Identification and Extrapolation with Instrumental Variables[*]

Magne Mogstad[†] Alexander Torgovitsky[‡]

September 19, 2017

Abstract

Instrumental variables (IV) are widely used in economics to address selection on unobservables. Standard IV methods produce estimates of causal effects that are specific to individuals whose behavior can be manipulated by the instrument at hand. In many cases, these individuals are not the same as those who would be induced to treatment by an intervention or policy of interest to the researcher. The average causal effect for the two groups can differ significantly if the effect of the treatment varies systematically with unobserved factors that are correlated with treatment choice. We review the implications of this type of unobserved heterogeneity for the interpretation of standard IV methods and for their relevance to policy evaluation. We argue that drawing inference about policy relevant parameters typically requires extrapolating from the individuals affected by the instrument to the individuals who would be induced to treatment by the policy under consideration. We discuss a variety of alternatives to standard IV methods that can be used to perform this extrapolation rigorously. We show that many of these approaches can be nested as special cases of a general framework that embraces the possibility of partial identification.

[*] This paper is prepared for the Annual Review of Economics. We thank Christian Brinch and Jack Mountjoy for useful comments and suggestions.

[†] 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-1426882.

1 Introduction

Instrumental variable (IV) methods are widely used in empirical work in economics and other fields. Their attraction stems from the hope that an instrument provides a source of exogenous variation that can be used to infer the causal impact of an endogenous treatment variable on an outcome of interest. IV methods attack the problem of selection on unobservables by only using variation in the treatment that is induced by the instrument. This variation only represents individuals whose treatment choice would be affected by changes in the instrument. As a consequence, standard IV methods, such as two-stage least squares, produce estimates of causal effects that are specific to these individuals.

In some cases, these estimates can be of intrinsic interest, for example if the instrument itself represents an intervention or policy change of interest. In many other cases, the group of individuals who would be affected by the available instrument are different from the group of individuals who would be affected by a policy.[1] If the effect of the treatment varies between the two groups, then the estimates produced by standard IV methods may differ dramatically from the parameters relevant for drawing inference about the effects of a policy of interest.[2] This raises concerns about the external validity of estimates produced by standard IV methods, and about their relevance for policy evaluation.

In this paper, we first review the implications of unobserved heterogeneity in treatment effects for the interpretation and policy relevance of standard IV methods. Then, we discuss alternatives to the standard methods that can be used to learn about policy relevant parameters in settings with unobserved heterogeneity in treatment effects. We argue that drawing inference about such parameters typically requires extrapolating from the individuals affected by the instrument to the individuals who would be induced to treatment by the policy under consideration. The external validity and policy relevance of IV methods turns on the ability to do this extrapolation reliably.

The structure of our review is as follows. In the next section, we review a widely

[1] For example, many instruments are based on the lack of pattern or predictability in certain natural events that cannot be shifted by policy, such as the weather (Angrist, Graddy, and Imbens, 2000; Miguel, Satyanath, and Sergenti, 2004), or the gender composition of children (Angrist and Evans, 1998).

[2] A number of studies in diverse fields report evidence of this type of unobserved heterogeneity in treatment effects. Heckman (2001) compiled a list of studies. More recent papers include Bitler, Gelbach, and Hoynes (2006), Doyle Jr. (2007), Moffitt (2008), Carneiro and Lee (2009), Firpo, Fortin, and Lemieux (2009), Carneiro, Heckman, and Vytlacil (2011), Maestas, Mullen, and Strand (2013), Bitler, Hoynes, and Domina (2014), Walters (2014), Felfe and Lalive (2014), French and Song (2014), Havnes and Mogstad (2015), Kirkeboen, Leuven, and Mogstad (2016), Kline and Walters (2016), Hull (2016), Carneiro, Lokshin, and Umapathi (2016), Cornelissen, Dustmann, Raute, and Schönberg (forthcoming), Nybom (2017), and Brinch, Mogstad, and Wiswall (2017), among many others.

studied IV model with a binary treatment. The model maintains the existence of an exogenous instrument that has a monotonic effect on treatment in the sense developed by Imbens and Angrist (1994). This monotonicity condition gives rise to the important related concepts of the marginal treatment effect and response functions developed by Heckman and Vytlacil (1999, 2005). Our review focuses on this model due to its central role in the recent literature on IV methods.

In Section 3, we introduce a general definition of a target parameter as a weighted average of the marginal treatment response functions. We view the target parameter as an object chosen by the researcher to answer a specific well-defined policy question. We argue here, and throughout the paper, that some conventional treatment parameters, such as the average treatment effect, often represent an uninteresting policy counterfactual, and so make for uninteresting target parameters. We recommend that researchers focus instead on target parameters in the class of policy relevant treatment effects (PRTEs) introduced by Heckman and Vytlacil (2001a). These parameters allow researchers to consider interventions that influence (but may not fully determine) an individual's treatment choice, for example by changing the costs associated with the treatment alternatives. We discuss specific examples of PRTEs, and show that the local average treatment effect of Imbens and Angrist (1994) can be viewed as a special case of a PRTE.

In Section 4, we discuss two conditions under which the PRTE and other target parameters are non-parametrically point identified. We argue that both of these conditions are too restrictive for many settings that involve policies that represent meaningful departures from the status quo. Evaluating such a policy requires extrapolating from the individuals whose treatment choice is affected by the available instrument to the individuals whose treatment choice would be affected by the policy.

This need to extrapolate motivates Section 5 where we consider a general framework proposed by Mogstad, Santos, and Torgovitsky (2017), in which data and a priori assumptions can be flexibly combined to produce bounds on PRTEs and other target parameters. We show that the tightness of the bounds—that is, the strength of the conclusions that one can obtain—naturally depends both on the extent of extrapolation required, and on the strength of the a priori assumptions that are maintained. As a result, the framework allows the researcher to achieve bounds that are as narrow as they desire, while requiring them to honestly acknowledge the strength of their assumptions and the degree of extrapolation involved in their counterfactual. In Section 6, we discuss the relationship between the general framework of Mogstad et al. (2017) and previous work, showing that it nests several previous approaches to extrapolation as special cases. In Section 7, we summarize and conclude with some directions for future research.

Our review focuses on the identification problem of using the distribution of the observed data to learn about parameters of interest. In practice, researchers do not know the population distribution of the observed data with certainty. Features of this distribution need to be estimated from the available sample, and most researchers would agree that it is important to formally account for statistical uncertainty in these estimates. We set these issues of statistical inference aside in our review. We view the identification problem as both distinct from—and primary to—the problem of statistical inference, since the conclusions one can make under imperfect knowledge of the population distribution of the data are a subset of those that can be drawn under perfect knowledge. Having said this, the general framework that we discuss in Section 5 involves some challenges for statistical inference. Mogstad et al. (2017) provide a complete discussion of these challenges and develop a method for addressing them.

2 Model

2.1 Potential Outcomes

Our discussion focuses on the canonical program evaluation problem with a binary treatment $D \in \{0, 1\}$ and a scalar, real-valued outcome, Y. Corresponding to the two treatment arms are potential outcomes, Y_0 and Y_1. These represent the realizations of Y that would have been experienced by an agent had their treatment status been exogenously set to 0 or 1. The relationship between observed and potential outcomes is given by

$$Y = DY_1 + (1 - D)Y_0. \tag{1}$$

In economic applications with observational data, it is often implausible to assume that D is exogenously determined relative to Y_0 and Y_1, especially if D is a choice variable. When D is endogenous, contrasting the distribution of Y for the treated $(D = 1)$ and control $(D = 0)$ groups confounds the effect of the treatment with other differences across these groups. Conditioning on observed covariates, X, can conceivably unconfound the effect of D on Y. However, one often expects that there are important factors that influence the choice of D, such as an agent's beliefs about Y_0 and Y_1, that are fundamentally difficult to observe, and therefore not part of X. The idea of an IV method is to use the variation from an instrument, Z, to indirectly shift D while holding X fixed. If Z is exogenous, then the resulting variation in Y results solely from the causal effect of D on Y, i.e. from the difference between Y_1 and Y_0.

2.2 Assumptions

A key theme of the literature, and of this paper, is that considering how Z affects the choice of D is crucial when there is unobserved heterogeneity in the causal effect of D on Y. Intuitively, if different individuals stand to gain or lose differently from receiving treatment, then it is important to model which individuals select into treatment.

In an influential paper, Imbens and Angrist (1994) introduced a simple model of choice behavior summarized by what they called the monotonicity condition. This condition says that, given X, an exogenous shift of Z from one value to another either weakly increases the choice of D for every agent, or else it weakly decreases it for every agent. Vytlacil (2002) showed that under the standard exogeneity assumption on Z, the monotonicity condition is equivalent to the existence of a weakly separable selection (or choice) equation,

$$D = \mathbb{1}[\nu(X, Z) - U \geq 0], \tag{2}$$

where ν is an unknown function and U is a continuously distributed random variable.

Our review is focused on approaches that maintain the monotonicity condition or, equivalently, the choice model (2). This choice model is widely used, but of course it is not beyond criticism. In Section 6.5, we compare these approaches with another influential framework for extrapolation that does not maintain a choice model and therefore does not use the monotonicity condition. Our view is that maintaining some choice model (although not necessarily (2)) is crucial for considering counterfactual policies that do not mandate a choice of treatment.

The period since Imbens and Angrist (1994) has witnessed the evolution of a large literature that explores the implications of choice model (2) for IV methods. The following set of assumptions are commonly maintained in this literature. We will maintain them throughout our discussion as well.[3]

Assumptions IV

IV.1 *D is determined by* (2).

IV.2 $(Y_0, Y_1, U) \perp\!\!\!\perp Z | X$, *where* $\perp\!\!\!\perp$ *denotes conditional independence.*

IV.3 *U is continuously distributed, conditional on X.*

Assumption IV.2 requires Z to be exogenous with respect to both the selection and outcome processes after conditioning on covariates, X. If one is only concerned with

[3] Our discussion also requires some mild technical conditions involving the existence of moments that we do not explicitly mention, but which will be clear from the context.

mean outcomes, then this assumption can be weakened to the combination of $U \perp\!\!\!\perp Z|X$ and $E[Y_d|U,X,Z] = E[Y_d|U,X]$ for $d = 0,1$. In applications, it can be difficult to think of reasons for which this weaker assumption would hold while IV.2 would fail. For simplicity, we maintain the stronger assumption throughout our discussion.

Given IV.3, one can normalize the distribution of $U|X = x$ to be uniformly distributed over $[0,1]$ for every x.[4] Under this normalization, and given IV.2, it is straightforward to show that $\nu(x,z)$ is equal to the propensity score,

$$p(x,z) \equiv P[D = 1|X = x, Z = z]. \tag{3}$$

Hence, the normalization allows (2) to be rewritten as

$$D = \mathbb{1}[U \leq p(X,Z)] \quad \text{where} \quad U|X = x, Z = z \sim \text{Unif}[0,1] \text{ for all } x, z. \tag{4}$$

Working with (4) instead of (2) simplifies the subsequent expressions, without changing the empirical implications of any of the results we discuss. It is worth repeating that the work of Vytlacil (2002) proves that (4) together with Assumptions IV is equivalent to the influential IV model introduced by Imbens and Angrist (1994).

2.3 Marginal Treatment Response Functions

An important unifying concept for IV methods that maintain the weakly separable choice model (4) is the marginal treatment effect (MTE), which was developed in a series of papers by Heckman and Vytlacil (1999, 2001a,b,c, 2005, 2007a,b).[5] The MTE is defined as

$$\text{MTE}(u,x) \equiv E[Y_1 - Y_0|U = u, X = x]. \tag{5}$$

The dependence of the MTE on u for a fixed x reflects unobserved heterogeneity in treatment effects, as indexed by an agent's latent propensity to choose treatment, u. The choice equation (4) implies that, given X, individuals with lower values of U are more likely to take treatment, regardless of their realization of Z.[6] An MTE function that is declining in u would therefore indicate that individuals who are more likely to choose $D = 1$ also experience larger gains in Y from receiving the treatment. The

[4] This type of normalization argument appears in many guises in the literature on nonparametric identification. It is one of many possible normalizations, see e.g. Matzkin (2007) for a complete discussion.

[5] As Heckman and Vytlacil recognize, the key ideas behind the MTE can be found in an earlier paper by Björklund and Moffitt (1987), albeit in a parametric context.

[6] This is only a convention; if (2) were written instead as $\mathbb{1}[\nu(X,Z) + U \geq 0]$, then higher values of U would be more likely to take treatment.

case of no unobserved treatment effect heterogeneity corresponds to an MTE function that is constant in u. Similarly, observed treatment effect heterogeneity is described through the dependence of the MTE function on x for a fixed u.

Instead of working with the MTE function directly, we consider treatment parameters that can be expressed as functions of the two marginal treatment response (MTR) functions, defined as

$$m_0(u,x) \equiv E\left[Y_0 \mid U=u, X=x\right] \quad \text{and} \quad m_1(u,x) \equiv E\left[Y_1 \mid U=u, X=x\right]. \qquad (6)$$

Each pair $m \equiv (m_0, m_1)$ of MTR functions generates an associated MTE function $m_1(u,x) - m_0(u,x)$, so there is no cost in generality from working with MTR functions directly. As we discuss later, an important advantage of working with MTR functions instead of MTE functions is that it allows one to consider parameters and estimands that depend on m_0 and m_1 asymmetrically. For example, the OLS estimand can be written as a weighted average of m_0 and m_1, whereas this interpretation is not available when working only with their difference.

2.4 A Running Numerical Illustration

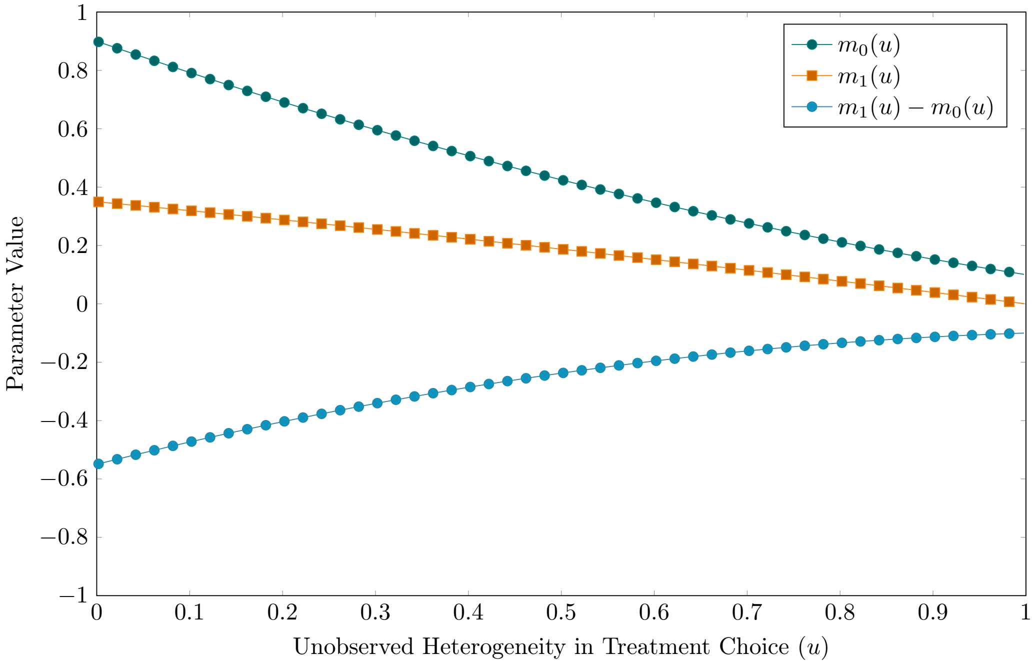
Throughout the paper, we will use a running numerical example to provide graphical explanations of the key concepts. The example is loosely based on the empirical application in Mogstad et al. (2017). They analyze how a class of potential subsidy regimes can promote the use of the health product, and compare increases in usage to the costs of subsidization. In their application, D is a binary indicator for purchasing a mosquito net (the health product), Z is an experimentally varied subsidy for the net and (for simplicity) there are no covariates X. The data is taken from Dupas (2014) and features a variety of different subsidy levels.

