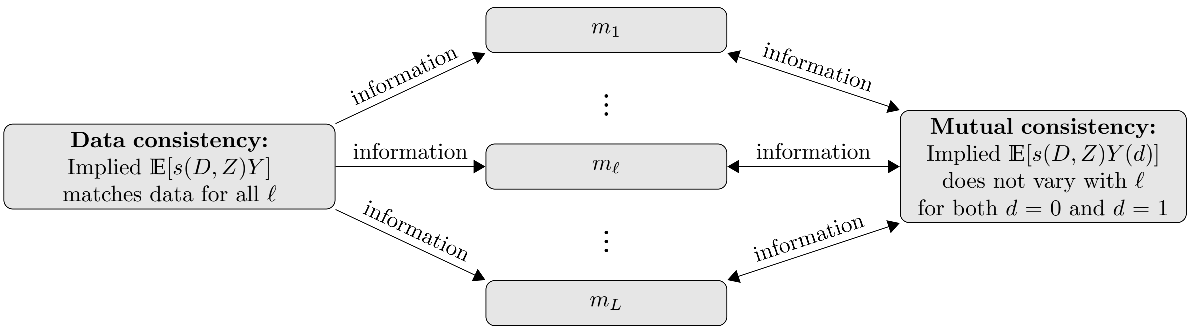


Figure 3: *The data consistency condition* (18) *constrains each m_ℓ to be consistent with the observed data in isolation. Mutual consistency* (23) *ties the m_ℓ together across $\ell = 2, \ldots, L$. This allows the information contained in different instruments to flow in the direction of the arrows, and therefore be combined across models that use different instruments.*

hand side should not either. Thus, we say that a collection of MTRs $m \equiv (m_1, \ldots, m_L)$ is mutually consistent under $\mathcal{S}$ if

$$\overbrace{\mathbb{E}\left[\int m_\ell(d|u_\ell, Z_{-\ell})\bar{\omega}^s(u_\ell, Z)\, du_\ell\right]}^{\mathbb{E}[s(D,Z)Y(d)] \text{ implied by } m_\ell} = \overbrace{\mathbb{E}\left[\int m_{\ell'}(d|u_{\ell'}, Z_{-\ell'})\bar{\omega}^s(u_{\ell'}, Z)\, du_{\ell'}\right]}^{\mathbb{E}[s(D,Z)Y(d)] \text{ implied by } m_{\ell'}}$$
$$\text{for } d = 0, 1, \text{ all } s \in \mathcal{S}, \text{ and all } \ell, \ell' \in \{1, \ldots, L\}. \tag{23}$$

Given a set of IV–like specifications, $\mathcal{S}$, the set of mutually consistent MTRs is

$$\mathcal{M}^{\text{mc}} \equiv \{m \equiv (m_1, \ldots, m_L) : m \text{ satisfies (23)}\}.$$

To combine multiple instruments together, we focus on the identified set

$$\mathcal{M}^{\text{id}} \equiv \mathcal{M} \cap \mathcal{M}^{\text{obs}} \cap \mathcal{M}^{\text{mc}},$$

where $\mathcal{M}^{\text{obs}} \equiv \mathcal{M}_1^{\text{obs}} \times \cdots \times \mathcal{M}_L^{\text{obs}}$. The identified set for the target parameter is then the projection of $\mathcal{M}^{\text{id}}$ under $\beta^\star$, or

$$\mathcal{B}^{\text{id}} \equiv \left\{\beta^\star(m) : m \in \mathcal{M}^{\text{id}}\right\}.$$

We conjecture that this identified set is sharp if $\mathcal{S}$ is a sufficiently rich class of functions (in the sense of Proposition 3 in MST), however we have been unable to prove it.

Figure 3 illustrates how the mutual consistency condition allows information to flow between different MTR functions. Intuitively, (18) places restrictions on m_ℓ for each ℓ by requiring it to match the observed data, whereas the mutual consistency condition propagates these restrictions from m_ℓ to $m_{\ell'}$. The result is a sort of equilibrium in which none of the MTR functions contradict each other on their implications for the instrument-invariant quantities $\mathbb{E}[s(D,Z)Y(d)]$ equated in (23). Limiting attention to the smaller set of MTRs that are consistent with this equilibrium mechanically shrinks the identified set for the target parameter as well.

The mutual consistency condition is a collection of linear equality constraints, so adding it does not fundamentally alter the theory or computational procedure in MST. In particular, if $\mathcal{M}$ is a convex set, then a minor change to Proposition 2 of MST shows that $\mathcal{B}^{\text{id}}$ is an interval, $[\underline{\beta}^\star, \overline{\beta}^\star]$, with endpoints given by

$$\underline{\beta}^\star \equiv \inf_{m \in \mathcal{M}^{\text{id}}} \beta^\star(m) \qquad \text{and} \qquad \overline{\beta}^\star \equiv \sup_{m \in \mathcal{M}^{\text{id}}} \beta^\star(m),$$

If the linear–in–parameters representation (19) is maintained for all $\ell = 1, \ldots, L$ with $\theta \equiv (\theta_1, \ldots, \theta_L) \in \Theta$, then $\underline{\beta}^\star$ can be found by modifying (20) to incorporate all MTRs as well as the mutual consistency constraint:

$$\begin{aligned} \underline{\beta}^\star_\ell = \min_{\theta \in \Theta} & \sum_{\ell=1}^{L} \sum_{k=1}^{K_\ell} \theta_{\ell k} \Gamma^\star_{\ell k} \\ \text{s.t. } & \sum_{k=1}^{K_\ell} \theta_{\ell k} \Gamma^s_{\ell k} = \mathbb{E}[s(D,Z)Y] \quad \text{for all } s \in \mathcal{S},\ \ell = 1, \ldots, L, \\ & \sum_{k=1}^{K_\ell} \theta_{\ell k} \bar{\Gamma}^s_{\ell k}(d) = \sum_{k=1}^{K_1} \theta_{1k} \bar{\Gamma}^s_{1k}(d) \quad \text{for all } s \in \mathcal{S},\ \ell = 2, \ldots, L, \text{ and } d = 0, 1, \end{aligned} \tag{24}$$

where for shorthand we have defined

$$\bar{\Gamma}^s_{\ell k}(d) \equiv \mathbb{E}\left[\int_0^1 b_{\ell k}(d|u_\ell, Z_{-\ell}) \bar{\omega}^s(u_\ell, Z) du_\ell\right].$$

If Θ consists only of linear equalities and inequalities, then (24) remains a finite-dimensional linear program in terms of quantities that are all point identified.

