Annual Review of Economics

Identification and Extrapolation of Causal Effects with Instrumental Variables

Magne Mogstad[1,2,3] and Alexander Torgovitsky[1]

[1]Department of Economics, University of Chicago, Chicago, Illinois 60637, USA; email: magne.mogstad@gmail.com

[2]Statistics Norway, 0177 Oslo, Norway

[3]National Bureau of Economic Research, Cambridge, Massachusetts 02138, USA

Annu. Rev. Econ. 2018. 10:577–613

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-101617-041813

JEL codes: C21, C26, C36, C61

Keywords

causal inference, instrumental variables, extrapolation, local average treatment effect, marginal treatment effects, partial identification

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 making inferences 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 rigorously perform this extrapolation. We show that many of these approaches can be nested as special cases of a general framework that embraces the possibility of partial identification.

1. INTRODUCTION

Instrumental variables (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 effect of an endogenous treatment variable on an outcome of interest. IV methods attack the problem of selection on unobservables by using only variation in the treatment that is induced by the instrument. However, 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 (TSLS), produce estimates of causal effects that are specific to these individuals.

The goal of this article is to review IV methods that can be used to make inferences about causal effects for individuals other than those affected by the instrument at hand. This requires extrapolating away from the variation in treatment induced by the available instrument. There are at least two common empirical situations in which such extrapolation is needed.

The first is when the instrument is based on the lack of pattern or predictability in a natural event that cannot be shifted by policy, such as the weather (Angrist et al. 2000, Miguel et al. 2004) or the gender composition of children (Angrist & Evans 1998, Black et al. 2005). Instruments derived from such natural experiments have gained popularity over the past few decades, in part because it is often relatively easy to argue that the instrument is exogenous. However, this advantage comes at the cost of external validity. In particular, the group of individuals whose behavior is affected by a natural experiment is often different from the group of individuals who would be affected under an interesting policy counterfactual. If the causal effect of the treatment varies between the two groups, then the estimates produced by standard IV methods can differ dramatically from the parameters relevant for the policy counterfactual.[1] Methods for extrapolation can be used to rigorously address this critique by analyzing the sensitivity of the empirical conclusions to the size or composition of the group affected by the natural experiment.

A second setting in which extrapolation is needed is one in which the instrument represents a policy change, and the researcher is interested in the effect of expanding or contracting the policy. In this case, standard IV methods already identify the causal effect for individuals whose treatment choice is affected by the past policy change. However, using the past policy change to inform about a new counterfactual change under which the policy is expanded requires extrapolation. For example, Evans & Ringel (1999) show that smoking during pregnancy declines when state-level taxes on cigarettes are increased. They use these results to form an IV estimate of the impact of maternal smoking on birth weight. Their IV estimate can be interpreted as an average causal effect for mothers whose smoking behavior was changed as a result of the small tax changes observed in their data. Among the states that changed cigarette taxes during the sample period, the average difference in taxes was only approximately $0.05 per pack. Yet Evans & Ringel (1999) use their IV estimate to draw conclusions about a proposed bill that would raise cigarette taxes by $1.10 per pack. To conduct this exercise coherently requires rigorously extrapolating the causal effect for mothers with high price elasticity to those who are less price elastic.

The structure of our review is as follows. In Section 2, we review a widely 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 & Angrist (1994). This

[1]Several studies in diverse fields report evidence of unobserved heterogeneity in causal effects. Heckman (2001) compiles a list of studies performed prior to 2001. More recent work includes that of Bitler et al. (2006, 2014), Doyle (2007), Moffitt (2008), Carneiro & Lee (2009), Firpo et al. (2009), Carneiro et al. (2011, 2016), Maestas et al. (2013), Walters (2014), Felfe & Lalive (2014), French & Song (2014), Havnes & Mogstad (2015), Kirkeboen et al. (2016), Kline & Walters (2016), Hull (2016), Cornelissen et al. (2018), Nybom (2017), and Brinch et al. (2017), among many others.

monotonicity condition gives rise to the important related concepts of the marginal treatment effect and response functions developed by Heckman & 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 in Section 3, and throughout the review, that some conventional treatment parameters, such as the average treatment effect, often represent uninteresting policy counterfactuals and thus 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 & 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. We discuss specific examples of PRTEs and show that the local average treatment effect (LATE) of Imbens & Angrist (1994) can be viewed as a special case of a PRTE.

In Section 4, we discuss two conditions under which the target parameter is already point identified and no extrapolation is needed. 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 et al. (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 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 the general framework 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 that one can draw 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 discussion of these challenges and develop a method for addressing them.

2. MODEL

2.1. Potential Outcomes and Endogeneity

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 individual had their treatment status been exogenously set to 0 or 1, respectively. The relationship between observed and potential outcomes is given by

$$Y = DY_1 + (1 - D)Y_0. \quad 1.$$

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, comparing the distribution of Y for the treated ($D = 1$) and control ($D = 0$) groups confounds the effect of the treatment with other differences between 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 individual'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 is solely due to the causal effect of D on Y, i.e., from the difference between Y_1 and Y_0.

2.2. Selection Into Treatment

A key theme of the literature, and of this review, 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 & Angrist (1994) introduce a simple model of treatment choice summarized by what they call 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 individual or weakly decreases it for every individual. Vytlacil (2002) shows that, under the standard exogeneity assumption on Z, the monotonicity condition is equivalent to the existence of a weakly separable selection (or choice) equation,

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

where ν is an unknown function and U is a continuously distributed random variable. The Imbens & Angrist (1994) monotonicity condition can be seen clearly to arise from Equation 2, since the separability between $\nu(X, Z)$ and U implies that a change in Z induces a shift either toward or away from treatment for all values of U.

Our review focuses on approaches that maintain this monotonicity condition or, equivalently, the choice model in Equation 2. This 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 Equation 2) is crucial for considering counterfactual policies that do not mandate a choice of treatment.

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

Assumption 1. D is determined by Equation 2.

[2]Our discussion also requires some mild technical conditions involving the existence of moments that we do not explicitly mention but that are clear from the context.

Assumption 2. $(Y_0, Y_1, U) \perp\!\!\!\perp Z|X$ holds true, where $\perp\!\!\!\perp$ denotes conditional independence.

Assumption 3. U is continuously distributed, conditional on X.

Assumption 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 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 Assumption 2 would fail. For simplicity, we maintain the stronger assumption throughout our discussion.

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

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

Therefore, the normalization allows Equation 2 to be rewritten as

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

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

2.3. Marginal Treatment Effect and Response Functions

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

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

In words, $\text{MTE}(u, x)$ is the average causal effect of D on Y for individuals with selection unobservable $U = u$ and observed characteristics $X = x$.

The dependence of the MTE on u for a fixed x allows for unobserved heterogeneity in treatment effects, as indexed by an individual's latent propensity to choose treatment, u. The choice equation (Equation 4) implies that, given X, individuals with lower values of U are more likely to take treatment, regardless of their realization of Z.[5] 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 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.

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

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

[5] This is only a convention; if Equation 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.

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

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

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 below, 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 ordinary least squares (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 this review, we use a running numerical example to provide graphical explanations of the key concepts. The example is loosely based on the empirical application of Mogstad et al. (2017). They analyze how a class of potential subsidy regimes can promote the use of a preventive 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 are taken from Dupas (2014) and feature 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 $P[Z = z] = 1/4$ for each of $z = 1,2,3,4$. We take the propensity score in our simulation to be equal to the estimated propensity score in the data, which is given by

$$p(1) = 0.12, \qquad p(2) = 0.29, \qquad p(3) = 0.48, \qquad \text{and} \qquad p(4) = 0.78.$$

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. To generate the distribution of Y, we set the MTR (and implied MTE) functions to be quadratic in u:

$$\begin{aligned} m_0(u) = 0.9 - 1.1u + 0.3u^2 \qquad &\text{and} \qquad m_1(u) = 0.35 - 0.3u - 0.05u^2, \\ \text{so that} \qquad m_1(u) - m_0(u) &= -0.55 + 0.8u - 0.35u^2. \end{aligned} \tag{7.}$$

As shown in **Figure 1**, these MTR functions are decreasing in u for both the treated and untreated states. Recalling that higher values of u correspond to lower propensities to choose treatment, this means that individuals less likely to purchase the mosquito net are also less likely to be afflicted by malaria regardless of whether they purchase the mosquito net. This situation 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 nonconstant 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 private knowledge of how likely they are to benefit from a mosquito net—for example, due to the prevalence of mosquitoes in their sleeping areas—and partly base their purchase decisions on this knowledge.

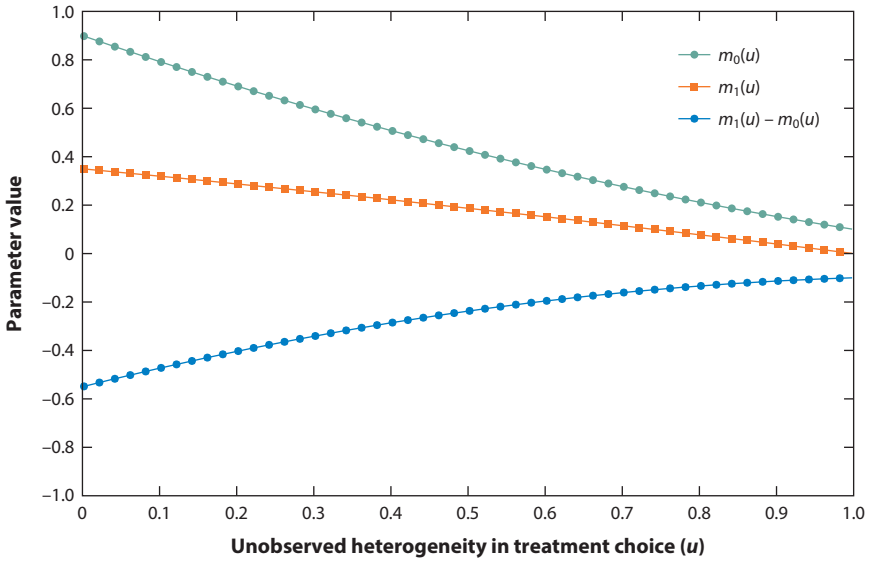

Figure 1

Marginal treatment response and marginal treatment effect functions used to generate data in the running numerical illustration.