For the numerical illustration, we bin these subsidies into four ascending groups, so that $Z \in \{1,2,3,4\}$, with $Z=4$ denoting the most generous subsidy. The groups are approximately equally likely, so we take $P[Z=z] = \frac{1}{4}$ for each z. We take the propensity score in our simulation to be equal to the estimated propensity score in the data, which is given by

$$p(1) = .12 \qquad p(2) = .29 \qquad p(3) = .48 \qquad p(4) = .78.$$

We take the outcome in our numerical example to be binary, i.e. $Y \in \{0,1\}$. To fix ideas, we think of Y as an indicator for whether an individual is infected by malaria.

Figure 1: MTR and MTE Functions Used to Generate Data in the Numerical Illustration



We take the MTR (and implied MTE) functions to be quadratic functions of u:

$$\begin{aligned} m_0(u) = .9 - 1.1u + .3u^2 \qquad &\text{and} \qquad m_1(u) = .35 - .3u - .05u^2, \\ \text{so that} \qquad m_1(u) - m_0(u) &= -.55 + .8u - .35u^2. \end{aligned} \tag{7}$$

As shown in Figure 1, the MTR functions for both the treated and untreated states are decreasing in u. Recalling that higher values of u correspond to lower propensities to choose treatment, this means that individuals less likely to purchase the mosquito net ($D = 1$) are also less likely to be afflicted by malaria ($Y = 1$) regardless of whether they purchase the mosquito net. This could arise because individuals differ in their degree of susceptibility to malaria and have some private knowledge of their personal vulnerability to the disease. Figure 1 shows that the m_1 function is larger than the m_0 function for all values of u, which means that the mosquito net reduces the incidence of malaria for all individuals. However, the difference between m_1 and m_0 (the MTE) is non-constant, and is larger for individuals who are more likely to purchase the net. This increasing pattern in $m_1 - m_0$ could arise if individuals have an idea of how likely they are to benefit from a mosquito net—for example, due to the prevalence of mosquitoes in their sleeping area—and partly base their purchase decision on this knowledge.

3 What We Want to Know: Target Parameters

3.1 Definition

Before considering identification, the researcher needs to define their parameter of interest, which we refer to as the target parameter, $\beta^\star$. We assume that the researcher has a specific well-defined policy question that they are interested in, and that this question suggests one or more relevant target parameters. A central theme of our discussion is that different target parameters can be relevant for different applications and policy questions. This motivates a framework in which the researcher is allowed wide latitude in how they can specify the target parameter. To do this, we only require that $\beta^\star$ can be written as a weighted average of the unknown MTR functions. Formally, we assume that

$$\beta^\star \equiv E\left[\int_0^1 m_0(u, X)\omega_0^\star(u, X, Z)\, du\right] + E\left[\int_0^1 m_1(u, X)\omega_1^\star(u, X, Z)\, du\right], \qquad (8)$$

for some identified weighting functions, $\omega_0^\star$ and $\omega_1^\star$.

Different target parameters are generated by choosing different pairs of $(\omega_0^\star, \omega_1^\star)$. We discuss several types of target parameters in the following sections. Our Tables 1–4 provide an extensive catalog of the weighting functions that correspond to these parameters. Of course, it is impossible to specify the universe of target parameters that could be of possible interest for an application. Fortunately, deriving the weighting functions $(\omega_0^\star, \omega_1^\star)$ that generate a given parameter can be accomplished relatively easily by appropriately modifying the arguments in Heckman and Vytlacil (2005).[7]

3.2 Conventional Target Parameters

The average treatment effect (ATE) is a widely studied target parameter. As shown in Table 1, the ATE can be written as (8) by specifying the weight functions as $\omega_1^\star(u, x, z) = 1$ and $\omega_0^\star(u, x, z) = -1$. This equally weights the individual level treatment effects regardless of differences across individuals. The ATE can be interpreted as the average change in outcomes that would be realized if all individuals were required to choose $D = 1$, compared to the regime in which all individuals are forbidden to choose $D = 1$.

[7] Most of the expressions in Tables 1–4 can be found in Heckman and Vytlacil (2005). The expressions for parameters with asymmetric weights (i.e. $m_0 \neq -m_1$) were derived by Mogstad et al. (2017). Note also that Mogstad et al. (2017) consider a slightly more general version of (8) in which the integrating measure (i.e. "du") can be something other than Lebesgue measure. For example, this allows one to define the target parameter to be the MTE at a given value $\widetilde{u}$, i.e. $\beta^\star = E[m_1(\widetilde{u}, X) - m_0(\widetilde{u}, X)]$.

Table 1: Weights for Conventional Treatment Effect Parameters

		Weights	
Target Parameter	Expression	$\omega_0^\star(u,x,z)$	$\omega_1^\star(u,x,z)$
Average Untreated Outcome	$E[Y_0]$	1	0
Average Treated Outcome	$E[Y_1]$	0	1
Average Treatment Effect (ATE)	$E[Y_1 - Y_0]$	-1	1
ATE given $X = \overline{x}$ where $P[X = \overline{x}] > 0$	$E[Y_1 - Y_0 \mid X = \overline{x}]$	$-\omega_1^\star(u,x,z)$	$\dfrac{\mathbb{1}[x = \overline{x}]}{P[X = \overline{x}]}$
Average Treatment on the Treated (ATT)	$E[Y_1 - Y_0 \mid D = 1]$	$-\omega_1^\star(u,x,z)$	$\dfrac{\mathbb{1}[u \leq p(x,z)]}{P[D = 1]}$
Average Treatment on the Untreated (ATU)	$E[Y_1 - Y_0 \mid D = 0]$	$-\omega_1^\star(u,x,z)$	$\dfrac{\mathbb{1}[u > p(x,z)]}{P[D = 0]}$
Local Average Treatment Effect (LATE) for $z \to z'$ given $X = x$, where $p(x,z') > p(x,z)$	$E[Y_1 - Y_0 \mid p(x,z) < U \leq p(x,z'), X = x]$	$-\omega_1^\star(u,x,z)$	$\dfrac{\mathbb{1}[p(x,z) < u \leq p(x,z')]}{p(x,z') - p(x,z)}$

Another commonly considered target parameter is the average treatment on the treated (ATT). Figure 2 plots the average $d = 1$ weights for the ATT and other conventional target parameters discussed in this section using our running numerical illustration. The horizontal axis indexes the weights by unobserved heterogeneity in treatment choice $u \in [0,1]$, with smaller values of u corresponding to individuals that are more likely to choose $D = 1$, c.f. (4). The vertical axis reports the average weights, i.e. $E[\omega_1^\star(u,X,Z)]$ in regions where they are non-zero. For all of these parameters, the weights are symmetric in the sense that $\omega_0^\star(u,x,z) = -\omega_1^\star(u,x,z)$, so we only plot the average weights for $d = 1$. Figure 2 shows that the average weights for the ATT are decreasing in u, indicating that this parameter places more weight on individuals that are more likely to choose $D = 1$. This property can be confirmed by analyzing the corresponding formula in Table 1.

The ATT provides the average decrease in outcomes that would be experienced by the treated group by switching from a regime in which the treatment is optional to a regime that forbids treatment. This counterfactual can be relevant for evaluating optional government programs, such as active labor market programs, since it measures the benefit to those who choose (or are chosen) to receive training (Heckman and Smith, 1998). Similarly, the average treatment on the untreated (ATU) measures the average increase in outcomes that would be experienced by the control group if treatment were made mandatory. This counterfactual would be relevant for evaluating the impact of requiring non-participants to participate in a program.

The ATE, ATT and ATU can all be defined without reference to the choice model (4).[8] Maintaining a choice model allows one to also consider parameters that are defined in terms of choice behavior under actual or counterfactual manipulations of the instrument. An important and well-known example of such a parameter is the local average treatment effect (LATE), which was first studied by Imbens and Angrist (1994). The LATE is informative about the average causal effect for the set of individuals whose choice of D would be altered by a given change in the instrument.

For example, Table 1 shows the weights for the LATE that corresponds to an instrument shift from $Z = z$ to $Z = z'$ with $p(x, z') > p(x, z)$, and conditional on $X = x$. These weights are only non-zero over the region $(p(x, z), p(x, z')]$. Examining (4), one can see that this region corresponds to realizations of U for which an agent with $X = x$ would choose $D = 1$ if assigned $Z = z'$, but would choose $D = 0$ if assigned $Z = z$. Imbens and Angrist (1994) refer to this unobservable subgroup as the (z to z') compliers. In the next section, we show that LATEs are specific examples of the more general concept of a policy relevant treatment effect.

3.3 Policy Relevant Treatment Effects

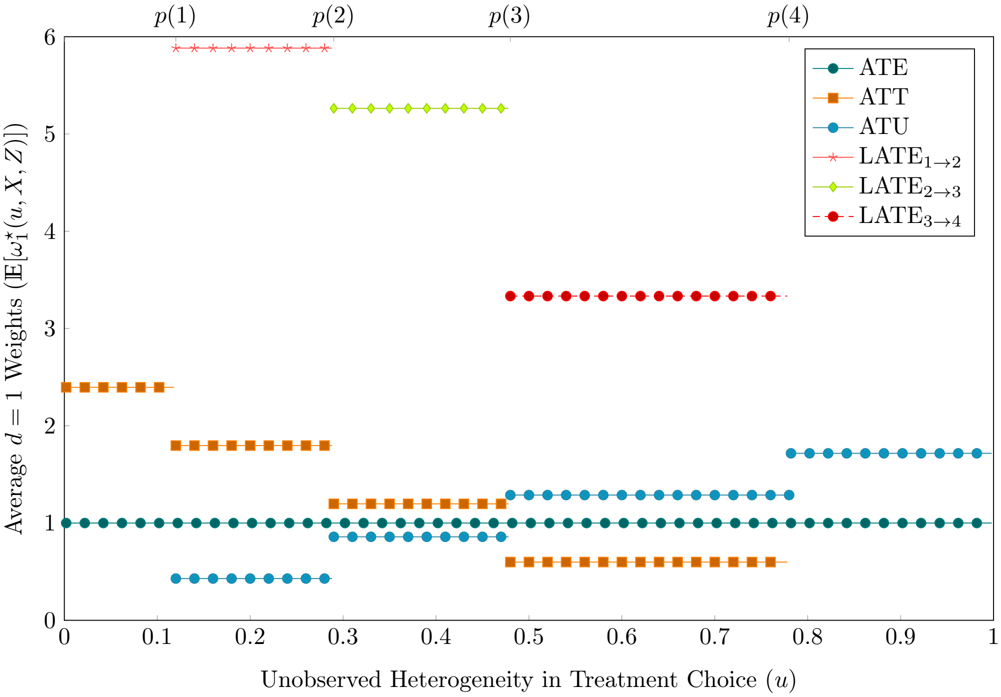
The ATE, ATT and ATU all measure the average effect on outcomes for policy counterfactuals that hypothesize mandating a choice of treatment. The relevance of these parameters, and the policy counterfactuals they address, is dubious when requiring or preventing treatment is conceptually or ethically infeasible. Indeed, many policy discussions are focused on interventions that change the costs or benefits of choosing certain activities, while still allowing individuals to freely select into these activities.

For example, consider the important empirical question of the labor market returns to investing in human capital, say through enrolling in higher education ($D = 1$). The ATE, ATT and ATU all correspond to counterfactuals that conjecture *mandating* enrollment or non-enrollment in higher education. These parameters do not speak to ongoing debates over higher education policy. Instead, these debates are about interventions that influence the decision to enroll in higher education, for example by increasing the availability of colleges, or expanding student loan or tuition subsidies.

Another example, considered in more depth in Mogstad et al. (2017), is the decision

[8] Although, as we will see in Sections 5 and 6, the choice model facilitates thinking about identification of these parameters as an extrapolation problem.

Figure 2: Weights for Conventional Target Parameters in the Numerical Illustration



to own a mosquito net ($D = 1$). This is an important preventative health care measure in many parts of the developing world. Mandating non-ownership—which is implicitly conjectured in the ATE and ATT—is not an interesting counterfactual. The ATU conjectures mandating ownership, which is perhaps conceivable through a policy of free provision, although this would still require full take-up. A more relevant policy intervention would be to provide subsidies to purchase a mosquito net, taking into account the potential benefits of usage and costs of subsidization.[9]

A choice model like (4) provides a framework for considering the effect of a policy intervention that influences (but may not fully determine) choice behavior. We follow Heckman and Vytlacil (1999, 2005) in considering policies that change the propensity score, p, and/or the instrument, Z, but which are assumed to have no impact on the model unobservables, (Y_0, Y_1, U), or the observed covariates X. For example, this assumption requires that a policy that alters the effective price of a mosquito net—modeled here as changing p and/or Z—would have no impact on the latent propensity to buy a mosquito net, U, or on whether an individual would be afflicted by malaria in either treatment state, (Y_0, Y_1). A policy a in this class can be summarized by a pair

[9] See e.g. Dupas and Zwane (2016) for a discussion of various policies that promote access to (and usage of) preventive health products. None of these policies involve mandating ownership or usage of preventive health products.

(p^a, Z^a) consisting of a function p^a that maps (X, Z^a) to $[0, 1]$, and a random variable Z^a that satisfies IV.2. Both the function, p^a, and the joint distribution of (X, Z^a), are assumed to be known or identified.

A policy with these properties generates random variables representing treatment choice and outcomes. Treatment choice under a policy a is given by

$$D^a \equiv \mathbb{1}[U \leq p^a(X, Z^a)]. \tag{9}$$

The outcome of Y that would be observed under policy a is therefore

$$Y^a = D^a Y_1 + (1 - D^a) Y_0. \tag{10}$$

Given two policies, a_1 and a_0, Heckman and Vytlacil (1999, 2005) define the policy relevant treatment effect (PRTE) of a_1 relative to a_0 as

$$\text{PRTE} \equiv \frac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]}, \tag{11}$$

where we assume that $E[D^{a_1}] \neq E[D^{a_0}]$, i.e. that the policy change also changes the overall proportion of individuals who receive treatment.[10]

3.4 Examples of PRTEs

PRTEs can be expressed as target parameters with form (8). The choice of weights, $(\omega_0^\star, \omega_1^\star)$, depends on the policies being compared.[11] Table 2 shows how different policy comparisons translate into different weights by way of three specific examples considered by Carneiro et al. (2011). Each of the examples sets a_1 to be a hypothetical policy, and takes a_0 to be the status quo policy observed in the data, i.e. $(p^{a_0}, Z^{a_0}) = (p, Z)$. The hypothetical policies are: (i) an additive α change in the propensity score, i.e. $p^{a_1} = p + \alpha$; (ii) a proportional $(1 + \alpha)$ change in the propensity score, i.e. $p^{a_1} = (1 + \alpha)p$; and (iii) an additive α shift in the distribution the jth component of Z, i.e. $Z^{a_1} = Z + \alpha e_j$, where e_j is the jth unit vector. The first and second of these represent policies that increase (or decrease) participation in the treatment by a given amount α or a proportional amount $(1 + \alpha)$. The third policy represents the effect of shifting the distribution of an exogenous variable that impacts treatment

[10] The purpose of this assumption is simply to adjust the units of the PRTE to be per net change in treatment participation. If this assumption is questionable, one can alternatively define the PRTE as $E[Y^{a_1}] - E[Y^{a_0}]$, see Heckman and Vytlacil (2001a) or Carneiro, Heckman, and Vytlacil (2010, pp. 380–381).