3.5 Testable Implications

Under Assumption IAM, the nonparametric IV model in Section 2 has testable implications (Balke and Pearl, 1997; Imbens and Rubin, 1997), and several authors have developed formal statistical tests of these implications in different forms (Huber and Mellace, 2014; Kitagawa, 2015; Mourifié and Wan, 2016). MST observe that these testable implications manifest themselves in the MTE methodology through the possibility that the identified set is empty. When using each instrument separately, as in Section 3.3, this would mean that $\mathcal{M}_\ell^{\text{id}}$ is empty, and would imply that either the ℓth instrument does not satisfy Assumption IAM, conditional on $Z_{-\ell}$, or that some aspect of Assumptions E are false. If the researcher specified $\mathcal{M}_\ell$ to include additional assumptions, then finding $\mathcal{M}_\ell^{\text{id}}$ to be empty could also call these into question.

This logic also holds when combining instruments, as in Section 3.4. Suppose that Assumptions E and the researcher's specification of $\mathcal{M}$ are beyond question. Then finding $\mathcal{M}^{\text{id}}$ empty implies that Assumption PM is violated. Suppose it is not empty and add the following assumption as part of the specification of $\mathcal{M}$:

$$\begin{aligned} &m_\ell(d|u, z_{-\ell}) = m_1(d|u, z_1) \equiv m_0(d|u) \\ &\qquad \text{for all } d = 0, 1,\ u \in [0, 1], \text{ all } l, \text{ and } z_{-\ell} \in \mathcal{Z}_{-\ell}. \end{aligned} \tag{25}$$

This assumption says that all of the L MTR functions are in fact the same, which among other things implies that they cannot depend on any component of z. When (25) is imposed, the mutual consistency condition (23) is immediately satisfied, and $\mathcal{M}^{\text{id}}$ reduces to the identified set obtained by using all instruments together under Assumption IAM. Thus, finding an empty identified set when (25) is imposed, but not when it is not, is evidence that Assumption IAM can be rejected, but that the strictly weaker Assumption PM cannot be rejected.

4 Aggregating Multiple Instruments

In this section, we demonstrate how the mutual consistency condition allows multiple instruments to be aggregated for more informative inference. In Section 4.1, we provide an algebraic example that shows how a model that would normally be just–identified becomes over–identified when mutual consistency is imposed. In Section 4.2, we show that the same principles are at work even without point identification. In Section 4.3, we conduct a numerical simulation that shows how mutual consistency interacts with the choice of target parameter and the auxiliary identifying assumptions maintained by the researcher. In Section 4.4, we discuss the implications of using IAM instead of PM when IAM does not hold, and provide an illustration of the bias this can cause when estimating a PRTE.

4.1 An Illustrative Example with Point Identification

The following example demonstrates how the mutual consistency condition yields additional over–identifying information that can be used to relax assumptions or to power a specification test.

Consider a setting with two binary instruments, so that $\mathcal{Z} = \{0,1\}^2$. As discussed, these two instruments give rise to two selection equations like (13) with two unobservables U_1 and U_2, and therefore two marginal treatment response functions, m_1 and m_2. To simplify the example, we will focus solely on the MTR function evaluated at the treated state, $d = 1$, so that our objects of concern are $m_1(1|u_1, z_2)$ and $m_2(1|u_2, z_1)$, viewed as functions of $(u_1, z_2) \in [0,1] \times \{0,1\}$ and $(u_2, z_1) \in [0,1] \times \{0,1\}$, respectively.

Consider the assumption that $m_1(1|u_1, z_2)$ is a linear function of u_1 for each value of z_2, so that

$$m_1(1|u_1, z_2) = \theta_{11} + \theta_{12}u_1 + \theta_{13}z_2 + \theta_{14}z_2u_1, \tag{26}$$

for some unknown parameters $\theta_1 \equiv (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14})$. Brinch, Mogstad, and Wiswall (2012; 2017) showed that θ_1 is point identified as long as $p(1,0) \neq p(0,0)$, and $p(1,1) \neq p(0,1)$. Their argument uses the implications of (26) for the observed mean of the treated group:

$$\begin{aligned}
&\mathbb{E}[Y|D = 1, Z_1 = z_1, Z_2 = z_2] \\
&\quad = \mathbb{E}[Y(1)|U_1 \leq p(z_1, z_2), Z_2 = z_2] \\
&\quad = \frac{1}{p(z_1, z_2)} \int_0^{p(z_1,z_2)} m_1(1|u_1, z_2)\, du_1 \\
&\quad = \theta_{11} + \frac{1}{2}p(z_1, z_2)\theta_{12} + z_2 \left[\theta_{13} + \frac{1}{2}p(z_1, z_2)\theta_{14}\right].
\end{aligned} \tag{27}$$

Thus, if $p(1,0) \neq p(0,0)$, then θ_{11} and θ_{12} are point identified by a linear regression of Y on a constant and $\frac{1}{2}p(Z_1, Z_2)$ in the $Z_2 = 0$ subpopulation, while $p(1,1) \neq p(0,1)$ ensures that θ_{13} and θ_{14} can be point identified off of the same linear regression in the subpopulation with $Z_2 = 1$.