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

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

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 sections below. The tables below 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 of Heckman & Vytlacil (2005).[6]

[6]Most of the expressions in **Tables 1–4** can be found in the work of Heckman & Vytlacil (2005). The expressions for parameters with asymmetric weights (i.e., $m_0 \neq -m_1$) are derived by Mogstad et al. (2017). Note also that Mogstad et al. (2017) consider a slightly more general version of Equation 8 in which the integrating measure (i.e., du) can be something

Table 1 Weights for conventional treatment effect parameters

Target parameter	Expression	Weights	
		$\omega_0^\star(u,x,z)$	$\omega_1^\star(u,x,z)$
Average untreated outcome	$E[Y_0]$	1	0
Average treated outcome	$E[Y_1]$	0	1
ATE	$E[Y_1 - Y_0]$	-1	1
ATE given $X = \bar{x}$, where $P[X = \bar{x}] > 0$	$E[Y_1 - Y_0 \mid X = \bar{x}]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[x=\bar{x}]}{P[X=\bar{x}]}$
ATT	$E[Y_1 - Y_0 \mid D = 1]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u \le p(x,z)]}{P[D=1]}$
ATU	$E[Y_1 - Y_0 \mid D = 0]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u > p(x,z)]}{P[D=0]}$
LATE for $z_0 \to z_1$ given $X = x$, where $p(x,z_1) > p(x,z_0)$	$E[Y_1 - Y_0 \mid p(x,z_0) < U \le p(x,z_1), X = x]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[p(x,z_0) < u \le p(x,z_1)]}{p(x,z_1) - p(x,z_0)}$

Abbreviations: ATE, average treatment effect; ATT, average treatment on the treated; ATU, average treatment on the untreated; LATE, local average treatment effect.

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 Equation 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 required to choose $D = 0$.

Another commonly considered target parameter is the average treatment on the treated (ATT). **Figure 2** plots the average of the $d = 1$ weights over observables, i.e., $E[\omega_1^\star(u,X,Z)]$, as a function of u for the ATT and several other conventional target parameters listed in **Table 1** for our running numerical example. All of the parameters in **Table 1** have symmetric weights 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 using the corresponding formula in **Table 1** to compute

$$E\left[\omega_1^\star(u,X,Z)\right] = \frac{E[\mathbb{1}[p(X,Z) \ge u]]}{P[D=1]} = \frac{P[p(X,Z) \ge u]}{P[D=1]}, \qquad 9.$$

which is necessarily decreasing as a function of u.

The ATT provides the average change in outcomes that would be experienced by the treated group if it switched 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 & 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 nonparticipants to participate in a program.

other than the Lebesgue measure. For example, this allows one to define the target parameter to be the MTE at a given value $\tilde{u}$, i.e., $\beta^\star = E[m_1(\tilde{u},X) - m_0(\tilde{u},X)]$.

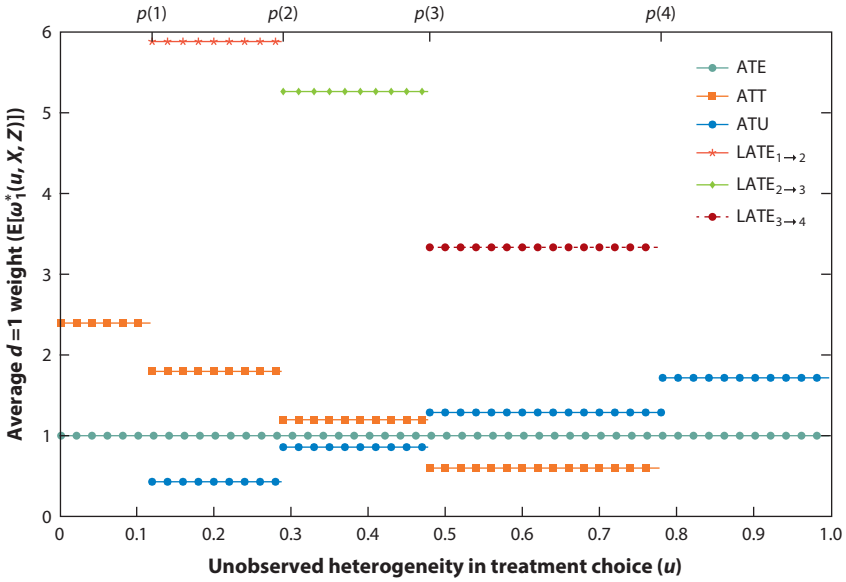

Figure 2

Weights for conventional target parameters in the running numerical illustration. The horizontal axis indexes the average weight functions by unobserved heterogeneity in treatment choice $u \in [0, 1]$, with smaller values of u corresponding to individuals that are more likely to choose $D = 1$ (see Equation 4). The vertical axis reports the average of the $d = 1$ weights $E[\omega_1^*(u, X, Z)]$ in regions where they are nonzero. Abbreviations: ATE, average treatment effect; ATT, average treatment on the treated; ATU, average treatment on the untreated; LATE, local average treatment effect.

The ATE, ATT, and ATU can all be defined without reference to the choice model (Equation 4).[7] 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 LATE, which was first studied by Imbens & 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 $Z = z_0$ to $Z = z_1$, with $p(x, z_1) > p(x, z_0)$ and conditional on $X = x$. These weights are only nonzero over the region $(p(x, z_0), p(x, z_1)]$. Examining Equation 4, one can see that this region corresponds to realizations of U for which an individual with $X = x$ would choose $D = 1$ if assigned $Z = z_1$ but would choose $D = 0$ if assigned $Z = z_0$. Imbens & Angrist (1994) refer to this unobservable subgroup as the (z_0 to z_1) compliers. In the next section, we show that LATEs are specific examples of the more general concept of a PRTE.

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, as well as the policy counterfactuals that they address, is dubious when requiring or preventing treatment is

[7] However, as we demonstrate in Sections 5 and 6, the choice model facilitates thinking about identification of these parameters as an extrapolation problem.

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 nonenrollment 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 by Mogstad et al. (2017), is the decision to own a mosquito net ($D = 1$). This is an important preventive health care measure in many parts of the developing world. Mandating nonownership—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 feasible, interesting, and 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.[8]

A choice model like Equation 4 provides a framework for considering the effect of a policy intervention that influences (but may not fully determine) choice behavior. We follow Heckman & Vytlacil (1999, 2005) in considering policies that change the propensity score, p, and/or the instrument, Z, but that 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 (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 assumption 2 (see Section 2.2). 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 under the policy. Treatment choice under a policy a is given by

$$D^a \equiv \mathbb{1}[U \leq p^a(X, Z^a)]. \tag{10.}$$

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

$$Y^a = D^a Y_1 + (1 - D^a) Y_0. \tag{11.}$$

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

$$\text{PRTE} \equiv \frac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]}, \tag{12.}$$

where it is assumed 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.[9]

[8]For example, Dupas et al. (2016) provide 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.

[9]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, then one can alternatively define the PRTE as $E[Y^{a_1}] - E[Y^{a_0}]$ (see Carneiro et al. 2010, pp. 380–81; Heckman & Vytlacil 2001a).

Table 2 Weights for PRTEs

Target parameter	Expression	$\omega_1^\star(u,x,z) = -\omega_0^\star(u,x,z)$
Generalized LATE for $U \in (\underline{u}, \overline{u}]$	$E[Y_1 - Y_0 \mid U \in (\underline{u}, \overline{u}]]$	$\frac{\mathbb{1}[u \in (\underline{u}, \overline{u}]]}{\overline{u} - \underline{u}}$
PRTE for policy (p^{a_1}, Z^{a_1}) relative to policy (p^{a_0}, Z^{a_0})	$\frac{E[Y^{a_1}] - E[Y^{a_0}]}{E[D^{a_1}] - E[D^{a_0}]}$	$\frac{P[u \le p^{a_1}(x, Z^{a_1}) \mid X = x] - P[u \le p^{a_0}(x, Z^{a_0}) \mid X = x]}{E[p^{a_1}(X, Z^{a_1})] - E[p^{a_0}(X, Z^{a_0})]}$
Additive PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(x,z) = p(x,z) + \alpha$	$\frac{\mathbb{1}[u \le p(x,z) + \alpha] - \mathbb{1}[u \le p(x,z)]}{\alpha}$
Proportional PRTE with magnitude α	PRTE with $Z^\star = Z$ and $p^\star(x,z) = (1+\alpha)p(x,z)$	$\frac{\mathbb{1}[u \le (1+\alpha)p(x,z)] - \mathbb{1}[u \le p(x,z)]}{\alpha E[p(X,Z)]}$
PRTE for an additive α shift of the jth component of Z	PRTE with $Z^\star = Z + \alpha e_j$ and $p^\star(x,z) = p(x,z)$	$\frac{\mathbb{1}[u \le p(x, z+\alpha e_j)] - \mathbb{1}[u \le p(x,z)]}{E[p(X, Z+\alpha e_j)] - E[p(X,Z)]}$

Abbreviations: LATE, local average treatment effect; PRTE, policy-relevant treatment effect.

3.4. Examples of Policy-Relevant Treatment Effects

PRTEs can be expressed as target parameters with the same form as Equation 8. The choice of weights, $(\omega_0^\star, \omega_1^\star)$, depends on the policies being compared.[10] **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 (*a*) an additive α change in the propensity score, i.e., $p^{a_1} = p + \alpha$; (*b*) a proportional $(1+\alpha)$ change in the propensity score, i.e., $p^{a_1} = (1+\alpha)p$; and (*c*) an additive α shift in the distribution of 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 policies 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 choice, such as a subsidy.

In all of these definitions, α is a quantity that could be either estimated or hypothesized by the researcher. Mogstad et al. (2017) consider PRTEs of the first type, 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 simply specify a value of α that represents an empirically interesting change in the probability of choosing treatment.

3.5. Local Average Treatment Effects are Policy-Relevant Treatment Effects

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 individual receives $Z = z_1$, against a policy a_0, under which every individual receives $Z = z_0$.[11]

[10]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 $a = a_0, a_1$.

[11]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_1 and z_0, respectively.

Choices under these policies are

$$D^{a_0} \equiv \mathbb{1}[U \leq p(z_0)] \quad \text{and} \quad D^{a_1} \equiv \mathbb{1}[U \leq p(z_1)],$$

where $p(z_1) > p(z_0)$ are the propensity score values in the observed data. The PRTE for this policy comparison is

$$\frac{E[Y^{a_1} - Y^{a_0}]}{E[D^{a_1} - D^{a_0}]} = \frac{E\left[(D^{a_1} - D^{a_0})(Y_1 - Y_0)\right]}{p(z_1) - p(z_0)} = E\left[Y_1 - Y_0 \mid p(z_0) < U \leq p(z_1)\right], \qquad 13.$$

where we use $D^{a_1} - D^{a_0} = \mathbb{1}[p(z_0) < U \leq p(z_1)]$. The right-hand side of Equation 13 is precisely the z_0 to z_1 LATE introduced by Imbens & Angrist (1994).