[11] Note that these weights are identified given the assumption that both p^a and the distribution of (X, Z^a) are known or identified with $Z^a \perp\!\!\!\perp U | X$ for $a = a_0, a_1$.

Table 2: Weights for Policy Relevant Treatment Effects

Target Parameter	Expression	$\omega^\star_1(u,x,z) = -\omega^\star_0(u,x,z)$
Generalized LATE for $U \in [\underline{u}, \overline{u}]$	$E[Y_1 - Y_0 \| U \in [\underline{u}, \overline{u}]]$	$\dfrac{\mathbb{1}[u \in [\underline{u}, \overline{u}]]}{\overline{u} - \underline{u}}$
Policy Relevant Treatment Effect (PRTE) for policy (p^{a_1}, Z^{a_1}) relative to policy (p^{a_0}, Z^{a_0})	$\dfrac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]}$	$\dfrac{P[u \leq p^{a_1}(x, Z^{a_1}) \| X = x] - P[u \leq p^{a_0}(x, Z^{a_0}) \| X = x]}{E[p^{a_1}(X, Z^{a_1})] - E[p^{a_0}(X, Z^{a_0})]}$
Additive PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(x,z) = p(x,z) + \alpha$	$\dfrac{\mathbb{1}[u \leq p(x,z) + \alpha] - \mathbb{1}[u \leq p(x,z)]}{\alpha}$
Proportional PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(x,z) = (1+\alpha)p(x,z)$	$\dfrac{\mathbb{1}[u \leq (1+\alpha)p(x,z)] - \mathbb{1}[u \leq p(x,z)]}{\alpha E[p(X,Z)]}$
PRTE for an additive α shift of the j^{th} component of Z	PRTE with $Z^\star = Z + \alpha e_j$ and $p^\star(x,z) = p(x,z)$	$\dfrac{\mathbb{1}[u \leq p(x, z + \alpha e_j)] - \mathbb{1}[u \leq p(x,z)]}{E[p(X, Z + \alpha e_j)] - E[p(X,Z)]}$

choice, such as a subsidy.

In all of these definitions, α is a quantity that could either be estimated or hypothesized by the researcher. Mogstad et al. (2017) consider PRTEs of type (i), and they estimate the value of α by parametrically extrapolating a demand curve fit off of experimentally varied prices. Since α is interpretable in terms of the change of treatment participation probability, a simpler approach is to just specify a value of α that represents an empirically interesting change in the probability of choosing treatment.

3.5 LATEs are PRTEs

The LATE is a particular example of a PRTE. To see this, suppose for simplicity that there are no covariates X, and consider the PRTE that results from comparing a policy a_1 under which every agent receives $Z = z'$ against a policy a_0 under which every agent receives $Z = z$.[12] Choices under these policies are

$$D^{a_0} \equiv \mathbb{1}[U \leq p(z)] \quad \text{and} \quad D^{a_1} \equiv \mathbb{1}[U \leq p(z')],$$

[12] More formally, let $p^{a_0} = p^{a_1} = p$, and take Z^{a_1} and Z^{a_0} to be deterministically equal to z' and z, respectively.

where $p(z') > p(z)$ are the propensity score values in the observed data. The PRTE for this policy comparison is

$$\frac{E[Y^{a_1} - Y^{a_0}]}{E[D^{a_1} - D^{a_0}]} = \frac{E\left[(D^{a_1} - D^{a_0})(Y_1 - Y_0)\right]}{p(z') - p(z)} = E\left[Y_1 - Y_0 \,\middle|\, p(z) < U \leq p(z')\right], \quad (12)$$

where we used $D^{a_1} - D^{a_0} = \mathbb{1}[p(z) < U \leq p(z')]$. The right-hand side of (12) is precisely the z to z' LATE introduced by Imbens and Angrist (1994).

More generally, Heckman and Vytlacil (2005) define a LATE as $E[Y_1 - Y_0 | U \in [\underline{u}, \overline{u}]]$ for two values $\underline{u}$ and $\overline{u}$. We refer to this parameter as a *counterfactual* LATE in order to distinguish it from a LATE for which $\underline{u}$ and $\overline{u}$ are given by values of the observed propensity score. The weights for a counterfactual LATE are shown in Table 2. They are equally weighted over $[\underline{u}, \overline{u}]$, zero elsewhere, and scaled to integrate to 1.

3.6 Extrapolating LATEs

Viewing the LATE as a specific example of a more general class of parameters is useful for thinking about parameters that represent subpopulations other than just the compliers under the observed instrument. For example, suppose that a researcher wants to perform a sensitivity analysis to investigate the robustness of the z to z' LATE to an expansion (or contraction) of the complier subpopulation. For this purpose, we define right- and left-hand α-extrapolations of the z to z' LATE as

$$\begin{aligned} &\text{LATE}^{+}_{z \to z'}(\alpha) \equiv E\left[Y_1 - Y_0 \mid p(z) < U \leq p(z') + \alpha\right] \\ \text{and} \quad &\text{LATE}^{-}_{z \to z'}(\alpha) \equiv E\left[Y_1 - Y_0 \mid p(z) - \alpha < U \leq p(z')\right]. \end{aligned} \quad (13)$$

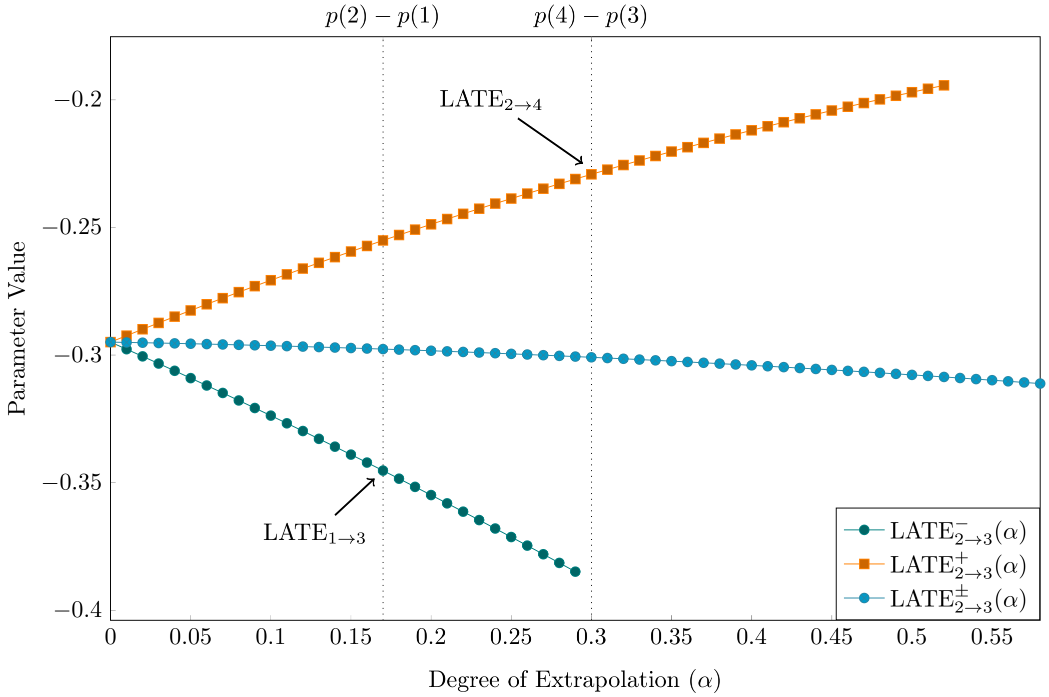
Similarly, we define a two-sided α-extrapolation as

$$\text{LATE}^{\pm}_{z \to z'}(\alpha) \equiv E\left[Y_1 - Y_0 \,\middle|\, p(z) - \frac{\alpha}{2} < U \leq p(z') + \frac{\alpha}{2}\right]. \quad (14)$$

These parameters are defined over subgroups that take the z to z' complier group and expand it by α, either to the left, right, or split between both sides. One could also allow $\alpha < 0$ in (13) and (14), in which case the parameters would be *interpolated* LATEs.

Imbens and Angrist (1994) showed that the z to z' LATE is nonparametrically point identified for any observed z and z', as long as $p(z') > p(z)$. This result has produced a focus on these types of LATEs as parameters of interest. Since these LATEs only reflect causal effects for z to z' compliers, their external validity (or generalizability) can be limited (Imbens, 2010). Some authors have criticized the practice of focusing

Figure 3: Extrapolated LATEs in the Numerical Illustration



on parameters with limited external validity, see e.g. Heckman (1996, 1997, 2010). Analyzing extrapolated LATEs allows one to bridge these two viewpoints, since it provides a precise way to gauge this lack of external validity. In particular, the extent to which a given LATE is externally valid depends on how different it can be from the extrapolated LATEs as α increases. As $\alpha \to 0$, an extrapolated z to z' LATE reduces back to the usual, point identified z to z' LATE.

Figure 3 illustrates this point for the $Z = 2$ to $Z = 3$ LATE in our running numerical example. The figure contains the values of the left-hand, right-hand, and two-sided extrapolations of this LATE as functions of the size of the extrapolation, α. As α increases from 0, these parameters cover increasingly large subpopulations. The graph shows that the $Z = 2$ to $Z = 3$ LATE is sensitive to extrapolation to either the left or right, but relatively insensitive when extrapolating on both sides simultaneously. For certain values of α, an extrapolated LATE can reduce to another ordinary LATE. For example, when $\alpha = p(4) - p(3)$, the right-hand extrapolated $Z = 2$ to $Z = 3$ LATE is equal to the usual $Z = 2$ to $Z = 4$ LATE, as indicated in Figure 3.

4 When is the Target Parameter Point Identified?

Once the researcher has defined the target parameter, the next step is to consider its identification. In this section, we consider two commonly discussed settings in which the target parameter is point identified without any additional assumptions.

4.1 When the Target Parameter is a LATE

Imbens and Angrist (1994) showed that under Assumptions IV their monotonicity condition (which, again, is equivalent to Assumptions IV and (4)), the z to z' LATE, conditional on $X = x$, is point identified by the Wald estimand, i.e.

$$\begin{aligned} & E\left[Y_1 - Y_0 | p(x,z) < U \leq p(x,z'), X = x\right] \\ & \quad = \frac{E[Y|X=x, Z=z'] - E[Y|X=x, Z=z]}{E[D|X=x, Z=z'] - E[D|X=x, Z=z]}. \end{aligned}$$

The LATE may be an interesting target parameter if the observed instrument variation from z to z' represents an intervention or policy change. For example, Angrist and Krueger (1991) report estimates of a LATE for which D is attaining an additional year of schooling, Y is a measure of future earnings, and the shift from z to z' represents the impact of a compulsory schooling law. This parameter would clearly be useful for evaluating how compulsory schooling laws affect labor market outcomes through their impact on raising educational attainment.

However, in many other situations, the observed variation in the instrument might be distinctly different than the variation relevant for the researcher's policy question. In such cases, the LATE is not a relevant target parameter. Consider, for example, the large body of empirical research that has examined the relationship between family size and observable child outcomes, such as educational attainment. Recent studies that use IV methods to address the possible endogeneity of family size, such as Black, Devereux, and Salvanes (2005), tend to conclude that family size has a small causal effect on child outcomes. Two instruments commonly used in these studies are twin births and the sex composition of prior births. The parameter estimates they report can be interpreted as reflecting the LATEs for these instruments.

When interpreting the estimated LATEs, it is natural to consider whether variation in these instruments can be used to address a counterfactual with interesting policy implications. An obvious concern in doing so is that families that would only have another child due to a twin birth, or due to the sex composition of their previous children, likely differ in unobservable ways from other families. As a consequence, families whose fertility decisions would be affected by these instruments may be dissimilar to families whose decisions would be affected by a proposed tax or transfer policy. For evaluating such a policy, LATEs for either of these instruments are not relevant target parameters. Arguing along these lines, Brinch et al. (2017) revisit the analysis of Black et al. (2005) using an extrapolation approach discussed in Section 6. Their findings suggest that there is a great deal of heterogeneity in the causal effect of family size on child outcomes. Their results warrant caution in using LATEs for twin or sex composition instruments as parameters for informing policy debates.

4.2 When There is Sufficient Variation in the Instrument

Heckman and Vytlacil (1999, 2001c) showed that if the random variable $P = p(X, Z)$ is continuously distributed, conditional on $X = x$, then under some regularity conditions the MTE is point identified for any $\widetilde{u}$ in the interior of its support. To see this, note that in general it can be shown by using (4) and IV.2 that

$$\begin{aligned} E\left[YD \mid p(x, Z) = u, X = x\right] &= \int_0^u m_1(u', x)\, du' \\ \text{and} \quad E\left[Y(1-D) \mid p(x, Z) = u, X = x\right] &= \int_u^1 m_0(u', x)\, du'. \end{aligned} \tag{15}$$

As a consequence, if the objects on the left-hand sides of (15) can be differentiated at $u = \widetilde{u}$, then

$$\begin{aligned} \frac{\partial}{\partial u} E\left[YD \mid p(x, Z) = u, X = x\right] \Big|_{u=\widetilde{u}} &= m_1(\widetilde{u}, x), \\ \frac{\partial}{\partial u} E\left[Y(1-D) \mid p(x, Z) = u, X = x\right] \Big|_{u=\widetilde{u}} &= -m_0(\widetilde{u}, x), \\ \text{and hence} \quad \frac{\partial}{\partial u} E\left[Y \mid p(x, Z) = u, X = x\right] \Big|_{u=\widetilde{u}} &= m_1(\widetilde{u}, x) - m_0(\widetilde{u}, x), \end{aligned} \tag{16}$$

so that the MTRs and MTE at $(\widetilde{u}, x)$ are point identified. The left-hand side of (16) is what Heckman and Vytlacil (1999, 2001c) describe as the local IV estimand. A consequence of their argument is that any target parameter is point identified if it has weights $(\omega_0^\star, \omega_1^\star)$ that are non-zero only for values of $(\widetilde{u}, x)$ for which $\widetilde{u}$ lies in the interior of the support of P, conditional on x. Viewed in reverse, a given target parameter is point identified if the distribution of P, given $X = x$, is continuous and exhibits sufficient variation to cover the support of $(\omega_0^\star, \omega_1^\star)$ for every x.

This support condition severely limits the types of target parameters that are point identified without additional assumptions. It requires a continuous instrument, for if Z is discrete, then the distribution of $P \equiv p(X, Z)$, conditional on X, will also be discrete, so that differentiation in (16) is not possible. Requiring an instrument to be continuous already eliminates perhaps the majority of instruments used in modern applications of IV methods. Moreover, even if the instrument is continuous, only target parameters with support contained within the observed support of P (conditional on X) can be non-parametrically point identified. PRTEs for policies that involve extrapolating beyond the currently available support will not be point identified without additional assumptions.[13] In many cases, however, these are precisely the types of policies that are likely to be relevant to decision makers.

For example, an important and largely unanswered question for developing countries is how to design cost effective policies that promote access to (and usage of) preventive health products. To analyze this question, Mogstad et al. (2017) use data from an experiment conducted in Kenya by Dupas (2014) in which the price for a preventative health product was randomly assigned. They view different subsidy regimes as different PRTEs and compare increases in usage to the cost of subsidization. For example, they estimate the PRTE that compares a policy of free provision to a policy under which all the product is offered at any given price. To do so, they use the randomly assigned prices as a (discrete) instrument for purchasing the health product. Many of the PRTEs they consider do not correspond to the variation in prices that were observed by the experiment. These PRTEs are not point identified without additional assumptions, but as Mogstad et al. (2017) show, informative bounds can still be constructed by using the method described in the next section.