The mutual consistency condition exploits the observation that (26) also has implications for the conditional mean of the treated outcome for the *untreated* group. This quantity is not observed, but it can be expressed in terms of θ_1 using an argument similar to (27):

$$\begin{aligned} &\mathbb{E}[Y(1)|D=0, Z_1=z_1, Z_2=z_2] \\ &\quad = \theta_{11} + \frac{1}{2}\left(1+p(z_1,z_2)\right)\theta_{12} + z_2\left[\theta_{13} + \frac{1}{2}\left(1+p(z_1,z_2)\right)\theta_{14}\right]. \end{aligned} \tag{28}$$

Since θ_1 is point identified, these counterfactual mean outcomes are also point identified. They could be used to evaluate treatment parameters that can be expressed in terms of the first selection model with unobservable U_1. The more surprising finding is that they could also be used as additional identifying information for the second selection model with unobservable U_2.

One way to see this is to consider a specification for $m_2(1|u_2, z_1)$ that would typically not be point identified in the current setting. For example, suppose that

$$m_2(1|u_2, z_1) = \theta_{21} + \theta_{22}u_2 + \theta_{23}z_1 + \theta_{24}z_1u_2 + \theta_{25}u_2^2, \tag{29}$$

so that $m_2(1|u_2, z_1)$ is more flexible than $m_1(1|u_1, z_2)$ in having an additional quadratic term. While this function now has five unknown parameters, $\theta_2 \equiv (\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{25})$, there are still only four observed conditional means: $\mathbb{E}[Y|D=1, Z_1=z_1, Z_2=z_2]$ for $(z_1, z_2) \in \{0,1\}^2$. If the selection model for U_2 were viewed in isolation, then θ_2 would not be point identified. However, the mutual consistency condition effectively provides four more moments via (28). Since θ_1 is point identified, these moments can be treated as known.

With eight moments total, it is possible to point identify (indeed, overidentify) the five parameters in θ_2. In analogy to (27) and (28), the system of equations is given by:

$$\begin{bmatrix} 1 & \frac{p(0,0)}{2} & 0 & 0 & \frac{p(0,0)^2}{3} \\ 1 & \frac{p(1,0)}{2} & 0 & 0 & \frac{p(1,0)^2}{3} \\ 1 & \frac{p(0,1)}{2} & 1 & \frac{p(0,1)}{2} & \frac{p(0,1)^2}{3} \\ 1 & \frac{p(1,1)}{2} & 1 & \frac{p(1,1)}{2} & \frac{p(1,1)^2}{3} \\ 1 & \frac{1+p(0,0)}{2} & 0 & 0 & \frac{1-p(0,0)^3}{3(1-p(0,0))} \\ 1 & \frac{1+p(1,0)}{2} & 0 & 0 & \frac{1-p(1,0)^3}{3(1-p(1,0))} \\ 1 & \frac{1+p(0,1)}{2} & 1 & \frac{(1+p(0,1))}{2} & \frac{1-p(0,1)^3}{3(1-p(0,1))} \\ 1 & \frac{1+p(1,1)}{2} & 1 & \frac{(1+p(1,1))}{2} & \frac{1-p(1,1)^3}{3(1-p(1,1))} \end{bmatrix} \begin{bmatrix} \theta_{21} \\ \theta_{22} \\ \theta_{23} \\ \theta_{24} \\ \theta_{25} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[Y|D=1, Z=(0,0)] \\ \mathbb{E}[Y|D=1, Z=(1,0)] \\ \mathbb{E}[Y|D=1, Z=(0,1)] \\ \mathbb{E}[Y|D=1, Z=(1,1)] \\ \mathbb{E}[Y(1)|D=0, Z=(0,0)] \\ \mathbb{E}[Y(1)|D=0, Z=(1,0)] \\ \mathbb{E}[Y(1)|D=0, Z=(0,1)] \\ \mathbb{E}[Y(1)|D=0, Z=(1,1)] \end{bmatrix}. \tag{30}$$

The entire right-hand side of (30) is known: The first four quantities are observed in the data, and the second set of four are identified using the selection model for the first instrument via (28), since θ_1 is point identified. The coefficient matrix on the left-hand side of (30) can be full rank, depending on the values of the propensity score.[7] When this is the case, the linear system of equations either has no solution, or a unique solution. If there is no solution, then the model is misspecified, while if there is a unique solution, then θ_2 is point identified.[8] Thus, the quadratic MTR specification (29) can be point identified even though the only source of exogenous variation in the second selection model is the binary instrument, Z_2. The reason is that the second selection model also harnesses some of the information from the first model—that is, some of the exogenous variation in Z_1—through the mutual consistency condition.

4.2 An Illustrative Example with Partial Identification

The next example shows that the implications of the previous example are not specific to cases with point identification.

Suppose again that there are two instruments, Z_1 and Z_2. As before, assume that $Z_1 \in \{0, 1\}$ is binary, but now suppose that Z_2 is continuous. Actually, assume that Z_2 is not only continuous, but that it has full support, in the sense that the support of $p(z_1, Z_2)|Z_1 = z_1$ is $[0, 1]$ for both $z_1 = 0, 1$. It is well–known that in this case the average treatment effect (ATE) is point identified, since for each z_1 there exists (at least in the limit) an instrument value $\bar{z}_2$ such that $p(z_1, \bar{z}_2) = 1$. This implies that

$$\mathbb{E}[Y(1)|Z_1 = z_1] = \mathbb{E}[Y(1)|Z_1 = z_1, Z_2 = \bar{z}_2] = \mathbb{E}[Y|D = 1, Z_1 = z_1, Z_2 = \bar{z}_2],$$

with a similar argument applying to the case with $d = 0$. See, for example, Heckman (1990), Manski (1990), and Heckman and Vytlacil (2001b).