More generally, Heckman & 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 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 Local Average Treatment Effects

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 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_0 to z_1 LATE to an expansion (or contraction) of the complier subpopulation. For this purpose, we define right- and left-hand α-extrapolations of the z_0 to z_1 LATE as

$$\begin{aligned} \text{LATE}^{+}_{z_0 \to z_1}(\alpha) &\equiv E\left[Y_1 - Y_0 \mid p(z_0) < U \leq p(z_1) + \alpha\right], \\ \text{LATE}^{-}_{z_0 \to z_1}(\alpha) &\equiv E\left[Y_1 - Y_0 \mid p(z_0) - \alpha < U \leq p(z_1)\right]. \end{aligned} \qquad 14.$$

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

$$\text{LATE}^{\pm}_{z_0 \to z_1}(\alpha) \equiv E\left[Y_1 - Y_0 \,\middle|\, p(z_0) - \frac{\alpha}{2} < U \leq p(z_1) + \frac{\alpha}{2}\right]. \qquad 15.$$

These parameters are defined over subgroups that take the z_0 to z_1 complier group and expand it by α to the left, to the right, or in a split between both sides. One could also allow $\alpha < 0$ in Equations 14 and 15, in which case the parameters would be interpolated LATEs.

Imbens & Angrist (1994) show that the z_0 to z_1 LATE is nonparametrically point identified for any observed z_0 and z_1 as long as $p(z_1) > p(z_0)$. This result has produced a focus on these types of LATEs as parameters of interest. Since these LATEs only reflect causal effects for z_0 to z_1 compliers, their external validity (or generalizability) can be limited (Imbens 2010). Some authors have criticized the practice of focusing 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_0 to z_1 LATE reduces back to the usual z_0 to z_1 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. **Figure 3** shows that the $Z = 2$ to $Z = 3$ LATE is sensitive

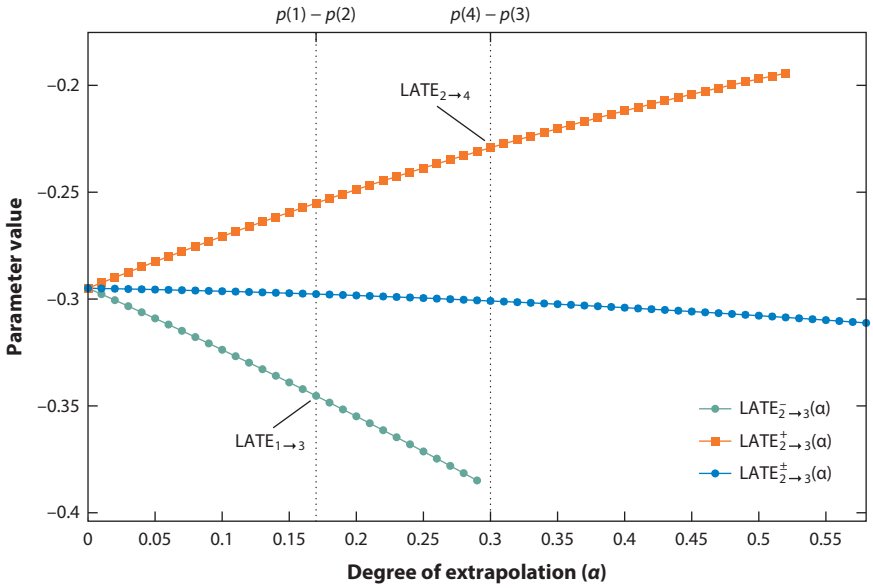

Figure 3

Extrapolated LATEs in the running numerical illustration. Abbreviation: LATE, local average treatment effect.

to extrapolation to either the left or the 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 NONPARAMETRICALLY 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. No extrapolation is necessary in these special cases.

4.1. When the Target Parameter is a Local Average Treatment Effect

Imbens & Angrist (1994) show that, under Assumptions 1–3 and their monotonicity condition (which, again, is equivalent to Equation 4), the z_0 to z_1 LATE, conditional on $X = x$, is point identified by the Wald estimand, i.e.,

$$E\left[Y_1 - Y_0 | p(x, z_0) < U \leq p(x, z_1), X = x\right]$$

$$= \frac{E[Y|X = x, Z = z_1] - E[Y|X = x, Z = z_0]}{E[D|X = x, Z = z_1] - E[D|X = x, Z = z_0]}.$$

The LATE may be an interesting target parameter if the observed instrument variation from z_0 to z_1 represents an intervention or policy change. For example, Angrist & 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_0 to z_1 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. Black et al. (2005) use twin births and the sex composition of prior births as instruments for family size. Their LATE estimates for these instruments suggest that family size has a small effect on a child's outcomes.

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 & Vytlacil (1999, 2001c) show 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 $\tilde{u}$ in the interior of its support. To see this, note that, in general, it can be shown using Equation 4 and assumption 2 (Section 2.2) that

$$\begin{aligned} E\left[YD \;\middle|\; p(x,Z) = u, X = x\right] &= \int_0^u m_1(u',x)\,\mathrm{d}u', \\ E\left[Y(1-D) \;\middle|\; p(x,Z) = u, X = x\right] &= \int_u^1 m_0(u',x)\,\mathrm{d}u'. \end{aligned} \tag{16.}$$

As a consequence, if the objects on the left-hand sides of Equation 16 can be differentiated at $u = \tilde{u}$, then we obtain

$$\begin{aligned} \frac{\partial}{\partial u} E\left[YD \;\middle|\; p(x,Z) = u, X = x\right]\Big|_{u=\tilde{u}} &= m_1(\tilde{u},x), \\ \frac{\partial}{\partial u} E\left[Y(1-D) \;\middle|\; p(x,Z) = u, X = x\right]\Big|_{u=\tilde{u}} &= -m_0(\tilde{u},x), \\ \text{and thus} \quad \frac{\partial}{\partial u} E\left[Y \;\middle|\; p(x,Z) = u, X = x\right]\Big|_{u=\tilde{u}} &= m_1(\tilde{u},x) - m_0(\tilde{u},x), \end{aligned} \tag{17.}$$

so that the MTRs and MTE at $(\tilde{u}, x)$ are point identified. The third line of Equation 17 is what Heckman & Vytlacil (1999, 2001c) refer to 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 nonzero only for values of $(\tilde{u}, x)$ for which $\tilde{u}$ 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.

Unfortunately, this support condition severely limits the types of target parameters that are point identified without additional assumptions. It requires a continuous instrument because, if Z is discrete, then the distribution of $P \equiv p(X, Z)$, conditional on X, will also be discrete, so that differentiation in Equation 17 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 nonparametrically point identified. PRTEs for policies that involve extrapolating beyond the currently available support will not be point identified without additional assumptions.[12] 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. Mogstad et al. (2017) analyze this question using the Dupas (2014) data (on which our running numerical illustration is styled) from a randomized controlled experiment in Kenya in which the price for a new type of mosquito net 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 individuals are offered the product at a given price. To do so, they use the randomly assigned prices as a (discrete) instrument for purchasing the health product. Many of the PRTEs that 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 methodology described in the next section.

5. A GENERAL FRAMEWORK FOR INFERENCE ABOUT CAUSAL EFFECTS

In the previous section, we discuss 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) to conduct this extrapolation.

5.1. What We Know: Instrumental Variables–Like Estimands

The starting point for Mogstad et al. (2017) is the observation that a rich class of identified quantities can also be written in the same form (Equation 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 IV regression that includes a constant term but that does not include any other covariates X. Assuming $\text{Cov}(D, Z) \neq 0$, this estimand is given by

$$\beta_{\text{IV}} \equiv \frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)}. \qquad 18.$$

[12]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.

Heckman & Vytlacil (2005) show that β_{IV} can be written as

$$\beta_{\text{IV}} = \int_0^1 [m_1(u,X) - m_0(u,X)]\ \omega_{\text{IV}}(u,X,Z)\,\mathrm{d}u, \tag{19.}$$

where ω_{IV} is an identified weighting function. The similarity between Equation 19 and Equation 8 suggests that β_{IV} carries some useful information about the possible values of $\beta^\star$.

Mogstad et al. (2017) 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 $\beta_s \equiv E[s(D,X,Z)Y]$. Mogstad et al. (2017) call the function s an IV-like specification and call the quantity β_s that s generates an IV-like estimand. Mogstad et al. (2017, proposition 1) show that, for any s,

$$\begin{aligned} \beta_s \;&=\; E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,X,Z)\,\mathrm{d}u\right] + E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,X,Z)\,\mathrm{d}u\right], \\ \text{where}\quad &\omega_{0s}(u,x,z) \equiv s(0,x,z)\mathbb{1}[u > p(x,z)] \\ \text{and}\quad &\omega_{1s}(u,x,z) \equiv s(1,x,z)\mathbb{1}[u \le p(x,z)]. \end{aligned} \tag{20.}$$

Intuitively, Equation 20 comes from writing $\beta_s = E[s(0,X,Z)Y_0] + E[s(1,X,Z)Y_1]$ and then using the selection equation (Equation 4) to express these quantities in terms of m_0 and m_1. Note that the propensity score $p(x,z) \equiv P[D=1|X=x,Z=z]$ is an identified quantity that is the same across different choices of s.

The weights in Equation 20 can be shown to nest the weighting expressions derived by Heckman & Vytlacil (2005). For example, their weights for β_{IV} can be generated by taking

$$s(d,x,z) = \frac{z - E[Z]}{\text{Cov}(D,Z)} \tag{21.}$$

and inserting this choice of s into the definitions in Equation 20. However, the expression in Equation 20 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).[13] Deriving these weights is 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 TSLS 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 thus are not themselves objects of interest. However, Equation 20 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. 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

[13] The phrases included instrument and excluded instrument are meant in the sense typically introduced in textbook treatments of the linear IV model with constant treatment effects.

Table 3 Common IV-like estimands

Estimand	β_s	$s(d, x, z)$	Notes
Wald (z_0 to z_1)	$\frac{E[Y\|Z=z_1]-E[Y\|Z=z_0]}{E[D\|Z=z_1]-E[D\|Z=z_0]}$	$\frac{\frac{\mathbb{1}[z=z_1]}{P[Z=z_1]}-\frac{\mathbb{1}[z=z]}{P[Z=z_0]}}{E[D\|Z=z_1]-E[D\|Z=z_0]}$	$P[Z=z_j]\neq 0, j=0,1$ and $E[D\|Z=z_1] \neq E[D\|Z=z_0]$
IV slope	$\frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(D,Z)}$	$\frac{z-E[Z]}{\mathrm{Cov}(D,Z)}$	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
TSLS (jth component)	$e_j'\left(\Pi E[\widetilde{Z}\widetilde{X}']\right)^{-1}\left(\Pi E[\widetilde{Z}Y]\right)$	$e_j'(\Pi E[\widetilde{Z}\widetilde{X}'])^{-1}\Pi\widetilde{Z}$	$\Pi\equiv E[\widetilde{X}\,\widetilde{Z}']E[\widetilde{Z}\widetilde{Z}']^{-1}$ Z vector
OLS slope	$\frac{\mathrm{Cov}(Y,D)}{\mathrm{Var}(D)}$	$\frac{d-E[D]}{\mathrm{Var}(D)}$	
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

Abbreviations: IV, instrumental variables; OLS, ordinary least squares; TSLS, two-stage least squares.

admissible set encodes the 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 that could also deliver the collection of identified IV estimands $\{\beta_s : s \in \mathcal{S}\}$ through Equation 20.