5 A General Framework for Inference about Treatment Effects

In the previous section, we discussed two cases in which the variation in the treatment that is induced by the instrument can be used to point identify the target parameter without additional assumptions. In many other cases, answering the policy question of interest requires extrapolation from the individuals whose treatment choice is affected by the available instrument to the individuals whose treatment choice would be affected by the policy. In this section, we discuss how to use the general framework proposed by Mogstad et al. (2017, "MST") to conduct this extrapolation.

[13] The marginal PRTE considered by Carneiro et al. (2010, 2011) provides a possible exception to this statement. This parameter can be viewed as the PRTE that results from contrasting the status quo, (p, Z), to a marginal change to the status quo. This marginal change is formally defined as an infinitesimally small change, so it is arguably not appropriate to view these parameters as conjecturing a significant departure from existing policies.

5.1 What We Know: IV-Like Estimands

The starting point for MST is the observation that a rich class of identified quantities can also be written in the same form ((8)) as the target parameter, $\beta^\star$. For example, consider the IV estimand that results from using Z as an instrument for D in a linear instrumental variables regression that includes a constant term, but which does not include any other covariates X. Assuming $\text{Cov}(D, Z) \neq 0$, this estimand is given by

$$\beta_{\text{IV}} \equiv \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}. \tag{17}$$

Heckman and Vytlacil (2005) showed that β_{IV} can be written as

$$\beta_{\text{IV}} = \int_0^1 [m_1(u, X) - m_0(u, X)]\ \omega_{\text{IV}}(u, X, Z)\, du, \tag{18}$$

where ω_{IV} is an identified weighting function. The similarity between (18) and (8) suggests that β_{IV} carries some useful information about the possible values of $\beta^\star$.

MST show that, more generally, any cross moment of Y with a known or identified function of (D, X, Z) can also be expressed as the weighted sum of the two MTR functions, m_0 and m_1. To be more precise, let s be a known or identified measurable function of (d, x, z) and define $\beta_s \equiv E[s(D, X, Z)Y]$. MST call the function s an IV–like specification, and they call the quantity β_s that s generates an IV–like estimand. Proposition 1 of MST shows that for any s,

$$\begin{aligned} \beta_s = E\left[\int_0^1 m_0(u, X)\omega_{0s}(u, X, Z)\, du\right] &+ E\left[\int_0^1 m_1(u, X)\omega_{1s}(u, X, Z)\, du\right], \\ \text{where}\ \ \omega_{0s}(u, X, Z) &\equiv s(0, X, Z)\mathbb{1}[u > p(X, Z)] \\ \text{and}\ \ \omega_{1s}(u, X, Z) &\equiv s(1, X, Z)\mathbb{1}[u \leq p(X, Z)]. \end{aligned} \tag{19}$$

The weights in (19) can be shown to reduce to the weights for β_{IV} derived by Heckman and Vytlacil (2005) by taking

$$s(d, x, z) = \frac{z - E[Z]}{\text{Cov}(D, Z)}, \tag{20}$$

which is an identified function of D, X (both trivially), and Z. However, the expression in (19) applies more broadly to include any well-defined weighted linear IV estimand that uses some function of (D, X, Z) as included and excluded instruments for a set of endogenous variables also constructed from (D, X, Z).[14] Deriving these weights is

[14] The phrases "included" and "excluded" instrument are meant in the sense typically introduced in

Table 3: Common IV–Like Estimands

Estimand	$\boldsymbol{\beta_s}$	$\boldsymbol{s(D, X, Z)}$	**Notes**
Wald (z to z')	$\dfrac{E[Y\|Z = z'] - E[Y\|Z = z]}{E[D\|Z = z'] - E[D\|Z = z]}$	$\dfrac{\frac{\mathbb{1}[Z=z']}{P[Z=z']} - \frac{\mathbb{1}[Z=z]}{P[Z=z]}}{E[D\|Z = z'] - E[D\|Z = z]}$	$P[Z = z'], P[Z = z] \neq 0$ and $E[D\|Z = z']$ $\neq E[D\|Z = z]$
IV slope	$\dfrac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}$	$\dfrac{Z - E[Z]}{\text{Cov}(D, Z)}$	Z scalar
IV (jth component)	$e_j' E[\widetilde{Z}\widetilde{X}']^{-1} E[\widetilde{Z}Y]$	$e_j' E[\widetilde{Z}\widetilde{X}']^{-1} \widetilde{Z}$	$\widetilde{X} \equiv [1, D, X']'$ $\widetilde{Z} \equiv [1, Z, X']'$ Z scalar e_j the jth unit vector
OLS slope	$\dfrac{\text{Cov}(Y, D)}{\text{Var}(D)}$	$\dfrac{D - E[D]}{\text{Var}(D)}$	—
OLS (jth component)	$e_j' E[\widetilde{X}\widetilde{X}']^{-1} E[\widetilde{X}Y]$	$e_j' E[\widetilde{X}\widetilde{X}']^{-1} \widetilde{X}$	$\widetilde{X} \equiv [1, D, X']'$ e_j the jth unit vector
TSLS (jth component)	$e_j' \left(\Pi E[\widetilde{Z}\widetilde{X}']\right)^{-1} \left(\Pi E[\widetilde{Z}Y]\right)$	$e_j' (\Pi E[\widetilde{Z}\widetilde{X}'])^{-1} \Pi \widetilde{Z}$	$\Pi \equiv E[\widetilde{X}\widetilde{Z}'] E[\widetilde{Z}\widetilde{Z}']^{-1}$ Included variables $\widetilde{X}$ Instruments $\widetilde{Z}$ e_j the jth unit vector

a matter of specifying the appropriate IV–like specification, s. Table 3 lists the IV–like specifications that generate several common IV–like estimands, such as the Wald estimand and the estimand corresponding to the two-stage least squares estimator.

5.2 From What We Know to What We Want

IV–like estimands are features of the observable data. In general, IV–like estimands are not equal to the target parameter, and so are not themselves objects of interest. On the other hand, equation (19) shows that any IV–like estimand is a weighted average of the underlying MTR functions. This implies that only some MTR functions are consistent with a given value of an IV–like estimand. Consequently, only some values of the target parameter, $\beta^\star$, are consistent with a given IV–like estimand. In this section, we show how to utilize this intuition to construct bounds on $\beta^\star$.

Let $\mathcal{S}$ denote some collection of IV–like specifications s chosen by the researcher.

textbook treatments of the linear IV model without heterogeneity.

Corresponding to each $s \in \mathcal{S}$ is an IV–like estimand, $\beta_s \equiv E[s(D, X, Z)Y]$. We assume that the researcher has restricted the pair of MTR functions $m \equiv (m_0, m_1)$ to lie in some admissible set, $\mathcal{M}$. The admissible set incorporates any a priori assumptions that the researcher wishes to maintain about the MTR functions, such as parametric or shape restrictions. Our goal is to characterize values of the target parameter $\beta^\star$ that could be generated by MTR functions that are elements of $\mathcal{M}$ and which also deliver the collection of identified IV–estimands $\{\beta_s : s \in \mathcal{S}\}$ through (19).

To do this, it is helpful to view the weighted integrals for the target parameter, (8), and the IV–like estimands, (19), as functions of m. Specifically, for the target parameter we define the function

$$\Gamma^\star(m) \equiv E\left[\int_0^1 m_0(u, X)\omega_0^\star(u, X, Z)du\right] + E\left[\int_0^1 m_1(u, X)\omega_1^\star(u, X, Z)du\right], \tag{21}$$

and for any IV–like specification s we define the function

$$\Gamma_s(m) \equiv E\left[\int_0^1 m_0(u, X)\omega_{0s}(u, X, Z)\, du\right] + E\left[\int_0^1 m_1(u, X)\omega_{1s}(u, X, Z)\, du\right]. \tag{22}$$

Now, suppose that the data were generated according to (1) and (4) under Assumptions IV with MTR pair $m \in \mathcal{M}$. Then m must satisfy $\Gamma_s(m) = \beta_s$ for every $s \in \mathcal{S}$. That is, m must lie in the set

$$\mathcal{M}_\mathcal{S} \equiv \{m \in \mathcal{M} : \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\}. \tag{23}$$

This in turn implies that $\beta^\star$ must belong to the set

$$\mathcal{B}^\star_\mathcal{S} \equiv \{b \in \mathbb{R} : b = \Gamma^\star(m) \text{ for some } m \in \mathcal{M}_\mathcal{S}\}. \tag{24}$$

Intuitively, $\mathcal{B}^\star_\mathcal{S}$ is the set of values for the target parameter that could have been generated by MTR functions that are consistent with both the assumptions of the model and the values of the IV–like estimands $\{\beta_s : s \in \mathcal{S}\}$ that were observed in the data. Given knowledge of the distribution of observables, $\mathcal{B}^\star_\mathcal{S}$ could be determined by checking for a candidate value b whether there exists an $m \in \mathcal{M}$ such that $\Gamma^\star(m) = b$ and $\Gamma_s(m) = \beta_s$ for all $s \in \mathcal{S}$. If such an m exists, then $b \in \mathcal{B}^\star_\mathcal{S}$; otherwise $b \notin \mathcal{B}^\star_\mathcal{S}$. Under weak conditions on $\mathcal{M}$, $\mathcal{B}^\star_\mathcal{S}$ will be a closed interval, say $[\underline{\beta}^\star, \overline{\beta}^\star]$. In this case, the process of characterizing $\mathcal{B}^\star_\mathcal{S}$ can be simplified to the task of solving two optimization problems:

$$\begin{aligned} &\underline{\beta}^\star \equiv \inf_{m \in \mathcal{M}} \Gamma^\star(m) \quad \text{subject to} \quad \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S} \\ \text{and} \qquad &\overline{\beta}^\star \equiv \sup_{m \in \mathcal{M}} \Gamma^\star(m) \quad \text{subject to} \quad \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}. \end{aligned} \tag{25}$$

5.3 Computing the Bounds

Both $\Gamma^\star$ and Γ_s are linear functions of m. This endows the optimization problems (25) with a great deal of structure that facilitates the speed and reliability of solving these problems. However, two computational obstacles remain. First, the variables of optimization in (25) are infinite dimensional. Second, (25) could be difficult to solve unless the admissible set $\mathcal{M}$ has enough structure.

MST show that both problems can be solved by replacing $\mathcal{M}$ with a finite dimensional linear basis. To see how this works, suppose that for every $m \equiv (m_0, m_1) \in \mathcal{M}$, there exists a finite dimensional vector $\theta \equiv (\theta_0, \theta_1) \in \mathbb{R}^{K_0+K_1}$ such that

$$m_d(u, x) = \sum_{k=0}^{K_d} \theta_{dk} b_{dk}(u, x) \quad \text{for } d = 0, 1, \tag{26}$$

where $b_{dk}(u, x)$ are known basis functions. Substituting (26) into the definition of $\Gamma^\star(m)$, we have

$$\begin{aligned} \Gamma^\star(m) &= \sum_{d \in \{0,1\}} \sum_{k=0}^{K_d} \theta_{dk} E\left[\int_0^1 b_{dk}(u, X)\omega_d^\star(u, X, Z)\, du\right] \\ &\equiv \sum_{d \in \{0,1\}} \sum_{k=0}^{K_d} \theta_{dk}\gamma_{dk}^\star \quad \text{where } \gamma_{dk}^\star \equiv E\left[\int_0^1 b_{dk}(u, X)\omega_d^\star(u, X, Z)\, du\right]. \end{aligned} \tag{27}$$

The $\gamma_{dk}^\star$ terms in (27) are identified population quantities that depend on the known basis functions, b_{dk}, and the known (or identified) weighting functions, $\omega_d^\star$, but which do not depend on θ. Imposing (26) therefore turns the objective of (25) into a linear function of the finite dimensional parameter, θ. Similarly, (26) implies that

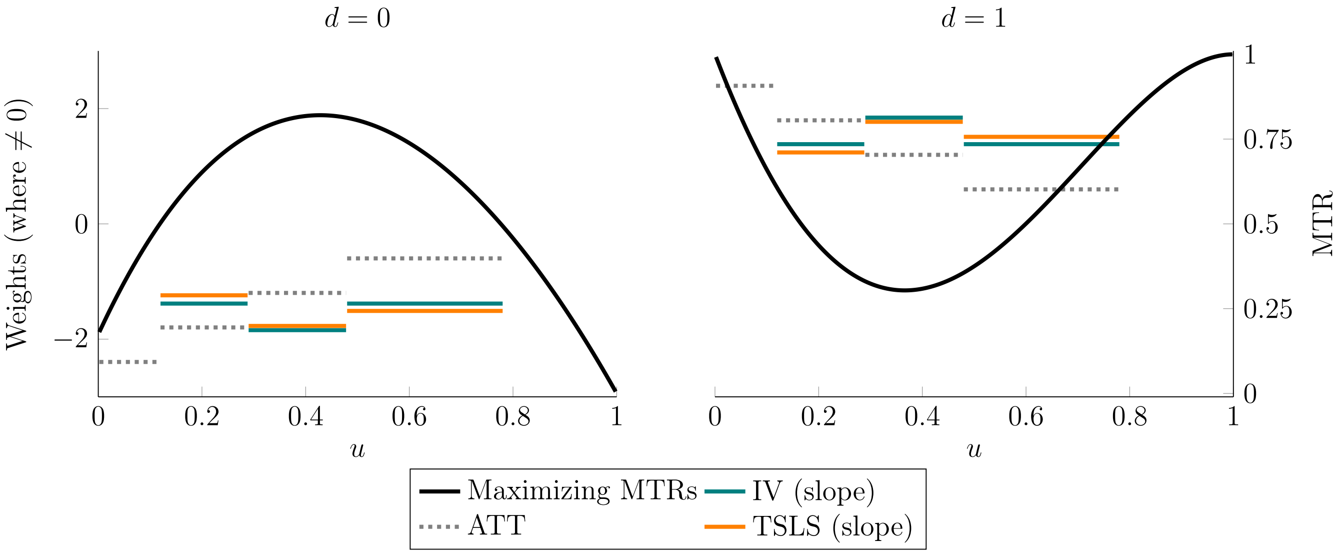
$$\Gamma_s(m) = \sum_{d \in \{0,1\}} \sum_{k=0}^{K_d} \theta_{dk}\gamma_{sdk} \quad \text{where } \gamma_{sdk} \equiv E\left[\int_0^1 b_{dk}(u, X)\omega_{ds}(u, X, Z)\, du\right],$$

for every $s \in \mathcal{S}$, so that the constraints for the IV–like specifications in (25) are also linear in θ.