However, suppose that our interest is not in the ATE, but in an instrument–specific target parameter involving the first instrument. For example, suppose that we are interested in the average treatment effect among the 99% of the population who are most likely to take treatment when measured according to Z_1. This parameter can be written as

$$\text{LATE}_1(0, .99) \equiv \mathbb{E}[Y(1) - Y(0)|U_1 \leq .99]. \tag{31}$$

This object is nearly the same as the ATE; it differs only by the 1% of the population excluded from the conditioning event. If we only had the first instrument at our

[7] For example, take $p(0, 0) = .3$, $p(1, 0) = .45$, $p(0, 1) = .55$, and $p(1, 1) = .7$.

[8] It is common to call θ_2 point identified regardless of which case holds, since the identified set consists of no more than a single element for both cases. The ambiguity comes from whether one is tacitly assuming that the model is correctly specified, which in our notation means $\mathcal{M}$ is not empty. We maintain a distinction between the two cases here just for clarity.

disposal, we would be trying to identify an object that is very nearly the ATE with only a binary instrument. The bounds could be expected to be quite wide if $p(0, Z_2)$ and $p(1, Z_2)$ are far from 0 and 1 for "many" realizations of Z_2.

On the other hand, this line of reasoning ignores the information that we have from the second instrument. That information is sufficient to point identify the ATE, which is nearly the same as the generalized LATE in (31). The relationship between the two objects can be written as

$$\text{LATE}_1(0, .99) = \frac{1}{.99}\Big(\text{ATE} - .01\,\overbrace{\mathbb{E}[Y(1) - Y(0)|U_1 > .99]}^{\equiv \text{LATE}_1(.99,1)}\Big).$$

Since the ATE is point identified from variation in the second instrument, this expression implies that the identified set for $\text{LATE}_1(0, .99)$ can actually be quite narrow. Indeed, if $\underline{y}$ and $\bar{y}$ are the logical bounds on Y, then the identified set for $\text{LATE}_1(0, .99)$ is contained in the interval

$$\left[\frac{1}{.99}\left(\text{ATE} - .01(\bar{y} - \underline{y})\right), \frac{1}{.99}\left(\text{ATE} + .01(\bar{y} - \underline{y})\right)\right],$$

which has width of only $\frac{.02}{.99}(\bar{y} - \underline{y})$.

4.3 Numerical Simulation

In this section, we illustrate how mutual consistency interacts with additional assumptions on the MTR functions using a numerical simulation.[9]

The simulation is like the example in Section 4.1 with two binary instruments. The joint distribution of (Z_1, Z_2) and the propensity score $p(z)$ are shown in Table 1. The propensity score is increasing in each component of Z, so that both instruments can be viewed as incentives that make choosing $D = 1$ more attractive, as in the college attendance example. We assume that $Y \in \{0, 1\}$ is binary, so that conditional expectations of Y are bounded between 0 and 1, and we generate the data using model $\ell = 1$ with an MTR that is linear in u_1 and does not depend on z_2:

$$m_1(0|u_1, z_2) = .5 - .1u_1 \quad \text{and} \quad m_1(1|u_1, z_2) = .8 - .4u_1.$$

In all results that follow, we use a saturated specification of $\mathcal{S}$, so that $\mathcal{S}$ consists of indicator functions $s(d, z) = \mathbb{1}[(D, Z) = (d, z)]$ for all possible combinations of d and z.

Figure 4 reports bounds on the average treatment on the treated (ATT). These

[9]Code for reproducing these simulations is available at `https://github.com/a-torgovitsky/MarginalTreatmentEffectsWithMultipleInstruments.jl`.

$z = (z_1, z_2)$	$\mathbb{P}[Z = z]$	$p(z)$
$(0, 0)$	.4	.3
$(0, 1)$	.3	.5
$(1, 0)$	.1	.6
$(1, 1)$	.2	.7

Table 1: *The joint distribution of (D, Z) for the numerical simulation in Section 4.3.*

bounds are derived under specifications of $m_\ell(d|u_\ell, z_{-\ell})$ that are J_ℓth degree polynomials in u_ℓ, and fully interacted in $z_{-\ell}$, with different parameters for $d = 0$ and $d = 1$. We implement these polynomials using the Bernstein basis so that it is easy to impose shape constraints (see e.g. Section S.2 of MST). There are three sets of bounds shown for increasing values of $J_1 = J_2$, as well as nonparametric bounds indicated with horizontal lines.

The two wider sets of bounds are derived using the $\ell = 1$ and $\ell = 2$ instruments in isolation. The bounds are different because with $\ell = 1$, the instrument is Z_1, with Z_2 serving as a control variable, while with $\ell = 2$ the instrument is Z_2, with Z_1 as a control. The third set of bounds is computed while also imposing mutual consistency between the two models. This substantially tightens both the nonparametric bounds and the polynomial bounds at all polynomial degrees.

Notice in particular that the mutual consistency bounds are tighter than the intersections of the $\ell = 1$ and $\ell = 2$ bounds. This shows that mutual consistency is *not* just a matter of taking intersection bounds across differ instruments used separately. Instead, it involves harmonizing the intricate common predictions about instrument-invariant quantities that one would obtain using each instrument separately, as formalized through the set of equalities (23). These equalities effectively combine the information from the two instruments into a whole that is greater than the sum of their parts. As Figure 4 shows, this can substantially tighten inference. For example, the nonparametric bounds under mutual consistency are as tight as the bounds using each instrument separately with a 5th degree polynomial.