To do this, it is helpful to view the weighted integrals for the target parameter (Equation 8) and the IV-like estimands (Equation 20) as functions of m. Specifically, for the target parameter, we define the function

$$\Gamma^\star(m) \equiv E\left[\int_0^1 m_0(u,X)\omega_0^\star(u,X,Z)\,\mathrm{d}u\right] + E\left[\int_0^1 m_1(u,X)\omega_1^\star(u,X,Z)\,\mathrm{d}u\right], \qquad 22.$$

and for any IV-like specification s, we define the function

$$\Gamma_s(m) \equiv E\left[\int_0^1 m_0(u,X)\omega_{0s}(u,X,Z)\,\mathrm{d}u\right] + E\left[\int_0^1 m_1(u,X)\omega_{1s}(u,X,Z)\,\mathrm{d}u\right]. \qquad 23.$$

Now, suppose that the data were generated according to Equations 1 and 4 under Assumptions 1–3 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

$$\mathcal{M}_\mathcal{S} \equiv \left\{m \in \mathcal{M} : \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}\right\}. \qquad 24.$$

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

$$\mathcal{B}_\mathcal{S}^\star \equiv \{b \in \mathbb{R} : b = \Gamma^\star(m) \text{ for some } m \in \mathcal{M}_\mathcal{S}\}. \qquad 25.$$

Intuitively, $\mathcal{B}_\mathcal{S}^\star$ 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, $\mathcal{B}_\mathcal{S}^\star$ 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 we have $b \in \mathcal{B}_\mathcal{S}^\star$; otherwise, we have $b \notin \mathcal{B}_\mathcal{S}^\star$. Under weak conditions on $\mathcal{M}$, it is possible to show that $\mathcal{B}_\mathcal{S}^\star$ will be a closed interval, say, $[\underline{\beta^\star}, \overline{\beta^\star}]$. In this case, the process of characterizing $\mathcal{B}_\mathcal{S}^\star$ can be simplified

to the task of solving two optimization problems, namely

$$\underline{\beta}^\star \equiv \inf_{m\in\mathcal{M}} \Gamma^\star(m) \quad \text{subject to} \quad \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}$$
$$\text{and } \overline{\beta}^\star \equiv \sup_{m\in\mathcal{M}} \Gamma^\star(m) \quad \text{subject to} \quad \Gamma_s(m) = \beta_s \text{ for all } s \in \mathcal{S}. \tag{26.}$$

5.3. Computing the Bounds

Both $\Gamma^\star$ and Γ_s are linear functions of m. This endows the optimization problems (Equation 26) 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 Equation 26 are infinite dimensional. Second, Equation 26 could be difficult to solve unless the admissible set $\mathcal{M}$ has enough structure.

Mogstad et al. (2017) solve both problems by replacing $\mathcal{M}$ with a finite dimensional linear space. 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

$$m_d(u,x) = \sum_{k=0}^{K_d} \theta_{dk} b_{dk}(u,x) \quad \text{for } d = 0, 1, \tag{27.}$$

where $b_{dk}(u,x)$ are known basis functions. Substituting Equation 27 into the definition of $\Gamma^\star(m)$, we have

$$\begin{aligned}\Gamma^\star(m) &= \sum_{d\in\{0,1\}} \sum_{k=0}^{K_d} \theta_{dk} E\left[\int_0^1 b_{dk}(u,X)\omega_d^\star(u,X,Z)\,\mathrm{d}u\right] \\ &\equiv \sum_{d\in\{0,1\}} \sum_{k=0}^{K_d} \theta_{dk}\gamma_{dk}^\star, \quad \text{where } \gamma_{dk}^\star \equiv E\left[\int_0^1 b_{dk}(u,X)\omega_d^\star(u,X,Z)\,\mathrm{d}u\right].\end{aligned} \tag{28.}$$

The $\gamma_{dk}^\star$ terms in Equation 28 are identified population quantities that depend on the known basis functions, b_{dk}, and the known (or identified) weighting functions, $\omega_d^\star$, but that do not depend on θ. Imposing Equation 27 therefore turns the objective of Equation 26 into a linear function of the finite dimensional parameter, θ. Similarly, Equation 27 implies that

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

for every $s \in \mathcal{S}$, so that the constraints for the IV-like specifications in Equation 26 are also linear in θ.

Under Equation 27, each $m \in \mathcal{M}$ is parameterized by a finite dimensional θ. In analogy 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 a closed convex polyhedron, i.e., a set determined by a finite collection of linear inequalities. In this case, the maximization problem in Equation 26 reduces to the linear program

$$\overline{\beta}^\star = \max_{\theta\in\Theta} \sum_{d\in\{0,1\}} \sum_{k=0}^{K_d} \gamma_{dk}^\star\theta_{dk} \quad \text{subject to} \quad \sum_{d\in\{0,1\}} \sum_{k=0}^{K_d} \gamma_{sdk}\theta_{dk} = \beta_s \text{ for all } s \in \mathcal{S}; \tag{29.}$$

a similar reduction is possible for the minimization problem. Linear programs like Equation 29 can be solved reliably and are routinely used in empirical work using quantile regressions (see e.g., Abadie et al. 2002, Buchinsky 1994, Koenker 2005). We view the computational benefits afforded

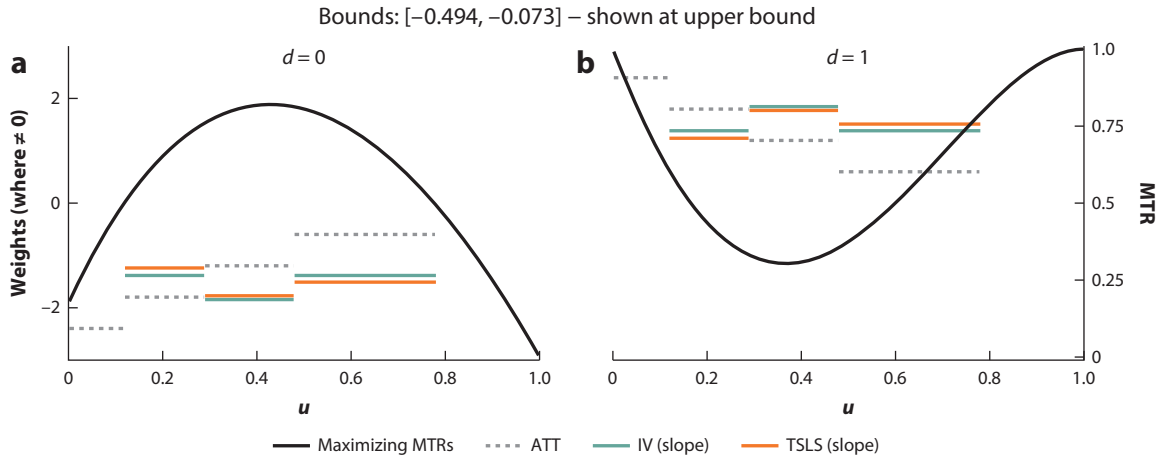

Figure 4

Fourth-degree polynomial bounds ($K_0 = K_1 = 4$) on the ATT. The left-hand vertical axis measures the weight functions for the target parameter and IV-like estimands; the right-hand vertical axis measures MTR functions. (*a*) Weights and a MTR function for $d = 0$. (*b*) Weights and a MTR function for $d = 1$. Abbreviations: ATT, average treatment on the treated; IV, instrumental variables; MTR, marginal treatment response; TSLS, two-stage least squares.

by linear programming as sufficiently important to restrict ourselves to this case in the following sections.

5.4. Parametric and Nonparametric Bounds

The interpretation of Equation 27 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 of K_d add more variables of optimization to Equation 26. 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. **Figure 4** is generated by solving the maximization problem (Equation 29) 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 are examples of functions m_0 and m_1 that yield the upper bound on $\beta^\star$, which we take to be the ATT. These choices are not unique. What is unique is the attained upper bound of .049 for $\beta^\star$. This upper bound is constrained by the requirement that IV-like estimands generated by this black curve are 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 are equal to the corresponding IV-like estimand, β_s. 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 still ensuring that this requirement is satisfied for all $s \in \mathcal{S}$.

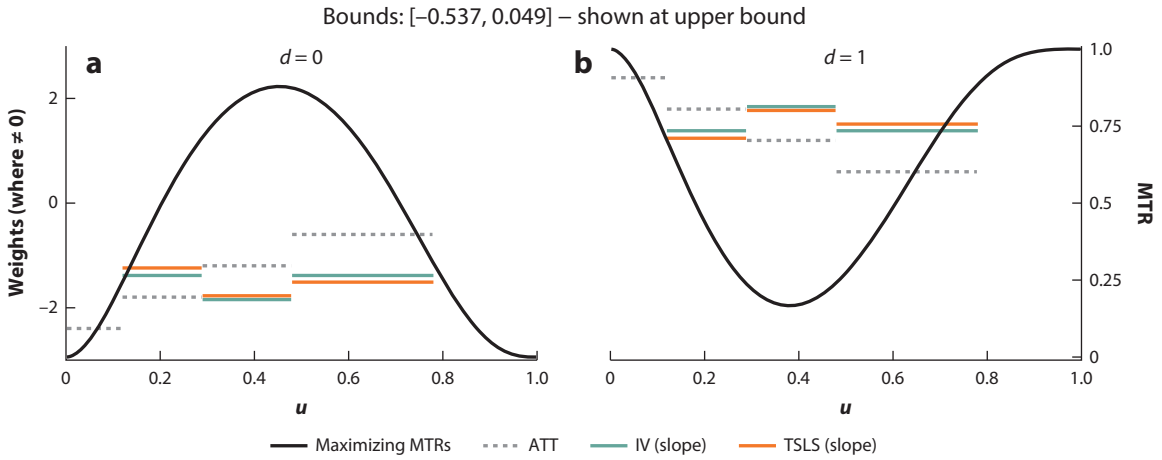

Figure 5

Ninth-degree polynomial bounds ($K_0 = K_1 = 9$) on the ATT. The left-hand vertical axis measures the weight functions for the target parameter and IV-like estimands; the right-hand vertical axis measures MTR functions. The black curves represent choices of m_0 and m_1 that yield the upper bound on $\beta^\star$, which we take to be the ATT. (*a*) Weights and a MTR function for $d = 0$. (*b*) Weights and a MTR function for $d = 1$. Abbreviations: ATT, average treatment on the treated; IV, instrumental variables; MTR, marginal treatment response; TSLS, two-stage least squares.