Under (26), each $m \in \mathcal{M}$ is parameterized by a finite dimensional θ. In analogy

Figure 4: Fourth Degree Polynomial Bounds ($K_0 = K_1 = 4$) on the ATT

Bounds: [-0.494,-0.073] – Shown at Upper Bound



to $\mathcal{M}$, one can specify an admissible set Θ to which θ is restricted to belong. For computation, it is advantageous to specify Θ to be closed convex polyhedron, i.e. a set determined by a finite collection of linear inequalities. In this case, the maximization problem in (25) reduces to the linear program

$$\overline{\beta}^{\star} = \max_{\theta \in \Theta} \sum_{d \in \{0,1\}} \sum_{k=0}^{K_d} \gamma^{\star}_{dk} \theta_{dk} \quad \text{subject to} \quad \sum_{d \in \{0,1\}} \sum_{k=0}^{K_d} \gamma_{sdk} \theta_{dk} = \beta_s \text{ for all } s \in \mathcal{S}, \quad (28)$$

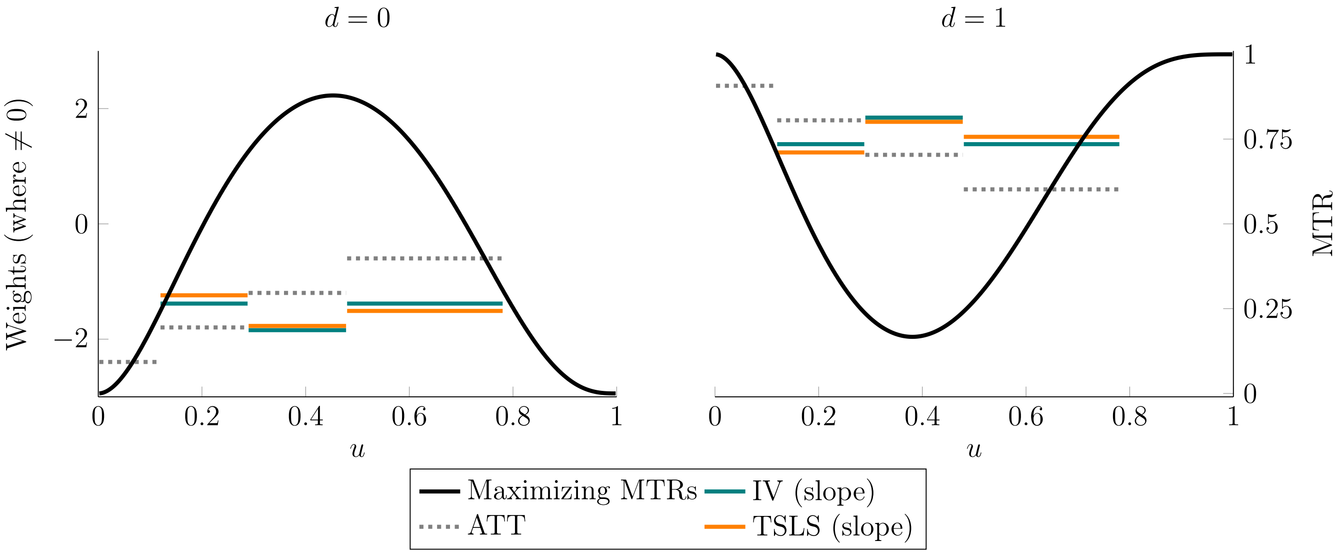
and similarly for the minimization problem. Linear programs like these can be solved reliably and are routinely used in empirical work using quantile regressions, see e.g. Buchinsky (1994), Abadie, Angrist, and Imbens (2002) and Koenker (2005). We view the computational benefits afforded by linear programming as sufficiently important to restrict ourselves to this case in the following.

5.4 Parametric and Nonparametric Bounds

The interpretation of (26) and Θ depends on the choice of basis functions. For example, suppose for simplicity that there are no covariates X, and that the basis functions are chosen to be polynomials, i.e. $b_{dk}(u) = u^{k-1}$ for $k = 1, \ldots, K_d$. With small values of K_d, this choice imposes a strong parametric restriction on the collection of admissible MTR pairs. The restriction becomes weaker for larger values of K_d, since larger values

Figure 5: Ninth Degree Polynomial Bounds ($K_0 = K_1 = 9$) on the ATT

Bounds: [-0.537,0.049] – Shown at Upper Bound

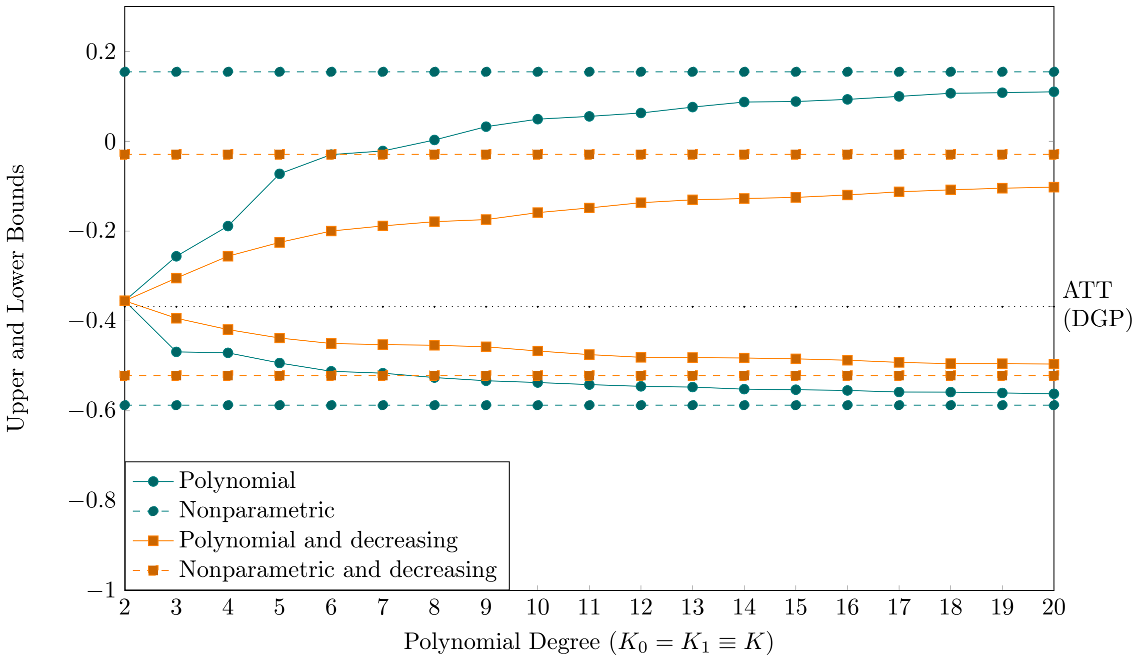


of K_d add more variables of optimization to (25). We view this as a natural and attractive property, since it allows a researcher to transparently trade off the strength of their assumptions with the strength of their conclusions.

Figures 4 and 5 demonstrate this property in our running numerical illustration. These graphs have two vertical axes, with the left-hand axis measuring the weight functions for the target parameter and IV–like estimands, and the right-hand axis measuring MTR functions. The graphs are split into two panels, with the left panel displaying weights and an MTR function for $d = 0$, and the right panel displaying these objects for $d = 1$. Figure 4 is generated by solving the maximization problem (28) when the target parameter is the ATT, the basis functions are fourth degree polynomials (so $K_0 = K_1 = 4$), and two IV–like estimands are included in $\mathcal{S}$. The two IV–like estimands are the slope terms for the IV estimand that uses Z as an instrument for D, and a TSLS estimand that uses $\{\mathbb{1}[Z = z]\}_{z=1}^{4}$ as instruments for D. In this example, these two IV–like estimands yield similar (although not identical) weights, shown by the colored curves in Figure 4.

The black curves in Figure 4 represent choices of m_0 and m_1 that yield the upper bound on $\beta^\star$, which we are taking here to be the ATT. These choices are not unique. What is unique is the attained upper bound of .049 for $\beta^\star$, which is indicated in the header of Figure 4 along with the analogous lower bound. This upper bound is constrained by the requirement that IV–like estimands generated by this black curve are

Figure 6: Bounds on the ATT for Different K



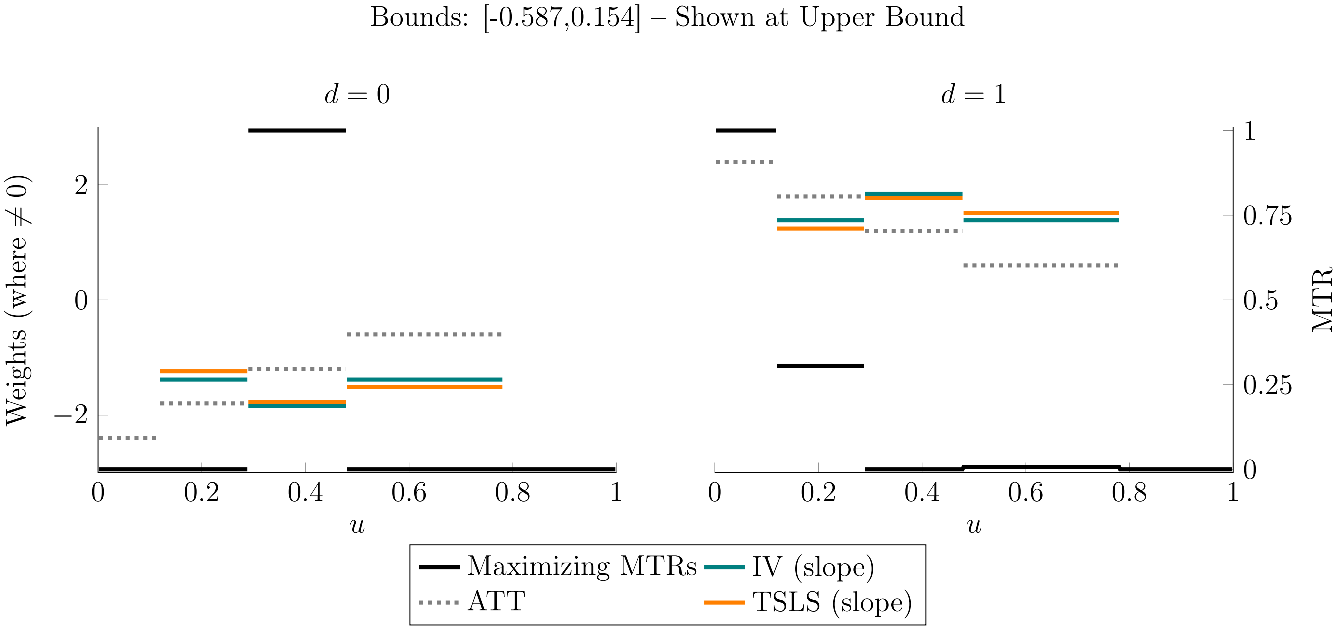
equal to the values observed in the data. Visually, this corresponds to a requirement that the integrals of the products of the black and colored functions attain a given value. The upper bound on the ATT is the largest that the integral of the product of the black and gray dotted curves could be while ensuring that this requirement is satisfied.

Figure 5 shows the result from the same problem with $K_0 = K_1 = 9$, so that the basis functions are ninth degree polynomials. The bounds necessarily become wider than in Figure 4, which reflects the fact that the set of fourth degree polynomials is a subset of the set of ninth degree polynomials. Figure 6 demonstrates this phenomenon for a large number of values of polynomial degrees, K. The upper and lower bounds for the current problem are shown as a solid green line with circle marks. Intuitively, by increasing the degree of the polynomial one is allowing for more wiggly MTR functions that can adjust to become larger more quickly in regions where the target parameter weights are most important.

For researchers who wish to remain fully nonparametric, MST show that (26) can also be used to recover *exact* nonparametric bounds by specifying the basis functions as segments of a constant spline with knots chosen at particular u values.[15] Figure

[15] They also provide a statistical inference framework in which the dimension of θ enters into the asymptotics as in sieve estimation (Chen, 2007).

Figure 7: Exact Nonparametric Bounds on the ATT



7 shows the impact of replacing the polynomial basis in Figures 4 and 5 with this constant spline basis. The bounds widen—as they must—since they are computed under strictly fewer restrictions than when a polynomial basis is maintained. Figure 6 shows that as K increases, the bounds using the polynomial basis approach the fully nonparametric bounds, depicted there as constant dotted green lines with circle marks.

5.5 Nonparametric Shape Restrictions

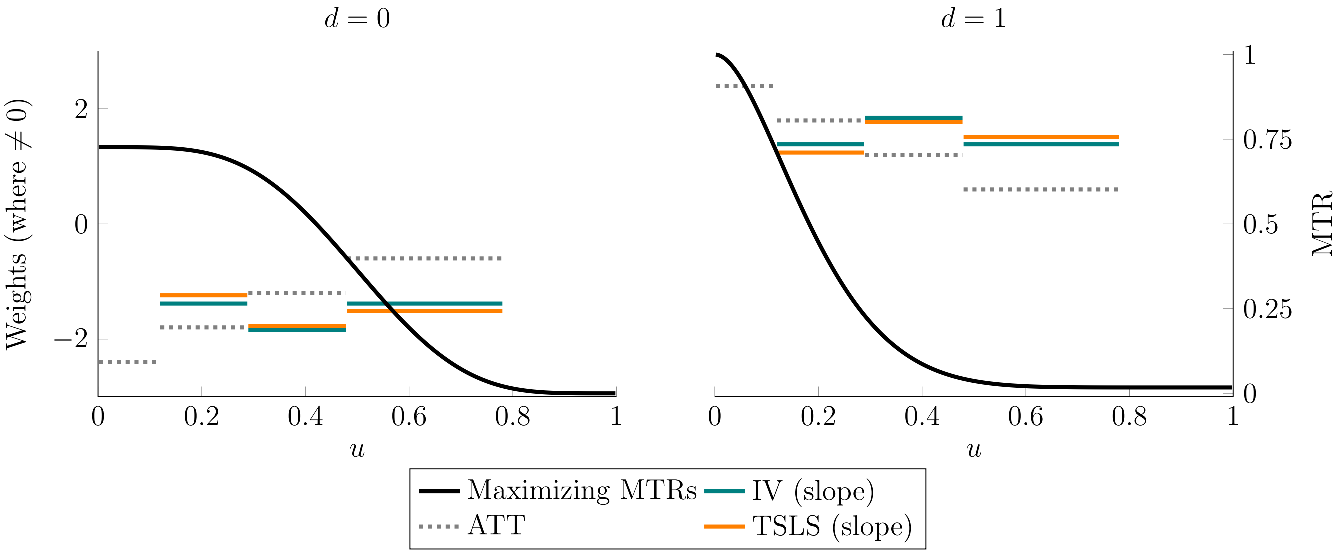
One attractive aspect of the general framework is that it allows researchers to easily incorporate nonparametric shape restrictions into their specification of the MTR functions. These restrictions can be imposed either on the MTR functions $m = (m_0, m_1)$ or directly on the MTE function $m_1 - m_0$. For example, in some applications one may be willing to assume that $m(\cdot, x)$ is weakly decreasing for every x. This restriction would reflect an assumption that those more likely to select into treatment (those with small realizations of U) are also more likely to have larger gains from treatment. This is similar to the monotone treatment selection assumption of Manski and Pepper (2000).[16]

Figure 8 demonstrates the effect of imposing the assumption that the MTR functions are decreasing in our running numerical example. In particular, the figure shows

[16] See Chernozhukov, Newey, and Santos (2015) for a discussion of various shape restrictions implied by economic theory in several empirical applications.

Figure 8: Ninth Degree Decreasing Polynomial Bounds on the ATT

Bounds: [-0.467,-0.159] – Shown at Upper Bound



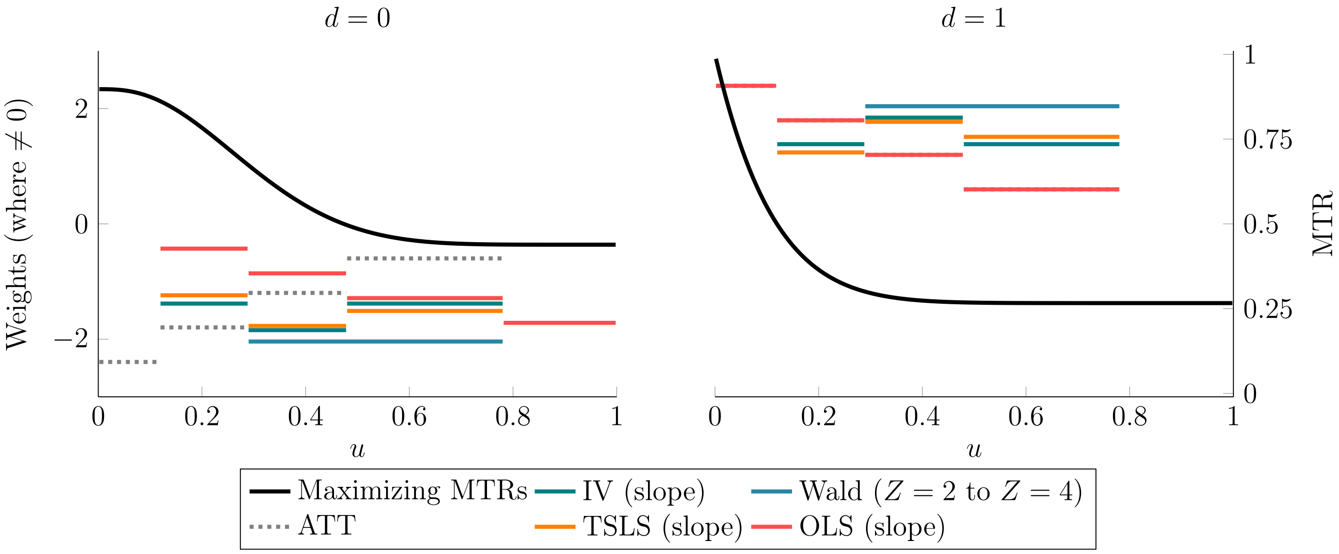
the result of using a ninth degree polynomial basis, as in Figure 5, but further restricting the admissible MTR pairs so that both m_0 and m_1 must be decreasing in u, as in Figure 1. The basis for this assumption would be an a priori belief in a selection story, for example the one we described in which individuals who are more likely to purchase mosquito nets would also be more likely to be afflicted by malaria due to variation in their personal immunity. The additional monotonicity restriction mechanically tightens the bounds by imposing an additional constraint on the optimization problem (28). In particular, it ensures that the maximizing MTR functions shown in Figure 5 are no longer feasible, since neither one is monotonically decreasing.