In Figure 5, we report bounds on $\text{LATE}_1(+\delta\%)$, as defined in (17) for $\delta = 20$. This quantity can only be expressed in terms of the unobservable U_1 for the first instrument. Nevertheless, comparing the four sets of bounds in Figure 5 shows that the second instrument provides information on $\text{LATE}_1(+20\%)$ through the mutual consistency condition. Thus, the mutual consistency condition allows information from the second instrument to inform the MTR functions for the first instrument. This extra information results in tighter bounds than would be possible using the first instrument

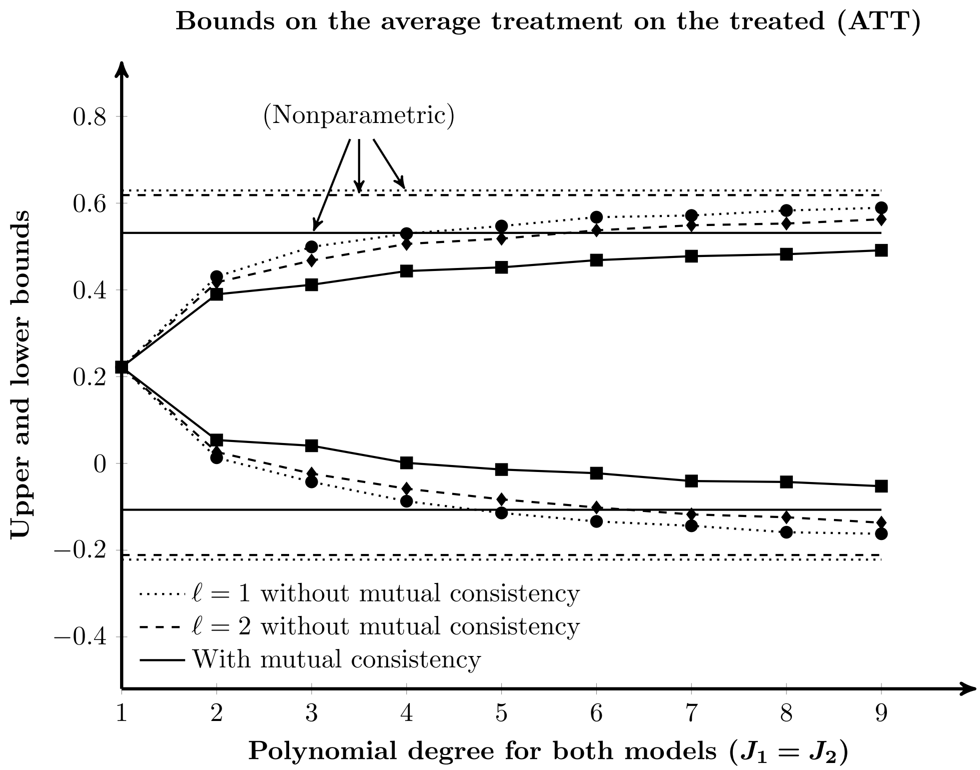


Figure 4: *Imposing mutual consistency tightens bounds on the average treatment on the treated (ATT) for both parametric and nonparametric specifications of the MTR functions.*

in isolation. In this data generating process, the additional information is small (but still present) when m_2 is left nonparametric. Adding the nonparametric shape constraints that $m_2(0|\cdot, z_1), m_2(1|\cdot, z_1)$ and $m_2(1|\cdot, z_1)-m_2(0|\cdot, z_2)$ are decreasing functions for every z_1 provides substantially more information.

A vivid case occurs when m_2 is specified as linear ($J_2 = 1$). Under this assumption, all instrument-invariant quantities are point identified using only variation in Z_2. A parameter that is specific to the first instrument, like $\text{LATE}_1(+20\%)$, generally remains partially identified. Suppose, however, that we impose the assumption that m_1 is quadratic ($J_1 = 2$). If we were using only variation in Z_1, then we would still expect $\text{LATE}_1(+20\%)$ to be partially identified. Indeed we can see that this is the case in Figure 5, where the bounds without imposing mutual consistency are approximately $[.075, .175]$ when $J_1 = 2$. Imposing mutual consistency with m_2 linear collapses these bounds into a single point, consistent with the example discussed in Section 4.1.