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 can be viewed as the subset of the set of ninth-degree polynomials by setting the coefficients on the fifth and higher terms to zero. **Figure 6** demonstrates this phenomenon for a range 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, Mogstad et al. (2017) show that Equation 27 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.[14] **Figure 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 assumptions 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 in **Figure 6** 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

[14]Mogstad et al. (2017) also provide a statistical inference framework in which the dimension of θ grows asymptotically as in sieve estimation (Chen 2007).

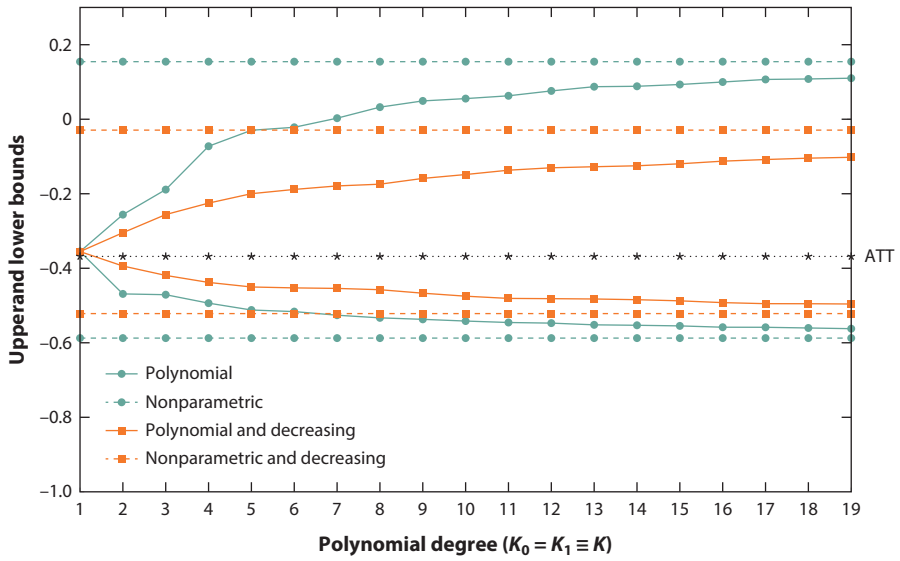

Figure 6

Bounds on the ATT for different K with and without requiring the MTR functions to be decreasing. The dotted line indicates the value of the ATT in the data generating process. Abbreviations: ATT, average treatment on the treated.

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 $m_1(\cdot, x) - m_0(\cdot, x)$ 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

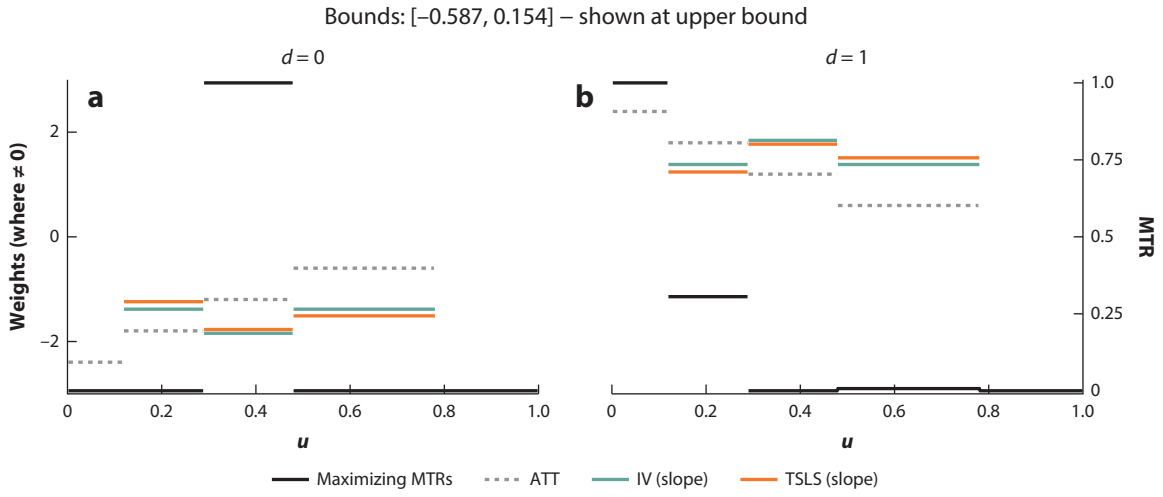

Figure 7

Exact nonparametric bounds on the ATT. Abbreviations: ATT, average treatment on the treated; IV, instrumental variables; MTR, marginal treatment response; TSLS, two-stage least squares.

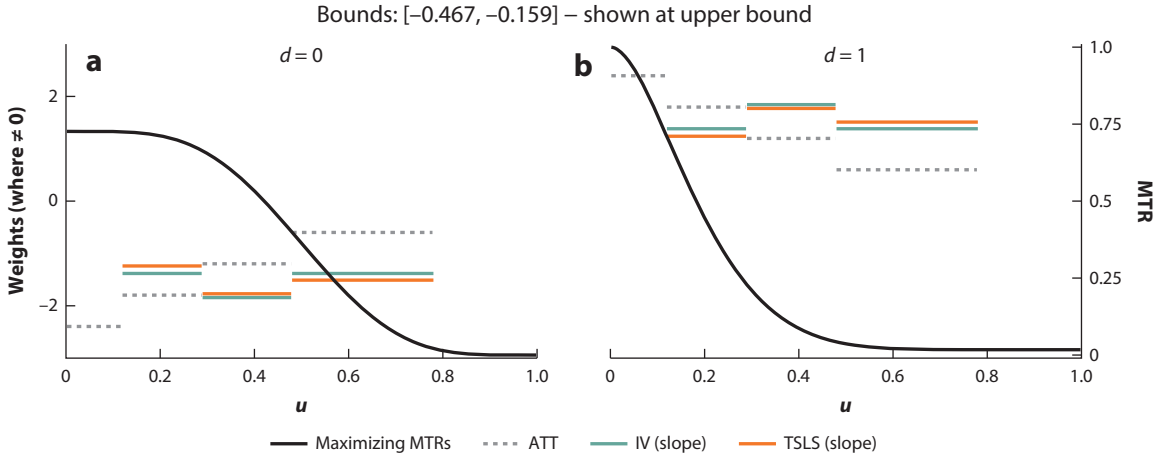

Figure 8

Ninth-degree decreasing polynomial bounds on the ATT. Abbreviations: ATT, average treatment on the treated; IV, instrumental variables; MTR, marginal treatment response; TSLS, two-stage least squares.

gains from treatment. This is similar to the monotone treatment selection assumption of Manski & Pepper (2000).[15]

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 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**. One justification for this assumption would a selection story 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 or home environment. The additional monotonicity restriction mechanically tightens the bounds by imposing an additional constraint on the optimization problem (Equation 29). In particular, it ensures that the maximizing MTR functions shown in **Figure 5** are no longer feasible, since neither 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. **Figure 6** 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 nonmonotone sixth-degree polynomials.

Another type of nonparametric shape restriction that is often used is separability between the observed (X) and unobserved (U) components, i.e., the assumption that

$$m_d(u,x) = m_d^U(u) + m_d^X(x) \quad \text{for } d = 0,1, \tag{30.}$$

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

[15] Chernozhukov et al. (2015) provide a discussion of various shape restrictions implied by economic theory in several empirical applications.

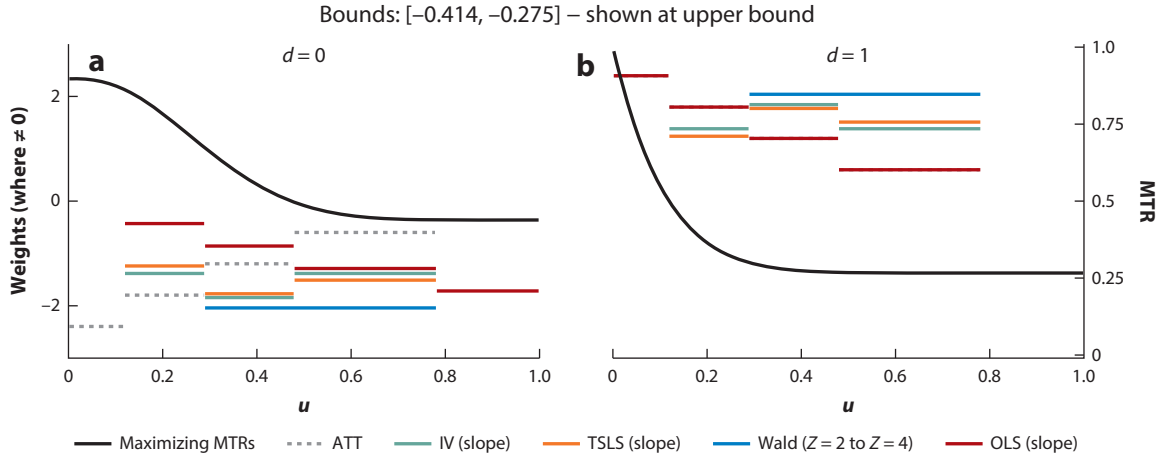

Figure 9

Ninth-degree decreasing polynomial bounds with more IV-like estimands. Abbreviations: ATT, average treatment on the treated; IV, instrumental variables; MTR, marginal treatment response; OLS, ordinary least squares; TSLS, two-stage least squares.

combinations of assumptions simultaneously (e.g., both monotonicity and separability) is simply a matter of imposing both restrictions at the same time.

In practice, these shape restrictions are imposed through the specification of Θ for a given finite basis (Equation 27). The restrictions involved in ensuring that a given θ generates a MTR pair with a particular set of shape properties depends on the choice of basis. As discussed by Mogstad et al. (2017), 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 Equation 26 remains a linear program.

5.6. Choosing Instrumental Variables–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 Equation 26 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. This results in a substantial decrease in the width of the bounds. Mogstad et al. (2017) 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 high noise will be unhelpful. A natural starting point is to choose IV-like specifications that generate the estimands that one would ordinarily be interested in when not concerned about endogeneity or unobserved heterogeneity. For example, one set of s would be the vector of OLS estimands, another would be the vector of IV estimands, and a third would be a vector of TSLS estimands from including an additional instrument.