Figure 6 illustrates the impact of enforcing monotonicity for different order polynomials. Monotonicity can also be imposed when using the fully nonparametric (constant spline) bounds. As expected, the polynomial monotone bounds are always narrower than the nonparametric monotone bounds, with the difference disappearing as the degree of the polynomial increases. The figure shows that shape restrictions such as monotonicity—which are inherently nonparametric—can contain a great deal of identifying content. Indeed, the bounds for nonparametric but decreasing MTRs are roughly the same as when allowing for MTRs that are non-monotone sixth degree polynomials.

Another type of nonparametric shape restriction that is often used is separability

Figure 9: Ninth Degree Decreasing Polynomial Bounds with More IV–Like Estimands

Bounds: [-0.414,-0.275] – Shown at Upper Bound



between the observed (X) and unobserved (U) components, i.e. the assumption that

$$m_d(u, x) = m_d^U(u) + m_d^X(x) \quad \text{for } d = 0, 1, \tag{29}$$

for some functions m_d^U and m_d^X. Separability implies that the slopes of the MTR functions with respect to u do not vary with x. We discuss separability more fully in Section 6.2. Maintaining combinations of assumptions simultaneously (e.g. both monotonicity and separability) is simply a matter of imposing both restrictions on $\mathcal{M}$ at the same time.

In practice, these shape restrictions are imposed through the specification of Θ for a given finite basis (26). The restrictions involved in ensuring that a given θ generates an MTR pair with a particular set of shape properties depends on the choice of basis. As discussed in MST, the Bernstein polynomial basis is particularly attractive in this regard, since many common shape restrictions can be phrased as linear constraints on the components of θ. For a nonparametric analysis, the constant spline basis discussed in the previous section is also easy to force into particular shapes by imposing linear constraints on θ. The linearity involved in these constraints is computationally helpful, since it ensures that (25) remains a linear program.

5.6 Choosing IV–Like Specifications

The set $\mathcal{S}$ of IV–like specifications is chosen by the researcher. Intuitively, one can think of $\mathcal{S}$ as the set of information from the data that the analyst uses to discipline their inference. Examining (25) shows that including more specifications in $\mathcal{S}$ mechanically reduces the identified set $[\underline{\beta}^\star, \overline{\beta}^\star]$ for the target parameter, $\beta^\star$. For example, in Figure 9, we recompute the bounds in Figure 8 after including two more IV–like estimands in $\mathcal{S}$: the OLS estimand, and the $Z = 2$ to $Z = 4$ Wald estimand. The effect is a substantial decrease in the width of the bounds. MST show how to choose $\mathcal{S}$ systematically so as to exhaust all of the information contained in the conditional mean of Y for any given choice of the admissible set $\mathcal{M}$.

For the purposes of identification, the only drawback to expanding $\mathcal{S}$ is increased computational difficulty. When considering statistical inference, the situation becomes more delicate, as including IV–like specifications with low content and large noise will be unhelpful. A natural starting point is to choose IV–like specifications that generate the estimands one would ordinarily be interested in when not being concerned about endogeneity or unobserved heterogeneity. For example, one set of s would be the vector of ordinary least squares (OLS) estimands, another would be the vector of IV estimands, and a third could be a vector of two-stage least squares (TSLS) estimands from including an additional instrument.

While this potentially leaves some information on the table, it has the interpretative benefit of being a departure from a well-understood baseline. An attractive property of this approach is that, by construction, any feasible value of the target parameter must also be consistent with these baseline IV–like estimands. This allows one to follow the advice of Imbens (2010, pp. 414–415), who recommends reporting both a standard LATE, as well as parameters with higher external validity, while maintaining a clear distinction between the assumptions that drive their identification. As long as one includes a Wald estimand corresponding to such a LATE in the set of IV–like specifications, all MTR pairs in $\mathcal{M}_{\mathcal{S}}$ and all potential values of the target parameter, $\mathcal{B}^\star_{\mathcal{S}}$, will necessarily be consistent with this LATE.[17]

5.7 Determinants of the Width of the Bounds

The width of the bounds is determined by three factors: The degree of extrapolation required to evaluate the target parameter, the strength of the a priori assumptions that the analyst maintains, and the information set of IV–like estimands, $\mathcal{S}$. The

[17] Kline and Walters (2017) note that some fully parametric models for binary treatments also happen to possess this property in certain settings. In contrast, our approach *imposes* this property.

Figure 10: Bounds on $\text{LATE}^{+}_{2\to 3}(\alpha)$ Under Different IV–Like Estimands

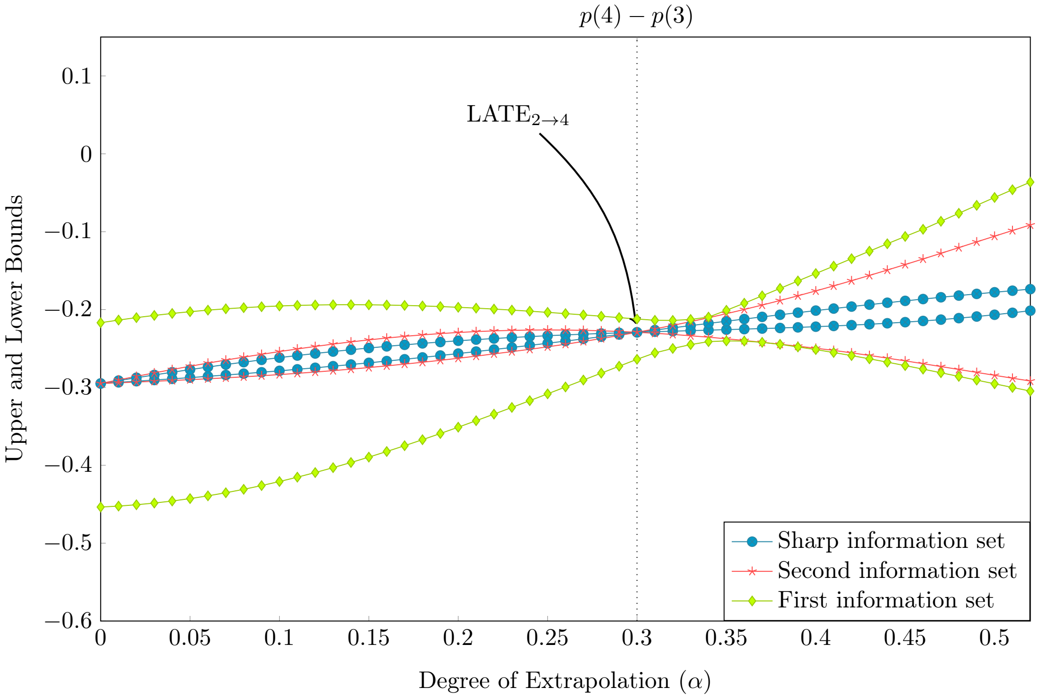
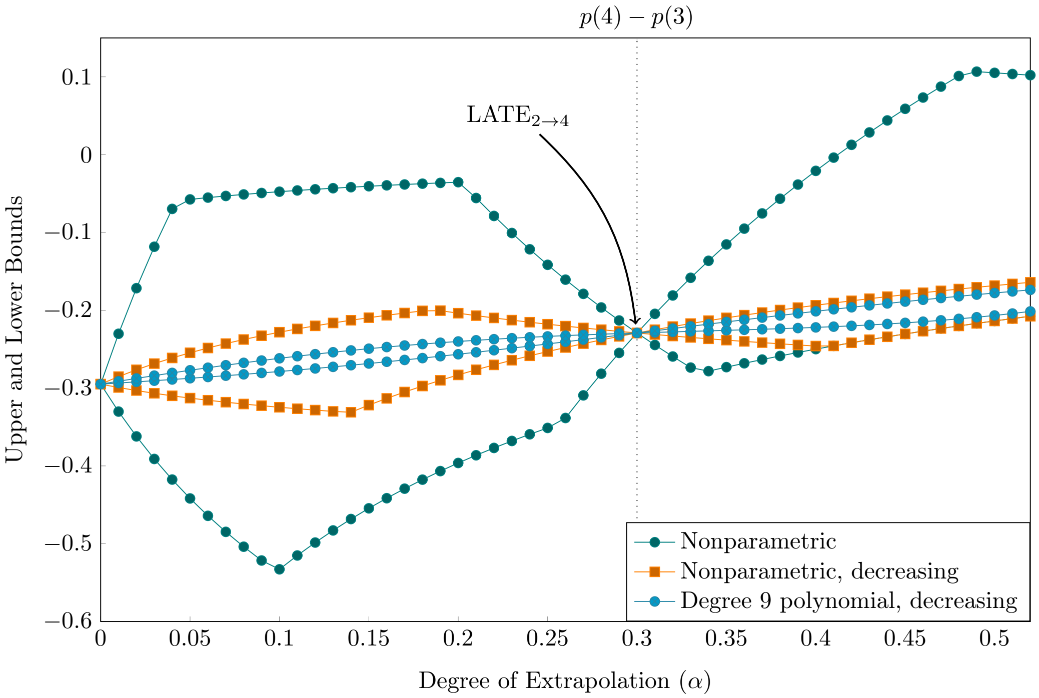


Figure 11: Sharp Bounds on $\text{LATE}^{+}_{2\to 3}(\alpha)$ Under Different Assumptions



trade-off between these factors can be demonstrated by considering bounds on the right-hand extrapolated $Z = 2$ to $Z = 3$ LATE, i.e. $\text{LATE}^{+}_{2\to 3}(\alpha)$, which was plotted in Figure 3. Figure 10 shows these bounds as a function of α for three information sets (specifications of $\mathcal{S}$) under the assumption that the MTR functions are decreasing ninth order polynomials. The first information set is the least restrictive one used in Figure 8, while the second information set is the one from Figure 9 that includes two additional IV–like estimands. The sharp information set represents the best possible bounds that can be achieved using the formulation that is discussed in MST.

As expected, the bounds are nested for any given value of α. For $\alpha = 0$, only the second and sharp information sets yield point identification of $\text{LATE}^{+}_{2\to 3}(0)$, which is just equal to the usual $Z = 2$ to $Z = 3$ LATE. This is simply because the first information set does not include either the $Z = 2$ to $Z = 3$ Wald estimand or a combination of other IV–like estimands that could generate this Wald estimand. Similarly, at $\alpha = p(4) - p(3) = .3$, the right-hand extrapolated $Z = 2$ to $Z = 3$ LATE is equal to the usual $Z = 2$ to $Z = 4$ LATE. Consequently, the bounds for the second and sharp information sets collapse to a point, reflecting the fact that this parameter is point identified. For other values of α, the second and sharp information set bounds are narrow, but not a point. Values of α that are farther away from 0 or .3 correspond to extrapolated LATEs that require more significant extrapolations (or interpolations) away from the instrument variation observed in the data. The intuition that these parameters should be more difficult to identify is visible in the bounds in Figure 10.

In Figure 11, we maintain the sharp information set from Figure 10 and consider a nested set of a priori assumptions on the MTR functions. Naturally, for any given α, weaker assumptions lead to wider bounds. For $\alpha = 0$ and $\alpha = .3$, even the nonmonotone nonparametric bounds yield point identification, again as a consequence of the results of Imbens and Angrist (1994). Figure 11 reveals that an analyst must face up to the compromise between the extent to which they wish to extrapolate (α) and the strength of the assumptions that they impose. There is no free lunch. Given a desired tightness of the bounds, a more ambitious extrapolation can be obtained only by imposing stronger assumptions. Given a set of assumptions, tighter bounds can be obtained only by less ambitious extrapolations. The utility of the general framework is that it gives the researcher the tools to decide exactly where they want to locate on this frontier between assumptions and external validity. The solution to this location problem is unlikely to be the corner solution of reporting only parameters that are nonparametrically point identified, such as the LATE.

Table 4: Weights for Measures of Selection

		Weights	
Quantity	Expression	$\omega_0^\star(u, x, z)$	$\omega_1^\star(u, x, z)$
Average Selection Bias	$E[Y_0 \mid D = 1]$ $- E[Y_0 \mid D = 0]$	$\frac{\mathbb{1}[u \leq p(x, z)]}{P[D = 1]} - \frac{\mathbb{1}[u > p(x, z)]}{P[D = 0]}$	0
Average Selection on the Level	$E[Y_1 \mid D = 1]$ $- E[Y_1 \mid D = 0]$	0	$\frac{\mathbb{1}[u \leq p(x, z)]}{P[D = 1]} - \frac{\mathbb{1}[u > p(x, z)]}{P[D = 0]}$
Average Selection on the Gain	$E[Y_1 - Y_0 \mid D = 1]$ $- E[Y_1 - Y_0 \mid D = 0]$	$-\omega_1^\star(u, x, z)$	$\frac{\mathbb{1}[u \leq p(x, z)]}{P[D = 1]} - \frac{\mathbb{1}[u > p(x, z)]}{P[D = 0]}$

5.8 Testable Implications

It is possible that no solution exists to the programs in (25) because the feasible set ($\mathcal{M}_\mathcal{S}$) is empty. This indicates that the model is misspecified: There does not exist a pair of MTR functions m that can satisfy the researcher's assumptions ($m \in \mathcal{M}$) while also generating the observed data ($\Gamma_s(m) = \beta_s$ for all $s \in \mathcal{S}$). This can happen even if $\mathcal{M}$ is unrestricted, since the choice equation (4) with Assumptions IV is known to have testable implications (Balke and Pearl, 1997; Imbens and Rubin, 1997; Kitagawa, 2015). On the other hand, if $\mathcal{M}$ is restricted, then misspecification could also be due to falsification of these additional restrictions on the MTR functions.