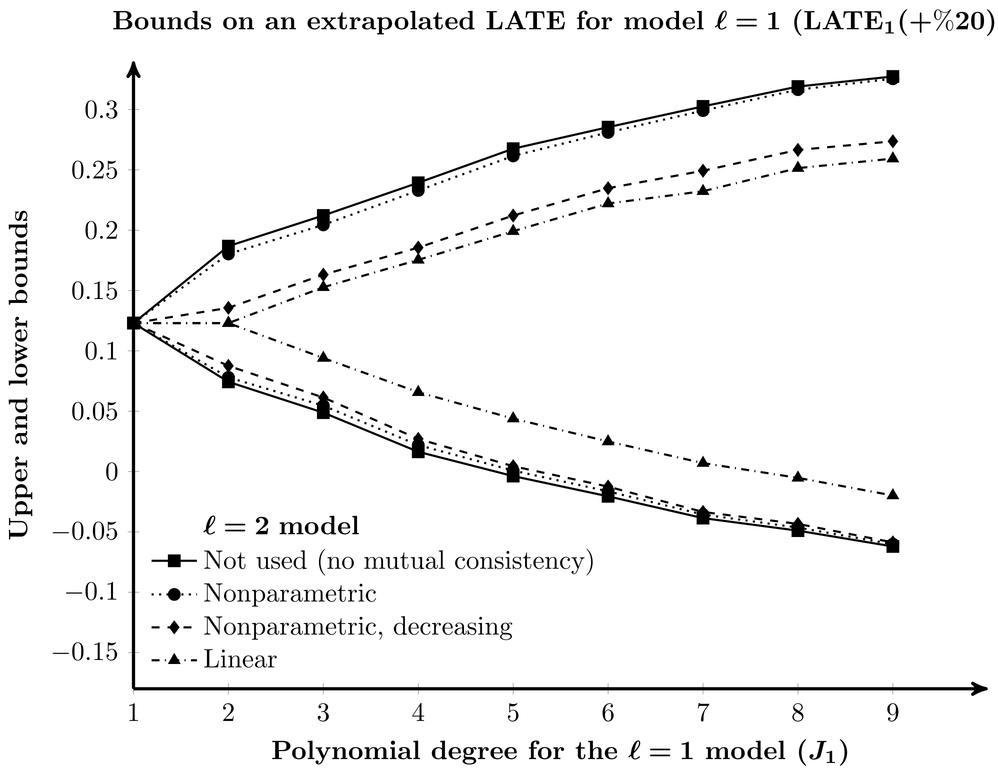


Figure 5: *The $\ell = 2$ model provides identifying content for parameters, such as $LATE_1(+\%20)$, that can only be defined using the $\ell = 1$ model.*

4.4 Using Assumption IAM When it Does Not Hold

As discussed in Section 3.5, Assumption IAM has testable implications given Assumptions E. Heckman and Vytlacil (2001b) showed that when these testable implications are not violated, the sharp nonparametric bounds on the ATE, ATT, and ATU are the same as Manski's (1990; 1994) IV bounds, which do not maintain any assumptions on the selection process. Since Assumption PM is weaker than Assumption IAM, the same conclusions also hold for Assumption PM. Thus, sharp nonparametric bounds on the ATE, ATT, and ATU are the same function of the joint support of $p(Z_1, Z_2)$ whether or not one maintains Assumptions IAM or PM. For example, if $p(z_1, z_2) = \frac{1}{2}(z_1 + z_2)$, with Z_1 uniform over $[0, 1]$ and Z_2 binary, then the ATE, ATT, and ATU are nonparametrically point identified with or without Assumptions IAM or PM, because the joint support of $p(Z_1, Z_2)$ is $[0, 1]$, even though the support induced when conditioning on either Z_1 or Z_2 is a proper subset of $[0, 1]$.

Assumptions IAM and PM start to be empirically useful when imposing additional assumptions (parametric structure or nonparametric shape restrictions; recall Figure

4), or when the target parameter is a quantity that depends on the choice model, such as a LATE or a PRTE. In either case, relying on Assumption IAM when it is violated can lead to incorrect inference.

We illustrate this point with a nonparametric analysis of a PRTE. As in the previous section, suppose that there are two binary instruments, Z_1 and Z_2. In the observed baseline, the support of (Z_1, Z_2) only has three points, $(0,0), (1,0)$, and $(1,1)$, with respective probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. In the new policy, the support of (Z_1, Z_2) is expanded to also include $(0,1)$ by taking half of the mass placed on $(1,0)$ and applying it to $(0,1)$, so that each of the four instrument values are equally likely. Letting $D^\star$ be the treatment decision in the new policy, and $Y^\star = D^\star Y(1) + (1 - D^\star)Y(0)$ the associated outcome, the (aggregate) PRTE is defined as

$$\text{PRTE} \equiv \mathbb{E}[Y^\star - Y],$$

see e.g. Heckman and Vytlacil (2007b) or Carneiro et al. (2010).

Suppose that the propensity score has the following order:

$$0 < p(0,0) < p(0,1) < p(1,0) < p(1,1) < 1. \tag{32}$$

Then under Assumption IAM, one would conclude that strictly fewer individuals will participate under the new policy, since half of those who previously received $Z = (1,0)$ now receive $Z = (0,1)$, and $p(0,1) < p(1,0)$. In terms of the single threshold-crossing selection model (3) with a single (normalized) unobservable U, one would conclude that

$$\text{PRTE} = -\frac{1}{4} \int_{p(0,1)}^{p(1,0)} m(1|u) - m(0|u) \, du, \tag{33}$$

where $m(d|u) \equiv \mathbb{E}[Y(d)|U = u]$. If, for example, one believed the effect of the treatment to be positive for everyone, then one would unambiguously conclude that the PRTE is negative.

However, that conclusion could be wrong if Assumption PM holds but Assumption IAM does not. To see why, consider the taxonomy of choice types under Assumption PM introduced by Mogstad et al. (2020), which is reproduced in Table 2. Under Assumption PM there are four types of compliers. For the Z_1 compliers, getting shifted from $Z = (1,0)$ to $Z = (0,1)$ means leaving treatment. But for the Z_2 compliers, that same shift means entering treatment. Whether the PRTE is positive or negative depends on the relative proportions of Z_1 and Z_2 compliers, and on their respective

Group name ($G = g$)	$D(0,0)$	$D(0,1)$	$D(1,0)$	$D(1,1)$	$\mathbb{P}[G = g]$	$\mathbb{E}[Y(1)\|G = g]$
Always-taker	1	1	1	1	.1	.6
Eager complier	0	1	1	1	.1	.6
Reluctant complier	0	0	0	1	.1	.2
Never-taker	0	0	0	0	.1	.1
Z_1 complier	0	0	1	1	.4	.1
Z_2 complier	0	1	0	1	.2	.6