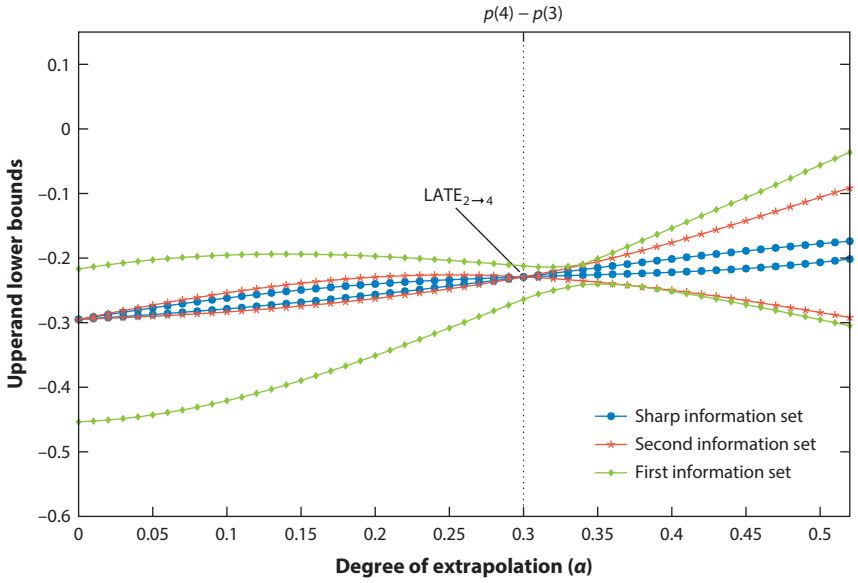

Figure 10

Bounds on $\text{LATE}^{+}_{2\to3}(\alpha)$ under different IV-like estimands. Abbreviation: LATE, local average treatment effect.

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. 414–15), who recommends reporting both a standard LATE and 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 $\mathcal{M}_S$ and all potential values of the target parameter, $\mathcal{B}^{\star}_S$, will necessarily be consistent with this LATE.[16]

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 trade-off among these factors can be demonstrated by considering bounds on the right-hand extrapolated $Z = 2$ to $Z = 3$ LATE, i.e., $\text{LATE}^{+}_{2\to3}(\alpha)$, which is 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 one used in **Figure 8**, while the second information set is the one from **Figure 9**, which includes two additional IV-like estimands. The sharp information set represents the best possible bounds that can be achieved using a formulation that is discussed by Mogstad et al. (2017).

[16] Kline & 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.

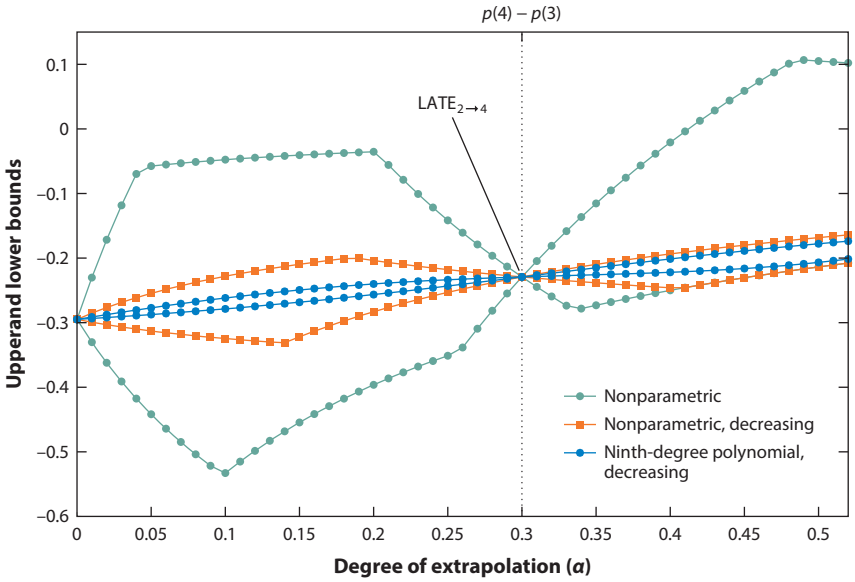

Figure 11

Sharp bounds on $\text{LATE}^{+}_{2\to3}(\alpha)$ under different assumptions. Abbreviation: LATE, local average treatment effect.

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 $\text{LATE}^{+}_{2\to3}(0)$, 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 $\alpha = p(4) - p(3) = 0.3$, 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 assumptions on the MTR functions. For any given α, weaker assumptions naturally lead to wider bounds. For $\alpha = 0$ and $\alpha = 0.3$, even the nonmonotone nonparametric bounds yield point identification, again as a consequence of the results of Imbens & Angrist (1994). **Figure 11** reveals that an analyst must acknowledge a 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 internal and external validity. It is unlikely that the researcher's optimal location is the corner solution of reporting only parameters that are nonparametrically point identified, such as the LATE.

Table 4 Weights for measures of selection

Quantity	Expression	Weights	
		$\omega_0^\star(u,x,z)$	$\omega_1^\star(u,x,z)$
Average selection bias	$E[Y_0\|D=1]-E[Y_0\|D=0]$	$\frac{\mathbb{1}[u\le p(x,z)]}{P[D=1]}-\frac{\mathbb{1}[u>p(x,z)]}{P[D=0]}$	0
Average selection on the level	$E[Y_1\|D=1]\ -E[Y_1\|D=0]$	0	$\frac{\mathbb{1}[u\le p(x,z)]}{P[D=1]}-\frac{\mathbb{1}[u>p(x,z)]}{P[D=0]}$
Average selection on the gain	$E[Y_1-Y_0\|D=1]\ -E[Y_1-Y_0\|D=0]$	$-\omega_1^\star(u,x,z)$	$\frac{\mathbb{1}[u\le p(x,z)]}{P[D=1]}-\frac{\mathbb{1}[u>p(x,z)]}{P[D=0]}$

5.8. Testable Implications

It is possible that no solution exists to the programs in Equation 26 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 (Equation 4) with Assumptions 1–3 is known to have testable implications (Balke & Pearl 1997, Imbens & Rubin 1997, Kitagawa 2015). However, 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

$$\Gamma_{\text{sel}}(m) \equiv E\left[\int_0^1 m_0(u,X)\left(\frac{\mathbb{1}[u \le p(X,Z)]}{P[D=1]} - \frac{\mathbb{1}[u > p(X,Z)]}{P[D=0]}\right) \mathrm{d}u\right] \qquad 31.$$

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 Equation 26 is empty when $\mathcal{M}$ is constrained in this way can be interpreted as evidence against the hypothesis that the treatment is exogenous. One could further restrict $\mathcal{M}$ to only contain m such that $\Gamma_{\text{gain}}(m) = 0$, where $\Gamma_{\text{gain}}(m)$ is defined as in Equation 31 using the weights for average selection on the gain given in **Table 4**. Finding the feasible set in Equation 26 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 Mogstad et al. (2017) framework discussed in Section 5 to several other approaches that have been used in the literature. We show that many of these approaches can be viewed as special cases of the general framework in which the set of admissible MTR functions, $\mathcal{M}$, is restricted to only contain 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.[17] 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_0 to z_1 Wald estimand (conditional on $X=x$) point identifies $\mathrm{MTE}(u,x)=\mathrm{MTE}(x)$ for all u, i.e.,

$$\begin{aligned}&\frac{E[Y|Z=z_1,X=x]-E[Y|Z=z_0,X=x]}{E[D|Z=z_1,X=x]-E[D|Z=z_0,X=x]}\\&=\frac{\int_{p(x,z_0)}^{p(x,z_1)}\left[m_1(u,x)-m_0(u,x)\right]\mathrm{d}u}{p(x,z_1)-p(x,z_0)}=\frac{\int_{p(x,z_0)}^{p(x,z_1)}\mathrm{MTE}(x)\,\mathrm{d}u}{p(x,z_1)-p(x,z_0)}=\mathrm{MTE}(x).\end{aligned}$$

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_0 to z_1 Wald estimand.[18]

As Heckman & 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 to assume that these individuals either do not know or do not act on their idiosyncratic differences in this causal effect. Consider the implications in our running example of the impact of a mosquito net on contracting malaria. The assumption that the causal effect of the mosquito net does not vary with unobservables is clearly a strong one, since it rules out heterogeneity in susceptibility and sleeping environments, which are known to be important. Given this, the MTE function will be constant in u only if individuals do not base their purchase decisions on these heterogeneous factors. The salience of mosquitos and malaria makes this assumption difficult to justify.

A dissenting opinion on the viability of assuming away unobserved heterogeneity in treatment effects is provided by Angrist & Fernández-Val (2013, p. 411), who argue that a version of this assumption, which they describe as "conditional effect ignorability," can be attractive.[19] We are not sympathetic to this view. Indeed, allowing for unobserved heterogeneity in the effect of D on

[17] These observations date back at least to Heckman & Robb (1985a,b). Angrist (2004) provides a modern revision.

[18] Using similar intuition, Angrist (2004) shows that, 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 explicitly parametric approaches discussed in Section 6.3.

[19] The assumption used by Angrist & Fernández-Val (2013) is that $\frac{\int_{p(x,z_0)}^{p(x,z_1)}\left[m_1(u,x)-m_0(u,x)\right]\mathrm{d}u}{p(x,z_1)-p(x,z_0)}=\int_0^1\left[m_1(u,x)-m_0(u,x)\right]\mathrm{d}u$ for all x and z. 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 this equation without making the stronger assumption.

Y is a key motivation in the modern program evaluation literature and one that 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 & Angrist (1994).

6.2. Separability of Observed and Unobserved Heterogeneity

In Section 4.2, we see 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

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

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 assumption 2 (see Section 2.2) to the assumption that $(V_0, V_1, U) \perp\!\!\!\perp (X, Z)$. Under this stronger independence assumption, we can write

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

which is an additively separable function of x and u. Returning to Equation 17, this implies that

$$\frac{\partial}{\partial u}\mathbf{E}[YD|p(x, Z) = u, X = x]\Big|_{u=\widetilde{u}} = \mu_1(x) + \mathbf{E}[V_1|U = \widetilde{u}], \tag{34.}$$

and similarly for $d = 0$.

Equation 34 shows that, under additive separability, variation in $P = p(X, Z)$ conditional on $X = x$ can be used to trace out the same function $\mathbf{E}[V_d|U = u]$, regardless of the value of 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.[20] 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 (see, e.g., Brinch et al. 2017; Carneiro & Lee 2009; Carneiro et al. 2011, 2016; Cornelissen et al. 2018; Eisenhauer et al. 2015; Kline & Walters 2016; Maestas et al. 2013). It is important to notice that 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 that does not need to obey the usual exclusion restriction. In applications, the types of variables usually included in X, such as sociodemographic controls, are unlikely to be exogenous in this way.

[20] For example, suppose that $\mu_1(x) = x'\tau_1$ is linear in parameters. Then, from Equation 6.2, 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 Equation 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 (for more details on this argument, see Carneiro et al. 2011).

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 Equation 33 can be achieved by writing Equation 32 and adding the assumption that $E[V_d|U,X] = E[V_d|U]$ to assumption 2 (Section 2.2). 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 show that separability can be imposed in the general Mogstad et al. (2017) framework by directly restricting the set of admissible MTR functions.