This observation can be used to test a variety of interesting hypotheses. For example, suppose that $\mathcal{M}$ is restricted to contain only MTR pairs with m_0 components consistent with $E[Y_0|D = 1] = E[Y_0|D = 0]$. This can be interpreted as the set of all MTR pairs that lead to no average selection bias. Table 4 shows that this restriction can be imposed as a linear constraint by defining

$$\Gamma_{\text{sel}}(m) \equiv E\left[\int_0^1 m_0(u, X)\left(\frac{\mathbb{1}[u \leq p(X, Z)]}{P[D = 1]} - \frac{\mathbb{1}[u > p(X, Z)]}{P[D = 0]}\right) du\right] \tag{30}$$

and then constraining $\mathcal{M}$ to satisfy $\Gamma_{\text{sel}}(m) = 0$. As long as no other assumptions in the model are deemed suspect, finding that the feasible set in (25) is empty when $\mathcal{M}$ is constrained in this way can be interpreted as evidence against the hypothesis of no selection bias. One could further restrict $\mathcal{M}$ to only contain m such that $\Gamma_{\text{gain}}(m) = 0$, where $\Gamma_{\text{gain}}(m)$ is defined like (30) using the weights for average selection on the gain given in Table 4. Finding the feasible set in (25) to be empty with both $\Gamma_{\text{sel}}(m) = 0$ and $\Gamma_{\text{gain}}(m) = 0$ is evidence against the hypothesis of no unobserved heterogeneity.

6 Other Approaches to Extrapolation

In this section, we compare the general MST framework discussed in Section 5 to several other approaches that have been used in prior research. We show that many of these approaches can be viewed as a special cases of this framework in which the set of admissible MTR functions, $\mathcal{M}$, is restricted to only contains functions with certain functional forms.

6.1 Independence, Constant Effects, and Random Choices

The primary motivation for using an IV method is the concern that D and (Y_0, Y_1) are dependent. In the notation of the choice model, this dependence arises from dependence between U and (Y_0, Y_1) that remains even after conditioning on X. If Y_0 and Y_1 were independent of U, conditional on X, then the MTR functions would be constant in u, i.e. $m_d(u, x) = m_d(x)$ for $d = 0, 1$. In this case, m_0 and m_1 could be directly recovered from the conditional means of Y, since

$$E[Y|D = 1, X = x] = E\left[m_1(U, x)|D = 1, X = x\right] = m_1(x)$$

and similarly for m_0. Any target parameter is then point identified. Indeed, most target parameters we have considered will be identical, since the potential outcomes do not vary systematically with the unobservable factors that are related to treatment status.[18] This independence condition is useful to keep in mind as an extreme case. However, it is unattractive as an assumption, since it assumes away the identification problem that originally motivated considering an IV strategy.

A slightly weaker alternative to independence is to assume that the MTE function $m_1(u, x) - m_0(u, x)$ is constant in u. While this assumption allows for selection bias, in the sense that m_0 and m_1 can still themselves be functions of u, it implies no selection on the unobserved gains from treatment. In other words, while Y_0 is still allowed to depend on D, the treatment effect $Y_1 - Y_0$ is assumed to be independent of D, conditional on X. Under this condition, the z to z' Wald estimand (conditional on $X = x$) point identifies $\text{MTE}(u, x) = \text{MTE}(x)$ for all u, i.e.

$$\begin{aligned}&\frac{E[Y|Z = z', X = x] - E[Y|Z = z, X = x]}{E[D|Z = z', X = x] - E[D|Z = z, X = x]}\\ &\quad = \frac{\int_{p(x,z)}^{p(x,z')} \left[m_1(u, x) - m_0(u, x)\right] du}{p(x, z') - p(x, z)} = \frac{\int_{p(x,z)}^{p(x,z')} \text{MTE}(x)\, du}{p(x, z') - p(x, z)} = \text{MTE}(x).\end{aligned}$$

[18] These observations date back at least to Heckman and Robb (1985a,b).

As a result, any target parameter that depends only on the MTE—but not on the MTRs per se—is point identified. This includes any target parameter with symmetric weights (i.e. $\omega_0^\star = -\omega_1^\star$), such as the ATE, ATT, ATU, and any counterfactual LATE. The intuition behind this is straightforward. If the average causal effect does not vary with unobservables, then it is sufficient to identify this effect for a single subgroup, such as the complier group picked up by the z to z' Wald estimand.[19]

As Heckman and Vytlacil (2007a,b) argue, justifying an MTE function that is constant in u requires strong economic assumptions. In particular, it requires one to assume either that the causal effect of D on Y is identical for all individuals with $X = x$, or else that these individuals either do not know (or do not act on) their idiosyncratic differences in this causal effect. A dissenting opinion is provided by Angrist and Fernández-Val (2013), who argue that this assumption, which they describe as "conditional effect ignorability," can be attractive.[20] We are not sympathetic to this view. Indeed, allowing for unobserved heterogeneity in the effect of D on Y is a key motivation in the modern program evaluation literature, and one which is supported by a large body of empirical work. Assuming it away also disposes of key conceptual distinctions, such as the difference between the LATE and the ATE discussed by Imbens and Angrist (1994).

6.2 Separability of Observed and Unobserved Heterogeneity

In Section 4.2, we saw that a key obstacle to nonparametric point identification is a lack of sufficient instrument variation. One way to ameliorate this problem is to exploit variation in the propensity score that arises from the covariates, X. Carneiro et al. (2011) show how to do this by first writing

$$Y_d = \mu_d(X) + V_d \quad \text{for } d = 0, 1, \tag{32}$$

[19] Using similar intuition, Angrist (2004) shows if the observed propensity score is symmetric around .5, then symmetry assumptions on (Y_0, Y_1, U) are sufficient to point identify the ATE. However, even if the propensity score is fortuitously symmetric in this way, it is not clear how one could motivate the symmetry assumption on unobservables without appealing to one of the explicit parametric approaches discussed in Section 6.3.

[20] The assumption used by Angrist and Fernández-Val (2013) is actually that

$$\frac{\int_{p(x,z)}^{p(x,z')} [m_1(u,x) - m_0(u,x)]\, du}{p(x,z') - p(x,z)} = \int_0^1 [m_1(u,x) - m_0(u,x)]\, du \quad \text{for all } x \text{ and } z. \tag{31}$$

While this is mathematically weaker than assuming that $m_1(u,x) - m_0(u,x)$ is constant in u, it is difficult to see how one could justify (31) without making the stronger assumption.

where $\mu_d(x) \equiv E[Y_d|X = x]$ and $E[V_d|X] = 0$. This by itself is not an assumption, since it is satisfied by letting $V_d = Y_d - \mu_d(X)$. However, Carneiro et al. (2011) then strengthen IV.2 to the assumption that $(V_0, V_1, U) \perp\!\!\!\perp (X, Z)$. Under this stronger independence assumption,

$$m_d(u, x) \equiv E[Y_d|U = u, X = x] = \mu_d(x) + E[V_d|U = u] \quad \text{for } d = 0, 1, \tag{33}$$

which is an additively separable function of x and u. Returning to (15), this implies that

$$E[YD|p(x, Z) = u, X = x] = u\mu_1(x) + \int_0^u E[V_1|U = u']\, du', \tag{34}$$

and similarly for $d = 0$.

Under additive separability, variation in $P = p(X, Z)$ conditional on $X = x$ traces out the same function $E[V_d|U = u]$ for *any* x. By parameterizing $\mu_d(x)$, this property can be exploited to point identify the MTR functions for every (u, x) with u on the interior of the *unconditional* support of P, using a modification of the idea behind Robinson's (1988) partially linear estimator.[21] In contrast, without separability the MTR functions are only point identified on the interior of the support of P, conditional on $X = x$, which is necessarily smaller. Continuous variation in the propensity score is still needed under separability, however the continuity is for the unconditional distribution of P, so it could in principle come from a continuous component of X, even if Z is discrete.

A growing empirical literature has started using this type of separability approach to circumvent limitations in instrument variation.[22] It is important to notice that

[21] For example, suppose that $\mu_1(x) = x'\tau_1$ is linear in parameters. Then from (34), one has

$$E\left[\widetilde{YD}|P, X\right] = P\widetilde{X}'\tau_1,$$

where $\widetilde{YD} \equiv YD - E[YD|P]$, $\widetilde{X} \equiv X - E[X|P]$, and $P \equiv p(X, Z)$ as usual. Given sufficient variation in $P\widetilde{X}$, this enables one to point identify τ_1, and therefore $\mu_1(x)$ for any x. Treating τ_1 as known, it follows that

$$E\left[YD - PX'\tau_1|P = u\right] = \int_0^u E[V_1|U = u']\, du',$$

so that $E[V_1|U = u]$ is point identified for any u in the interior of the support of P by differentiating the left-hand side. It follows from (33) that $m_1(u, x) = \mu_1(x) + E[V_1|U = u]$ is point identified for any x and any u in the interior of the unconditional support of P. See Carneiro et al. (2011) for more details on this argument.

[22] Examples include Carneiro and Lee (2009), Carneiro et al. (2011), Maestas et al. (2013), Eisenhauer, Heckman, and Vytlacil (2015), Carneiro et al. (2016), Kline and Walters (2016), Brinch et al. (2017), and

assuming $(V_0, V_1, U) \perp\!\!\!\perp (X, Z)$ does not imply that Y_0 or Y_1 are independent of X. Rather, the dependence of Y_0 and Y_1 on X is captured through the conditional mean function $\mu_d(X)$, which is often specified as linear-in-parameters in applications. Still, the stronger independence assumption implies, among other things, that X and U are independent. This nearly elevates X to the status of an instrument, albeit one which does not need to obey the usual exclusion restriction. In applications, the types of variables usually included in X, such as socio-demographic controls, are unlikely to be exogenous in this way.

Brinch et al. (2017) observe that the stronger independence assumption is not actually necessary for the purpose of expanding the effective support of the propensity score. Instead, the separability in (33) can be achieved by writing (32) and adding the assumption that $E[V_d|U, X] = E[V_d|U]$ to IV.2. This assumption still allows for X and U to be dependent in arbitrary ways, thereby addressing the previous concerns while still allowing the researcher to exploit the separability assumption. In Section 5.5, we showed that separability can be imposed in the general MST framework as a direct restriction on the set $\mathcal{M}$ of admissible MTR functions.

In some settings, the separability in (33) can be motivated by economic theory through standard classes of technologies or preferences. For example, suppose that m_d is a production function in state d, with Y_d denoting output and X denoting observed input factors. Additive separability in m_d is then implied by perfect substitutability between X and unobserved input factors. Alternatively, if input and output factors are measured in logs, then additive separability is implied by unit elasticity between observable and unobservable inputs, as in a Cobb-Douglas production function. More generally, additive separability in m_d is compatible with a production technology in which unobserved productivity differences across agents are factor neutral, which is a standard assumption for methods of estimating production functions.

6.3 Parametric Assumptions

Another natural response to the problem of limited instrument variation is to impose parametric structure. Using parametric assumptions to correct for unobserved heterogeneity has a long history, dating back to Gronau (1974) and Heckman (1974, 1976, 1979). Heckman, Tobias, and Vytlacil (2001, 2003) apply this approach to the binary treatment setting considered in this paper. The case they study, which is the most widely used, maintains (32) and the assumption that $(V_d, \Phi^{-1}(U))$ is bivariate normal and independent of X for $d = 0, 1$, where Φ^{-1} is the inverse of the standard normal

Cornelissen et al. (forthcoming).

cumulative distribution function (CDF).[23]

Under this assumption, (33) reduces to

$$m_d(u,x) = \mu_d(x) + \operatorname{Corr}(V_d, U)\operatorname{Var}(V_d)\Phi^{-1}(u) \quad \text{for } d = 0,1, \tag{35}$$

since the conditional mean function for bivariate normal random variables is linear in the conditioning value. Assuming that there is at least one value x for which $p(x,Z)$ has two support points, say $p(x,z') \equiv \widetilde{u}_1 > \widetilde{u}_2 \equiv p(x,z)$, it follows from (15) that

$$\begin{aligned} &E[YD|p(x,Z) = \widetilde{u}_1, X = x] - E[YD|p(x,Z) = \widetilde{u}_2, X = x] \\ &\qquad = \operatorname{Corr}(V_1, U)\operatorname{Var}(V_1)\int_{\widetilde{u}_2}^{\widetilde{u}_1} \Phi^{-1}(u)\,du, \end{aligned} \tag{36}$$

and similarly for $d = 0$. This implies that $\operatorname{Corr}(V_1,U)\operatorname{Var}(V_1)$ is identified, and hence that the functional form restriction in (35) is sufficient to point identify the MTR functions *everywhere*, at least as long as there is enough variation in X to identify the μ_d component.

This identification argument hinges heavily on the assumption of bivariate normality, which ensures that $E[V_d|U = u]$ is a function that is completely determined by the single unknown quantity, $\operatorname{Corr}(V_d,U)\operatorname{Var}(V_d)$. Two points of exogenous variation, i.e. z and z', are sufficient to identify this quantity. Once it is known, the functional form of the normal distribution is used to extrapolate to any other value required to evaluate a given target parameter. This argument should be concerning whenever normality of an unobserved error lacks an economic motivation. In our view, it is an exceptional case when one actually can motivate normality as anything other than a convenient functional form assumption.

There are other parametric approaches that yield the same payoff, but which may sometimes be easier to interpret and motivate than bivariate normality. For example, suppose that instead of (35), we assume that $m_d(u,x)$ is linear in its u component for every x, i.e.

$$m_d(x,u) = \mu_d(x) + \lambda_d(x)u \quad \text{for } d = 0,1, \tag{37}$$

where both μ_d and λ_d are unknown functions of x. From (15), we have

$$E[YD|p(x,Z) = u, X = x] = u\mu_1(x) + \frac{1}{2}u^2\lambda_1(x),$$

[23] Alternatively, and equivalently, the same assumption can be made about (V_d, U) using the pre-normalized choice equation (2).

and similarly for $d = 0$. Since $P[D = 1|p(x, Z) = u, X = x] = u$ by definition of the propensity score, it follows that

$$E[Y|D = 1, p(x, Z) = u, X = x] = \mu_1(x) + \frac{1}{2}u\lambda_1(x). \tag{38}$$

Using (38) with two values $\widetilde{u}_1 \equiv p(x, z') \neq p(x, z) = \widetilde{u}_2$ and $X = x$ fixed shows that both $\mu_1(x)$ and $\lambda_1(x)$ are point identified. The same argument could be repeated for any other x for which the distribution of $p(x, Z)|X = x$ has two support points. Alternatively, if separability is imposed (i.e. $\lambda_d(x) = 1$), then this propensity score variation is needed conditional on only a single value of x, as in (36).

This linearity assumption was first suggested by Brinch, Mogstad, and Wiswall (2012).[24] The assumption yields point identification through effectively the same extrapolation argument as bivariate normality. Linearity has a straightforward interpretation: Holding $X = x$ fixed, a one percentage point change in the unobserved willingness to pay for treatment u results in an average increase in Y_d of $\lambda_d(x)$. In contrast, under normality, a one unit increase in u results in a different average increase in Y_d depending on the base value of u, where the form of this difference is dictated by the shape of the inverse normal CDF. Since the two assumptions are not nested, their implications must be considered on a case by case basis.[25] However, at least in some applications, the comparative ease of interpreting linearity should make it easier to motivate.

Another benefit of considering a functional form restriction like linearity is that it is straightforward to relax the restriction. As discussed by Brinch et al. (2012, 2017), whereas a linear MTR can be point identified with a binary instrument, point identifying a quadratic MTR requires a ternary instrument, a cubic MTR requires a quaternary instrument, etc.[26] However, the notion that the richness of the data should constrain the assumptions of the model is, in our view, backward. The assumptions of the model should be considered on their own; if the data is insufficiently rich to point identify the desired model then this must be recognized.

[24] See Kowalski (2016) for a more recent application of the same idea.

[25] It should also be noted that bivariate normality imposes a restriction on the entire distributions of (Y_0, U) and (Y_1, U), while the linearity assumption (37) is a restriction only on the means, i.e. the MTR functions. That is, bivariate normality leads to a fully parametric model, whereas under (37) the model is still semiparametric. This engenders several differences for identification of other features of the distribution of Y_0 and Y_1, as well for the efficiency of statistical inference. A more direct comparison would be between (35) and (37) as different restrictions on the forms of the MTR functions.