Table 2: *Groups definitions under Assumption PM with $Z_1, Z_2 \in \{0,1\}$. The probabilities and average treatment effects are used in the numerical example in Section 4.4.*

average treatment effects. Formally, the PRTE computed under the first selection model with the first instrument can be written as

$$\begin{aligned}\text{PRTE}_1 = \frac{1}{4}&\left(\int_0^{p(0,1)} m_1(1|u_1,1) - m_1(0|u_1,1)\,du_1\right)\\ &- \frac{1}{4}\left(\int_0^{p(1,0)} m_1(1|u_1,0) - m_1(0|u_1,0)\,du_1\right), \qquad (34)\end{aligned}$$

or under the second model as

$$\begin{aligned}\text{PRTE}_2 = \frac{1}{4}&\left(\int_0^{p(0,1)} m_2(1|u_2,0) - m_2(0|u_2,0)\,du_2\right)\\ &- \frac{1}{4}\left(\int_0^{p(1,0)} m_2(1|u_2,1) - m_2(0|u_2,1)\,du_2\right). \qquad (35)\end{aligned}$$

Thus, under Assumption PM the PRTE can be either positive or negative, even if the treatment effect is positive for everyone.

Figure 6 illustrates this point with a numerical example. The data is drawn according to the type probabilities and conditional means shown in Table 2. For simplicity, $Y(0)$ is set to 0, and $Y(1) \in \{0,1\}$ is binary, so that the treatment effect is positive for everyone. Figure 6 shows that the nonparametric upper bound on the PRTE implied under Assumption IAM (equation (33)) is indeed 0, even though the true PRTE is positive. Nonparametric bounds on PRTE_1 or PRTE_2 that use each instrument separately under Assumption IAM contain 0 and the true PRTE, but are quite wide. Maintaining Assumption PM with both instruments and imposing mutual consistency narrows the nonparametric bounds considerably, while still containing both 0 and the true, positive value of the PRTE.

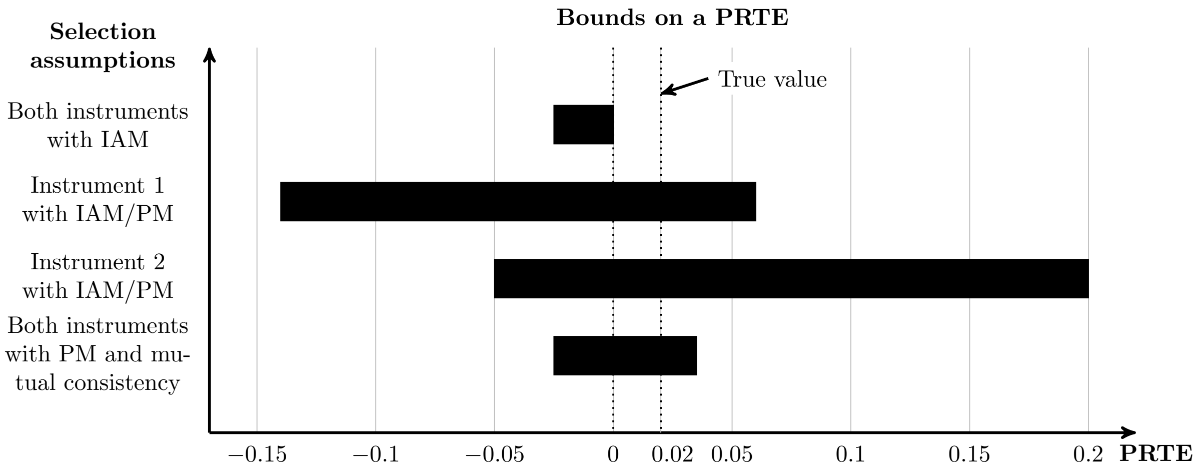


Figure 6: *Bounds on the PRTE of shifting $\frac{1}{4}$ mass from $Z = (1, 0)$ to $Z = (0, 1)$.*

5 Conclusion

A central conclusion of the modern IV literature is that the parameter estimated by a traditional linear IV estimator depends on the instrument itself. This conclusion creates a problem: certain instruments may lead to less relevant parameters and there might be no available instrument that answers the researcher's specific scientific or policy question. MTE methods address this problem by returning focus to the definition of the target parameter, leaving the specifics of how it can be identified (parametrically, nonparametrically, partially, etc.) as a separate and conceptually distinct issue. However, MTE methods crucially depend on the monotonicity condition (threshold-crossing equation) introduced by Imbens and Angrist (1994). This condition is extremely strong when there are multiple instruments, since it assumes away all meaningful unobserved choice heterogeneity (Mogstad et al., 2020).

In this paper, we have extended the MTE methodology under a weaker, partial monotonicity condition. Partial monotonicity allows for rich patterns of unobserved heterogeneity in choices, while still remaining rooted in an interpretable choice-theoretic model that is fundamentally nonparametric. We showed how to modify the general partial identification framework of Mogstad et al. (2018) to allow for partial monotonicity instead of the stronger, traditional monotonicity condition. The framework provides a general, flexible way for researchers to explore the assumptions–conclusion frontier through different parametric and nonparametric shape restrictions on the underlying marginal treatment response functions. It can be implemented at scale using linear programming.

An unusual feature of the framework is that it can be viewed as having multiple different selection models for the same treatment. In order to rationalize these models simultaneously, we imposed a condition called mutual consistency. The mutual consistency condition effectively allows information from one instrument about one marginal treatment response function to be transferred to another marginal treatment response function defined by a different instrument. This allows for the accumulation of identifying content from multiple instruments. The method provides a path for extracting and aggregating information about treatment effects from multiple different sources of exogenous variation while still maintaining a plausible model of choice behavior and allowing for rich unobserved heterogeneity.