In some settings, the separability in Equation 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 individuals 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 et al. (2001, 2003) apply this approach to the binary treatment setting considered in this review. The case that they study, which is the most widely used, maintains Equation 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 cumulative distribution function (CDF).[21]

Under this assumption, Equation 33 reduces to

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

because 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, $p(x, z_1) \equiv \widetilde{u}_1 > \widetilde{u}_0 \equiv p(x, z_0)$, it follows from Equation 16 that

$$\begin{aligned} E[Y\,D|p(x,Z) = \widetilde{u}_1, X = x] - E[Y\,D|p(x,Z) = \widetilde{u}_0, X = x] \\ = \text{Corr}(V_1, U)\text{Var}(V_1) \int_{\widetilde{u}_0}^{\widetilde{u}_1} \Phi^{-1}(u)\,\mathrm{d}u, \end{aligned} \qquad 36.$$

and similarly for $d = 0$. This implies that the product $\text{Corr}(V_d, U)\text{Var}(V_d)$ is identified for $d = 0, 1$, and thus, that the functional form restriction in Equation 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 components.

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, $\text{Corr}(V_d, U)\text{Var}(V_d)$. Two points of exogenous variation, i.e., z_0 and z_1, are sufficient to

[21] Alternatively, and equivalently, the same assumption can be made about (V_d, U) using the prenormalized choice equation (Equation 2).

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 variable 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.

Moreover, normality in particular has some unattractive economic implications. As Carneiro et al. (2011, p. 2767) note, normality implies that the limit as u tends to 0 or 1 of the MTR functions (Equation 35) is necessarily $\pm\infty$, since $\lim_{u\to 0}\Phi^{-1}(u) = -\infty$ and $\lim_{u\to 1}\Phi^{-1}(u) = +\infty$. That is, the normal model implies that individuals unlikely to take treatment and those very likely to take treatment experience arbitrarily high or low causal effects of the treatment. In most settings, this is implausible on its face.

A more subtle property of normality is that it requires the MTE function to be monotone decreasing or increasing as a function of u, i.e., it imposes one direction of the monotone treatment selection condition discussed in Section 5.5. While potentially appealing in some situations, some authors have found settings in which this assumption does not appear to hold. For example, Brinch et al. (2017) find evidence of a U-shaped MTE function for the causal effect of having an additional child on the educational outcomes of older children when using sex composition as an instrument. One attraction of the framework discussed in Section 5 is that it allows one to decouple parametric shape restrictions from fundamentally nonparametric restrictions such as monotonicity.

There are other parametric approaches that yield the same payoff as normality but that do not share all of these negative features and that can arguably be easier to interpret. For example, suppose that, instead of using Equation 35, we assume that $m_d(u, x)$ is linear as a function of u 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 Equation 16, we have

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

with a similar expression holding for $d = 0$. Since we have $P[D = 1|p(x, Z) = u, X = x] = u$ 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 Equation 38 with two values $\widetilde{u}_1 \equiv p(x, z_1) \neq p(x, z_0) = \widetilde{u}_0$ 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 Equation 36.

This linearity assumption was first suggested by Brinch et al. (2012).[22] The assumption yields point identification through effectively the same extrapolation argument as bivariate normality. Linearity has a straightforward interpretation: Holding $X = x$ fixed, a 1-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

[22] Kowalski (2016) provides a more recent application of the same idea.

considered on a case-by-case basis.[23] 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.[24] 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.

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 $\mathcal{M}_S$ in Equation 24 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 nonredundant IV-like estimands to exactly pin down a single $m \in \mathcal{M}$ is a generalization to the arguments in Equations 36 and 38. As always, whether such a specification is possible depends on both the richness of the data, i.e., how many distinct IV-like estimands can be found, and 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 et al. (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 individual'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.[25]

Chernozhukov & Hansen (2005) show that rank invariance can be used to point identify the ATE under a somewhat nonstandard relevance condition for the relationship between D and Z. Their model does not impose the choice equation (Equation 4). Vuong & Xu (2017) show that also imposing Equation 4 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

[23] 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 (Equation 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 Equation 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 as for the efficiency of statistical inference. A more direct comparison would be between Equation 35 and Equation 37 as different restrictions on the forms of the MTR functions.

[24] These observations are related to proposed series estimators of the local IV estimand (Equation 17), as in the work of Moffitt (2008) and French & 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 Equation 38.

[25] 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 & Yıldız (2007), Chesher (2010), Shaikh & Vytlacil (2011), Bhattacharya et al. (2012), Machado et al. (2013), Mourifié (2015), and Torgovitsky (2017b).

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 observed $Y = Y_1$ 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 due primarily to Manski (1989, 1990, 1994, 1997, 2003) and Manski & Pepper (2000, 2009).[26] Unlike Manski's work on IV methods, the Mogstad et al. (2017) approach maintains the choice Equation 4.[27] Maintaining some form of choice model (not necessarily Equation 4) is indispensable for evaluating the effects of policy interventions that do not mandate treatment or nontreatment.[28] As we argue in Section 3, we view such policies as being typical of interesting counterfactual questions in economic applications.

Another way in which the Mogstad et al. (2017) framework departs from Manski's work is more practical. Instead of deriving explicit expressions for bounds, the Mogstad et al. (2017) framework provides a computational characterization of bounds. 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. Such 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 specify 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 specify the MTR functions as ninth-degree polynomials. This narrows the bounds considerably by ruling out the discontinuous MTR functions that are permitted in **Figure 7**. We view this as attractive for many applications, since these discontinuous functions are unlikely to represent important cases to guard against in typical economic settings. However, analytic expressions for the bounds under a ninth-degree polynomial are unknown and seem difficult to derive. Using Mogstad et al.'s (2017) computational approach, this derivation is not necessary, and the bounds are returned almost instantaneously using standard software.

[26] Tamer (2010) provides an historical perspective on this literature.

[27] Incidentally, the Imbens & Angrist (1994) monotonicity assumption underlying the choice equation (Equation 4) is exactly Manski's (1997) monotone treatment response assumption, but applied to the counterfactual relationship between Z and D rather than that between D and Y.

[28] An interesting result due to Heckman & Vytlacil (2001b) shows that, when the implications of the choice model (Equation 4) are not rejected (see Section 5.8), the choice model has no impact on the sharp nonparametric bounds for the ATE derived by Manski (1994). Balke & Pearl (1997) and Kitagawa (2009) find related results. 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.

7. CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

Above, we discuss 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 review these approaches and argue that their reliance on point identification is a weakness. We discuss a general framework, developed fully by 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. F407–9). We view computational methods, such as the one discussed in Section 2, as important tools for answering this criticism. The flexibility of the Mogstad et al. (2017) method means that a researcher can smoothly adjust their policy question (target parameter), or the assumptions that they are willing to maintain, in a way that approaches point identification as a special case. As a result, the tightness of the bounds that the researchers 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.[29]

There are many avenues down which Mogstad et al.'s (2017) approach to identification and extrapolation of treatment effects can be further developed. While we focus 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 and involves many complications.[30] The issue of policy relevance and the corresponding 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 in this review to help ameliorate the local nature of regression discontinuity designs.[31] Similar ideas could potentially also be applied to more complicated evaluation settings involving dynamics, mediation, peer effects, or other challenges for identification.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Christian Brinch, Derek Neal, Jack Mountjoy, and an anonymous reviewer for useful comments and suggestions. The authors' research is supported in part by National Science Foundation grant SES-1426882.

[29]For a devoted defense of faith-based extrapolation, the reader is referred to Angrist (2016).

[30]For discussions of methods for multiple discrete treatments, the reader is referred to Angrist & Imbens (1995), Heckman et al. (2006), Heckman & Vytlacil (2007b), Heckman & Urzua (2010), Kirkeboen et al. (2016), Lee & Salanié (2016), and Heckman & Pinto (2016), among others. Methods for continuous treatments have been considered by Angrist et al. (2000), Chesher (2003), Florens et al. (2008), Imbens & Newey (2009), Torgovitsky (2015, 2017a), Masten (2015), and Masten & Torgovitsky (2016), among others.

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

LITERATURE CITED

Abadie A, Angrist J, Imbens G. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70:91–117

Angrist JD. 2004. Treatment effect heterogeneity in theory and practice. *Econ. J.* 114:C52–83

Angrist JD. 2016. *Sometimes you get what you need: discussion of Mogstad, Santos, and Torgovitsky*. Slides presented at the NBER Labor Studies Meeting, Cambridge, MA, July 25–29

Angrist JD, Evans WN. 1998. Children and their parents' labor supply: evidence from exogenous variation in family size. *Am. Econ. Rev.* 88:450–77

Angrist JD, Fernández-Val I. 2013. ExtrapoLATE-ing: external validity and overidentification in the LATE framework. In *Advances in Economics and Econometrics*, ed. D Acemoglu, M Arellano, E Dekel, pp. 401–34. Cambridge, UK: Cambridge Univ. Press

Angrist JD, Graddy K, Imbens GW. 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Rev. Econ. Stud.* 67:499–527

Angrist JD, Imbens GW. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J. Am. Stat. Assoc.* 90:431–42

Angrist JD, Krueger AB. 1991. Does compulsory school attendance affect schooling and earnings? *Q. J. Econ.* 106:979–1014

Angrist JD, Rokkanen M. 2015. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *J. Am. Stat. Assoc.* 110:1331–44

Balke A, Pearl J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92:1171–76

Bhattacharya J, Shaikh AM, Vytlacil E. 2012. Treatment effect bounds: an application to Swan-Ganz catheterization. *J. Econom.* 168:223–43

Bitler MP, Gelbach JB, Hoynes HW. 2006. What mean impacts miss: distributional effects of welfare reform experiments. *Am. Econ. Rev.* 96:988–1012

Bitler MP, Hoynes HW, Domina T. 2014. *Experimental evidence on distributional effects of Head Start*. NBER Work. Pap. 20434

Björklund A, Moffitt R. 1987. The estimation of wage gains and welfare gains in self-selection models. *Rev. Econ. Stat.* 69:42–49

Black SE, Devereux PJ, Salvanes KG. 2005. The more the merrier? The effect of family size and birth order on children's education. *Q. J. Econ.* 120:669–700

Brinch CN, Mogstad M, Wiswall M. 2012. *Beyond LATE with a discrete instrument*. Discuss. Pap. 703, Stat. Norway, Oslo

Brinch CN, Mogstad M, Wiswall M. 2017. Beyond LATE with a discrete instrument. *J. Political Econ.* 125:985–1039

Buchinsky M. 1994. Changes in the US wage structure 1963–1987: application of quantile regression. *Econometrica* 62:405–58

Carneiro P, Heckman JJ, Vytlacil E. 2010. Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica* 78:377–94