[26] These observations are related to proposed series estimators of the local instrumental variables estimand (16), as in Moffitt (2008) and French and Song (2014). Brinch et al. (2012, 2017) show that more flexible specifications of the MTE functions can be point identified by first point identifying the MTR functions separately, as in (38).

The general framework in Section 5 provides a disciplined solution to this criticism, since it allows researchers to maintain parametric restrictions without requiring point identification. Point identification is still allowed as a special case, however. In particular, notice that the set $\mathcal{M}_{\mathcal{S}}$ in (23) is a system of $|\mathcal{S}|$ linear equations, with the number of variables given by the combined dimensions of $m \equiv (m_0, m_1)$. The assumption that $\mathcal{S}$ can be specified to include enough non-redundant IV–like estimands to exactly pin down a single $m \in \mathcal{M}$ is a higher dimensional analog to the arguments in (36) and (38). As always, where such a specification is possible depends both on the richness of the data, i.e. how many distinct IV–like estimands can be found, as well as how flexibly the researcher wishes to specify $\mathcal{M}$.

6.4 Rank Invariance

Rank invariance is an assumption about unobserved heterogeneity that was introduced to the program evaluation literature by Heckman, Smith, and Clements (1997). The formal assumption is that $F_{0|x}(Y_0) = F_{1|x}(Y_1)$ (almost surely), where $F_{0|x}$ and $F_{1|x}$ denote the marginal distributions of Y_0 and Y_1, conditional on $X = x$. In words, $F_{0|x}(Y_0) \in [0, 1]$ can be viewed as an agent's rank (order) in the distribution of $Y_0|X = x$, and rank invariance postulates that this order remains the same in the $D = 1$ counterfactual outcome distribution. While rank invariance allows Y_0 and Y_1 to be dependent with D, conditional on X, it has the unusual implication that the joint conditional-on-X distribution of Y_1 and Y_0 is degenerate, since it implies that Y_1 is a deterministic function of Y_0 and X.[27]

Chernozhukov and Hansen (2005) showed that rank invariance can be used to point identify the ATE under a somewhat non-standard relevance condition for the relationship between D and Z. Their model does not impose the choice equation (4). Vuong and Xu (2017) show that also imposing a choice equation allows one to obtain point identification of conventional parameters, such as the ATE and ATT, under the usual relevance condition used to ensure the existence of Wald estimands. Their argument works by identifying the relationship (mapping) between Y_0 and Y_1 among the compliers, i.e. those individuals whose choices would be affected by a given shift in the instrument. Under rank invariance, one can then infer the distribution of Y_0 for the subpopulation that would always choose $D = 1$ by applying this mapping to their ob-

[27] Assuming rank invariance in this way only makes sense in settings where Y is continuously distributed. Rank invariance can be interpreted as a restriction on the dimension of unobserved heterogeneity. In this sense, it is intuitively similar to models for discrete outcomes with a threshold crossing form, as considered for example by Vytlacil and Yıldız (2007), Chesher (2010), Shaikh and Vytlacil (2011), Bhattacharya, Shaikh, and Vytlacil (2012), Machado, Shaikh, and Vytlacil (2013), Mourifié (2015), and Torgovitsky (2017), among others.

served $Y = Y_1$ outcomes. Similarly, one can infer the distribution of Y_1 for individuals who would always choose $D = 0$. This strategy effectively uses the rank invariance assumption to extrapolate from individuals whose treatment choices are affected by the instrument to those whose choices are not.

6.5 Analytical Bounds

The approach in Section 5 is influenced by an important line of work primarily due to Manski (1989, 1990, 1994, 1997, 2003) and Manski and Pepper (2000, 2009). Unlike most of Manski's work on IV methods, the MST approach maintains the choice equation (4).[28] Maintaining a choice equation such as this places a substantive restriction on behavior, but one that is indispensable for considering the effects of policy interventions that do not mandate treatment or non-treatment.[29] As we argued in Section 3, we view such policies as being typical of interesting counterfactual questions in economic applications.

Another difference between the MST framework and Manski's research is more practical. Instead of deriving explicit expressions for bounds, it takes a computational approach of solving linear programs. The benefit of the computational approach is flexibility: The same procedure can be used for a large class of target parameters under a wide range of assumptions without requiring new analytical derivations. These derivations can be extremely challenging for models that maintain multiple assumptions. The cost of a computational approach is that without analytical expressions for the bounds it is difficult to understand specific details of their structure. Our view is that the benefits of the computational approach outweigh this cost in many settings.

As an example of this benefit, recall Figures 5 and 7 of our numerical illustration. For Figure 7, we specified the MTR functions to be constant splines in a way that exactly replicates the nonparametric bounds. With some effort, one could derive the analytical bounds for this case. In contrast, for Figure 5, we specified the MTR functions as ninth degree polynomials. This narrowed the bounds considerably by ruling out the discontinuous MTR functions that are permitted in Figure 7. We view this

[28] Incidentally, the monotonicity assumption underlying the choice equation is exactly Manski's (1997) monotone treatment response assumption, but applied to the counterfactual relationship between Z and D, rather than between D and Y.

[29] An interesting result due to Heckman and Vytlacil (2001b) shows that when the implications of the choice model (4) are not rejected (c.f. Section 5.8), the choice model has no impact on the sharp nonparametric bounds for the ATE derived by Manski (1994). This result extends to the ATT and the ATU, but clearly not to parameters, such as PRTEs, that are defined only given a choice equation. Similarly, the result also loses meaning when placing assumptions on the MTR functions that have no clear interpretation in the absence of a choice equation.

as attractive for many applications, since these discontinuous functions are unlikely to represent important cases to guard against in many economic settings. However, analytic expressions for the bounds under a ninth degree polynomial are unknown, and seem difficult to derive. Using the MST computational approach, this derivation was not necessary, and the bounds were returned using standard software almost instantaneously.

7 Conclusion and Directions for Future Research

We have discussed the implications of unobserved heterogeneity in treatment effects for using IV methods to answer specific well-defined policy questions. The identification challenge inherent in doing this can be viewed as a problem of extrapolating from the individuals whose treatment choices are affected by the variation in the data to the individuals relevant for the counterfactual question. Several methods for formally conducting this extrapolation have been proposed in the literature. We reviewed these approaches, and argued that their reliance on point identification is a weakness. We discussed a general framework, developed fully in Mogstad et al. (2017), that nests these approaches but allows for more flexibility by recognizing the possibility of partial identification.

Partial identification approaches are sometimes criticized for yielding empirical conclusions that are insufficiently informative for practitioners (e.g. Imbens, 2013, pp. F407–F409). We view computational methods, such as the one discussed in Section 2, as important tools for answering this criticism. The flexibility of the MST method means that a researcher can smoothly adjust their policy question (target parameter), or the assumptions they are willing to maintain, in a way that approaches point identification as a special case. As a result, the tightness of the bounds they report is at their discretion, while still being disciplined by the reality that stronger conclusions require stronger assumptions. We view this as an important improvement over the current practice—common in applied work—of hoping that a given estimand is relevant for the policy change of interest to the researcher. This type of "faith-based extrapolation" is ad hoc and potentially misleading.[30]

There are many avenues down which the partial identification approaches to identification and extrapolation of treatment effects can be further developed. While we focused on the widely studied case of a binary treatment, applying similar ideas to models with continuous or discrete (ordered or unordered) treatments would be useful

[30] For a different view, see Angrist (2016).

and involves many complications.[31] The issues of policy relevance and need for extrapolation that arises in IV models is also a concern in other common program evaluation strategies. For example, it may be interesting to apply ideas similar to those discussed here to help ameliorate the local nature of regression discontinuity designs.[32] Similar ideas could also be applied to more complicated evaluation settings involving dynamics, mediation, peer effects, or other challenges for identification.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117. 24

Angrist, J. D. (2004): "Treatment Effect Heterogeneity in Theory and Practice," *The Economic Journal*, 114, C52–C83. 35

——— (2016): "Sometimes You Get What You Need: Discussion of Mogstad, Santos, and Torgovitsky," Slides; NBER Labor Studies Summer Meetings. 42

Angrist, J. D. and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*, 88, 450–477. 2

Angrist, J. D. and I. Fernández-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in *Advances in Economics and Econometrics*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, 401–434. 35

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527. 2, 43

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 43

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014. 17

Angrist, J. D. and M. Rokkanen (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff," *Journal of the American Statistical Association*, 110, 1331–1344. 43

[31] For discussions of methods for multiple discrete treatments, see Angrist and Imbens (1995), Heckman, Urzua, and Vytlacil (2006), Heckman and Vytlacil (2007b), Heckman and Urzua (2010), Kirkeboen et al. (2016), Lee and Salanié (2016), and Heckman and Pinto (2016), among others. Methods for continuous treatments have been considered by Angrist et al. (2000), Chesher (2003), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), Torgovitsky (2015), Masten (2015), and Masten and Torgovitsky (2016), among others.

[32] Various approaches to extrapolation in regression discontinuity designs have been proposed by Wing and Cook (2013), Dong and Lewbel (2015), Angrist and Rokkanen (2015) and Rokkanen (2015).

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 33

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012): "Treatment effect bounds: An application to SwanGanz catheterization," *Journal of Econometrics*, 168, 223–243. 40

Bitler, M., H. Hoynes, and T. Domina (2014): "Experimental Evidence on Distributional Effects of Head Start," Tech. rep. 2

Bitler, M. P., J. B. Gelbach, and H. W. Hoynes (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *The American Economic Review*, 96, 988–1012. 2

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. 6

Black, S. E., P. J. Devereux, and K. G. Salvanes (2005): "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education," *The Quarterly Journal of Economics*, 120, 669–700. 17, 18

Brinch, C. N., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," *Working paper.* 39

——— (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039. 2, 18, 36, 37, 39

Buchinsky, M. (1994): "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica*, 62, 405–458. 24

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. 13, 19

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 2, 13, 19, 35, 36

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. 2, 36

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. 2, 36

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632. 26

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. 40

Chernozhukov, V., W. K. Newey, and A. Santos (2015): "Constrained conditional moment restriction models," *arXiv preprint arXiv:1509.06311.* 27

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 43

——— (2010): "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 78, 575–601. 40

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (forthcoming): "Who benefits from universal childcare? Estimating marginal returns to early childcare attendance," *Journal of Political Economy*. 2, 37

Dong, Y. and A. Lewbel (2015): "Identifying the effect of changing the policy threshold in regression discontinuity models," *Review of Economics and Statistics*, 97, 1081–1092. 43

Doyle Jr., J. J. (2007): "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *The American Economic Review*, 97, 1583–1610. 2

Dupas, P., H. V. K. M. and A. P. Zwane (2016): "Targeting health subsidies through a non-price mechanism: A randomized controlled trial in Kenya," *Science*, 353, 889–895. 12

Dupas, P. (2014): "ShortRun Subsidies and LongRun Adoption of New Health Products: Evidence From a Field Experiment," *Econometrica*, 82, 197–228. 7, 19

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs," *Journal of Political Economy*, 123, 413–443. 36

Felfe, C. and R. Lalive (2014): "Does Early Child Care Help or Hurt Children's Development?" Tech. Rep. 8484. 2

Firpo, S., N. M. Fortin, and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, 77, 953–973. 2

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 43

French, E. and J. Song (2014): "The Effect of Disability Insurance Receipt on Labor Supply," *American Economic Journal: Economic Policy*, 6, 291–337. 2, 39

Gronau, R. (1974): "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143. 37

Havnes, T. and M. Mogstad (2015): "Is universal child care leveling the playing field?" *Journal of Public Economics*, 127, 100–114. 2

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694. 37

——— (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *The Journal of Human Resources*, 32, 441–462. 16

Heckman, J., J. L. Tobias, and E. Vytlacil (2001): "Four Parameters of Interest in the Evaluation of Social Programs," *Southern Economic Journal*, 68, 210. 37

——— (2003): "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85, 748–755. 37

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.* 37

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. 37

——— (1996): "Comment," *Journal of the American Statistical Association*, 91, 459–462. 16

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 2

——— (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–98. 16

Heckman, J. J. and R. Pinto (2016): "Unordered Monotonicity," *Working paper.* 43

Heckman, J. J. and R. Robb (1985a): "Alternative methods for evaluating the impact of interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer, Cambridge University Press. 34

——— (1985b): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics*, 30, 239–267. 34

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. 40

Heckman, J. J. and J. A. Smith (1998): "Evaluating the Welfare State," *NBER Working Paper 6542*, this was reprinted in a volume called "Frisch Centenary" which is not available online. 10

Heckman, J. J. and S. Urzua (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156, 27–37. 43

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. 43

Heckman, J. J. and E. Vytlacil (2001a): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–111. 3, 6, 13

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 3, 6, 9, 12, 13, 15, 20

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. 3, 6, 12, 13, 18

——— (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. 6, 41

——— (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. 6, 18

——— (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. 6, 35

——— (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. 6, 35, 43

Hull, P. (2016): "Estimating Hospital Quality with Quasi-Experimental Data," *Working paper.* 2

Imbens, G. (2013): "Book Review Feature: Public Policy in an Uncertain World: By Charles F. Manski (Cambridge, MA: Harvard University Press. 2013, pp. 224, $39.95. ISBN: 978-0674066892)," *The Economic Journal*, 123, F401–F411. 42

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423. 15, 30

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 3, 5, 6, 11, 15, 17, 32, 35

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 43

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. 33

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings, and Self-Selection *," *The Quarterly Journal of Economics*, 131, 1057–1111. 2, 43

Kitagawa, T. (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. 33

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. 2, 36

——— (2017): "Through the Looking Glass: Heckits, LATE, and Numerical Equivalence," *Working paper.* 30

Koenker, R. (2005): *Quantile Regression*, Cambridge University Press. 24

Kowalski, A. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," *NBER Working paper 22363.* 39

Lee, S. and B. Salanié (2016): "Identifying Effects of Multivalued Treatments," *Working paper.* 43

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2013): "Instrumental Variables and the Sign of the Average Treatment Effect," *Working paper.* 40

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. 2, 36

Manski, C. (1994): "The selection problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 41

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360. 41

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 41

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 41

——— (2003): *Partial identification of probability distributions*, Springer. 41

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 27, 41

——— (2009): "More on monotone instrumental variables," *Econometrics Journal*, 12, S200–S216. 41

Masten, M. A. (2015): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *cemmap working paper 25/15.* 43

Masten, M. A. and A. Torgovitsky (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics*, 98, 1001–1005. 43

Matzkin, R. L. (2007): "Chapter 73 Nonparametric identification," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5307–5368. 6

Miguel, E., S. Satyanath, and E. Sergenti (2004): "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy*, 112, 725–753. 2

Moffitt, r. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations," *Annales d'Economie et de Statistique*, 239–261. 2, 39

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *NBER Working Paper.* 3, 4, 7, 9, 11, 14, 19, 42

Mourifié, I. (2015): "Sharp bounds on treatment effects in a binary triangular system," *Journal of Econometrics*, 187, 74–81. 40

Nybom, M. (2017): "The Distribution of Lifetime Earnings Returns to College," *Journal of Labor Economics*, 000–000. 2

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. 36

Rokkanen, M. (2015): "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," *Working paper.* 43

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955. 40

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 43

——— (2017): "Partial Identification by Extending Subdistributions," *SSRN Electronic Journal.* 40

Vuong, Q. and H. Xu (2017): "Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity," *Quantitative Economics*, 8, 589–610. 40

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 5, 6

Vytlacil, E. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757–779. 40

Walters, C. (2014): "The Demand for Effective Charter Schools," Tech. rep. 2

Wing, C. and T. D. Cook (2013): "Strengthening the Regression Discontinuity Design using Additional Design Elements: A Within-Study Comparison," *Journal of Policy Analysis and Management*, 32, 853–877. 43