References

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. 14

Arnold, D., W. Dobbie, and C. S. Yang (2018): "Racial Bias in Bail Decisions," *The Quarterly Journal of Economics*, 133, 1885–1932. 1

Arnold, D., W. S. Dobbie, and P. Hull (2020): "Measuring Racial Discrimination in Bail Decisions," Working Paper 26999, National Bureau of Economic Research. 1

Autor, D., A. Kostøl, M. Mogstad, and B. Setzler (2019): "Disability Benefits, Consumption Insurance, and Household Labor Supply," *American Economic Review*, 109, 2613–54. 1

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 19

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2020): "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 128, 1269–1324. 1

Björklund, A. and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *The Review of Economics and Statistics*, 69, 42–49. 1

Brinch, C. N., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," *Working paper.* 21

——— (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039. 1, 21

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 L. N. Christofides, K. E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222. 6

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44, 361–422. 3

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, 377–394. 14, 15, 28

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 1

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. 1

Carneiro, P., M. Lokshin, and N. Umapathi (2016): "Average and Marginal Returns to Upper Secondary Schooling in Indonesia," *Journal of Applied Econometrics*, 32, 16–36. 1

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018): "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance," *Journal of Political Economy*, 126, 2356–2409. 1

Cunha, F., J. J. Heckman, and S. Navarro (2007): "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds," *International Economic Review*, 48, 1273–1309. 3

Depalo, D. (2020): "Explaining the Causal Effect of Adherence to Medication on Cholesterol through the Marginal Patient," *Health Economics*, n/a. 1

Doyle Jr., J. J. (2007): "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *The American Economic Review*, 97, 1583–1610. 1

Felfe, C. and R. Lalive (2018): "Does Early Child Care Affect Children's Development?" *Journal of Public Economics*, 159, 33–53. 1

French, E. and J. Song (2014): "The Effect of Disability Insurance Receipt on Labor Supply," *American Economic Journal: Economic Policy*, 6, 291–337. 1

Gautier, E. (2020): "Relaxing Monotonicity in Endogenous Selection Models and Application to Surveys," *arXiv:2006.10997 [math, stat]*. 3

Gautier, E. and S. Hoderlein (2015): "A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation," *arXiv:1109.0362 [math, stat]*. 3

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694. 1, 5

——— (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *The Journal of Human Resources*, 32, 441–462. 1

Heckman, J., J. L. Tobias, and E. Vytlacil (2001): "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, 68, 210. 5

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*. 1, 5

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

——— (1990): "Varieties of Selection Bias," *The American Economic Review*, 80, 313–318. 23

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 6

——— (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–98. 9

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35. 3

Heckman, J. J. and J. A. Smith (1998): "Evaluating the Welfare State," *NBER Working Paper 6542*. 1

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. 1, 3

——— (2008): "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case," *Annales d'Economie et de Statistique*, 91/92, 151–174. 3

Heckman, J. J. and E. Vytlacil (2001a): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–111. 1, 14

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 1, 6, 7, 12, 14, 15

Heckman, J. J. and E. J. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734. 1, 12, 14

——— (2001b): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner and F. Pfeiffer, Heidelberg and Berlin: Physica. 23, 27

——— (2001c): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by K. M. C Hsiao and J. Powell, Cambridge University Press. 1, 12

——— (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4779–4874. 12

——— (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4875–5143. 3, 4, 12, 14, 28

Huber, M. and G. Mellace (2014): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411. 19

Imbens, G. W. (2014): "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29, 323–358. 6

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 1, 4, 5, 6, 30

Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64, 555–574. 19

Kane, T. J. and C. E. Rouse (1993): "Labor Market Returns to Two- and Four-Year Colleges: Is a Credit a Credit and Do Degrees Matter?" Working Paper 4268, National Bureau of Economic Research. 6

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. 19

Kline, P. and C. R. Walters (2016): "Evaluating Public Programs with Close Substitutes: The Case of Head Start*," *The Quarterly Journal of Economics*, 131, 1795–1848. 1, 3

Kowalski, A. E. (2018): "Behavior within a Clinical Trial and Implications for Mammography Guidelines," Working Paper 25049, National Bureau of Economic Research. 1

Lee, S. and B. Salanié (2018): "Identifying Effects of Multivalued Treatments," *Econometrica*, Forthcoming. 3

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *The American Economic Review*, 103, 1797–1829. 1

Manski, C. (1994): "The Selection Problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 27

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 23, 27

Marx, P. (2020): "Sharp Bounds in the Latent Index Selection Model," *arXiv:2012.02390 [econ]*. 4

Moffitt, r. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations," *Annales d'Economie et de Statistique*, 239–261. 1

Moffitt, R. A. (2019): "The Marginal Labor Supply Disincentives of Welfare Reforms," Working Paper 26028, National Bureau of Economic Research. 1

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *NBER Working Paper*. 1

——— (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619. 2, 4, 12, 30

Mogstad, M. and A. Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10. 1, 5, 14, 17

Mogstad, M., A. Torgovitsky, and C. R. Walters (2020): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *Working paper*. 1, 2, 5, 7, 9, 14, 28, 30

Mountjoy, J. (2019): "Community Colleges and Upward Mobility," *Working paper*. 3, 7

Mourifié, I. and Y. Wan (2016): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99, 305–313. 19

Nybom, M. (2017): "The Distribution of Lifetime Earnings Returns to College," *Journal of Labor Economics*, 000–000. 1

Pinto, R. (2019): "Noncompliance as a Rational Choice: A Framework That Exploits Compromises in Social Experiments to Identify Causal Effects," . 3

Shea, J. and A. Torgovitsky (2019): "Ivmte: An R Package for Implementing Marginal Treatment Effect Methods," *Working Paper*. 17

Torgovitsky, A. (2019): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505. 17

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 1, 4, 5, 9

Walters, C. R. (2018): "The Demand for Effective Charter Schools," *Journal of Political Economy*, 126, 2179–2223. 1