Carneiro P, Heckman JJ, Vytlacil EJ. 2011. Estimating marginal returns to education. *Am. Econ. Rev.* 101:2754–81

Carneiro P, Lee S. 2009. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *J. Econom.* 149:191–208

Carneiro P, Lokshin M, Umapathi N. 2016. Average and marginal returns to upper secondary schooling in Indonesia. *J. Appl. Econom.* 32:16–36

Chen X. 2007. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, Vol. 6, ed. JJ Heckman, EE Leamer, pp. 5549–632. Amsterdam: Elsevier

Chernozhukov V, Hansen C. 2005. An IV model of quantile treatment effects. *Econometrica* 73:245–61

Chernozhukov V, Newey WK, Santos A. 2015. Constrained conditional moment restriction models. arXiv:1509.06311 [math.ST]

Chesher A. 2003. Identification in nonseparable models. *Econometrica* 71:1405–41

Chesher A. 2010. Instrumental variable models for discrete outcomes. *Econometrica* 78:575–601

Cornelissen T, Dustmann C, Raute A, Schönberg U. 2018. Who benefits from universal childcare? Estimating marginal returns to early childcare attendance. *J. Political Econ.* In press

Dong Y, Lewbel A. 2015. Identifying the effect of changing the policy threshold in regression discontinuity models. *Rev. Econ. Stat.* 97:1081–92

Doyle JJ Jr. 2007. Child protection and child outcomes: measuring the effects of foster care. *Am. Econ. Rev.* 97:1583–610

Dupas P. 2014. Short-run subsidies and long-run adoption of new health products: evidence from a field experiment. *Econometrica* 82:197–228

Dupas P, Hoffmann V, Kremer M, Zwane AP. 2016. Targeting health subsidies through a non-price mechanism: a randomized controlled trial in Kenya. *Science* 353:889–95

Eisenhauer P, Heckman JJ, Vytlacil E. 2015. The generalized Roy model and the cost-benefit analysis of social programs. *J. Political Econ.* 123:413–43

Evans W, Ringel JS. 1999. Can higher cigarette taxes improve birth outcomes? *J. Public Econ.* 72:135–54

Felfe C, Lalive R. 2014. *Does early child care help or hurt children's development?* Tech. Rep. 8484, Inst. Labor Econ., Bonn, Ger.

Firpo S, Fortin NM, Lemieux T. 2009. Unconditional quantile regressions. *Econometrica* 77:953–73

Florens JP, Heckman JJ, Meghir C, Vytlacil E. 2008. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica* 76:1191–206

French E, Song J. 2014. The effect of disability insurance receipt on labor supply. *Am. Econ. J. Econ. Policy* 6:291–337

Gronau R. 1974. Wage comparisons: a selectivity bias. *J. Political Econ.* 82:1119–43

Havnes T, Mogstad M. 2015. Is universal child care leveling the playing field? *J. Public Econ.* 127:100–14

Heckman JJ. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42:679–94

Heckman JJ. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* 5:475–92

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61

Heckman JJ. 1996. Comment. *J. Am. Stat. Assoc.* 91:459–62

Heckman JJ. 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *J. Hum. Resour.* 32:441–62

Heckman JJ. 2001. Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. *J. Political Econ.* 109:673–748

Heckman JJ. 2010. Building bridges between structural and program evaluation approaches to evaluating policy. *J. Econ. Lit.* 48:356–98

Heckman JJ, Pinto R. 2016. *Unordered monotonicity*. NBER Work. Pap. 23497

Heckman JJ, Robb R. 1985a. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, ed. JJ Heckman, B Singer, pp. 156–245. Cambridge, UK: Cambridge Univ. Press

Heckman JJ, Robb R. 1985b. Alternative methods for evaluating the impact of interventions: an overview. *J. Econom.* 30:239–67

Heckman JJ, Smith JA. 1998. *Evaluating the welfare state*. NBER Work. Pap. 6542

Heckman JJ, Smith JA, Clements N. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud.* 64:487–535

Heckman JJ, Tobias JL, Vytlacil E. 2001. Four parameters of interest in the evaluation of social programs. *South. Econ. J.* 68:210–23

Heckman JJ, Tobias JL, Vytlacil E. 2003. Simple estimators for treatment parameters in a latent-variable framework. *Rev. Econ. Stat.* 85:748–55

Heckman JJ, Urzua S. 2010. Comparing IV with structural models: what simple IV can and cannot identify. *J. Econom.* 156:27–37

Heckman JJ, Urzua S, Vytlacil E. 2006. Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Stat.* 88:389–432

Heckman JJ, Vytlacil EJ. 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *PNAS* 96:4730–34

Heckman JJ, Vytlacil EJ. 2001a. Policy-relevant treatment effects. *Am. Econ. Rev.* 91:107–11

Heckman JJ, Vytlacil EJ. 2001b. Instrumental variables, selection models, and tight bounds on the average treatment effect. In *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. M Lechner, F Pfeiffer, pp. 1–15. Heidelberg, Ger.: Physica

Heckman JJ, Vytlacil EJ. 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. KMC Hsiao, J Powell, pp. 1–46. Cambridge, UK: Cambridge Univ. Press

Heckman JJ, Vytlacil EJ. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73:669–738

Heckman JJ, Vytlacil EJ. 2007a. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In *Handbook of Econometrics*, Vol. 6, ed. JJ Heckman, EE Leamer, pp. 4779–874. Amsterdam: Elsevier

Heckman JJ, Vytlacil EJ. 2007b. 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*, Vol. 6, ed. JJ Heckman, EE Leamer, pp. 4875–5143. Amsterdam: Elsevier

Hull P. 2016. *Estimating hospital quality with quasi-experimental data*. Work. Pap., Dep. Econ., Mass. Inst. Technol., Cambridge, MA

Imbens GW. 2010. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* 48:399–423

Imbens GW. 2013. Book review feature: *Public Policy in an Uncertain World*: by Charles F. Manski (Cambridge, MA: Harvard University Press. pp. 224, $39.95. ISBN: 978-0674066892). *Econ. J.* 123:F401–11

Imbens GW, Angrist JD. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–75

Imbens GW, Newey WK. 2009. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77:1481–512

Imbens GW, Rubin DB. 1997. Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* 64:555–74

Kirkeboen LJ, Leuven E, Mogstad M. 2016. Field of study, earnings, and self-selection. *Q. J. Econ.* 131:1057–111

Kitagawa T. 2009. *Identification region of the potential outcome distributions under instrument independence*. Work. Pap., Cent. Microdata Methods Pract., London

Kitagawa T. 2015. A test for instrument validity. *Econometrica* 83:2043–63

Kline P, Walters CR. 2016. Evaluating public programs with close substitutes: the case of Head Start. *Q. J. Econ.* 131:1795–848

Kline P, Walters CR. 2017. *Through the looking glass: Heckits, LATE, and numerical equivalence*. Work. Pap., Dep. Econ., Univ. Calif., Berkeley

Koenker R. 2005. *Quantile Regression*. Cambridge, UK: Cambridge Univ. Press

Kowalski A. 2016. *Doing more when you're running LATE: applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments*. NBER Work. Pap. 22363

Lee S, Salanié B. 2016. *Identifying effects of multivalued treatments*. Work. Pap., Dep. Econ., Columbia Univ., New York

Machado C, Shaikh AM, Vytlacil EJ. 2013. *Instrumental variables and the sign of the average treatment effect*. Work. Pap., Dep. Econ., Univ. Chicago

Maestas N, Mullen KJ, Strand A. 2013. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *Am. Econ. Rev.* 103:1797–829

Manski CF. 1989. Anatomy of the selection problem. *J. Hum. Resour.* 24:343–60

Manski CF. 1990. Nonparametric bounds on treatment effects. *Am. Econ. Rev.* 80:319–23

Manski CF. 1994. The selection problem. In *Advances in Econometrics: Sixth World Congress*, ed. CA Sims, pp. 143–70. Cambridge, UK: Cambridge Univ. Press

Manski CF. 1997. Monotone treatment response. *Econometrica* 65:1311–34

Manski CF. 2003. *Partial Identification of Probability Distributions*. Berlin: Springer

Manski CF, Pepper JV. 2000. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 68:997–1010

Manski CF, Pepper JV. 2009. More on monotone instrumental variables. *Econom. J.* 12:S200–16

Masten MA. 2015. *Random coefficients on endogenous variables in simultaneous equations models*. Work. Pap. 25/15, Cent. Microdata Methods Pract., London

Masten MA, Torgovitsky A. 2016. Identification of instrumental variable correlated random coefficients models. *Rev. Econom. Stat.* 98:1001–5

Matzkin RL. 2007. Nonparametric identification. In *Handbook of Econometrics*, Vol. 6, ed. JJ Heckman, EE Leamer, pp. 5307–68. Amsterdam: Elsevier

Miguel E, Satyanath S, Sergenti E. 2004. Economic shocks and civil conflict: an instrumental variables approach. *J. Political Econ.* 112:725–53

Moffitt R. 2008. Estimating marginal treatment effects in heterogeneous populations. *Ann. Econ. Stat.* 91/92:239–61

Mogstad M, Santos A, Torgovitsky A. 2017. *Using instrumental variables for inference about policy relevant treatment parameters*. NBER Work. Pap. 23568

Mourifié I. 2015. Sharp bounds on treatment effects in a binary triangular system. *J. Econom.* 187:74–81

Nybom M. 2017. The distribution of lifetime earnings returns to college. *J. Labor Econ.* 35:903–52

Robinson PM. 1988. Root-*N*-consistent semiparametric regression. *Econometrica* 56:931–54

Rokkanen M. 2015. *Exam schools, ability, and the effects of affirmative action: latent factor extrapolation in the regression discontinuity design*. Work. Pap., Dep. Econ., Columbia Univ., New York

Shaikh AM, Vytlacil EJ. 2011. Partial identification in triangular systems of equations with binary dependent variables. *Econometrica* 79:949–55

Tamer E. 2010. Partial identification in econometrics. *Annu. Rev. Econ.* 2:167–95

Torgovitsky A. 2015. Identification of nonseparable models using instruments with small support. *Econometrica* 83:1185–97

Torgovitsky A. 2017a. Minimum distance from independence estimation of nonseparable instrumental variables models. *J. Econom.* 199:35–48

Torgovitsky A. 2017b. *Partial identification by extending subdistributions*. Work. Pap., Univ. Chicago

Vuong Q, Xu H. 2017. Counterfactual mapping and individual treatment effects in nonseparable models with binary endogeneity. *Quant. Econ.* 8:589–610

Vytlacil E. 2002. Independence, monotonicity, and latent index models: an equivalence result. *Econometrica* 70:331–41

Vytlacil E, Yldz N. 2007. Dummy endogenous variables in weakly separable models. *Econometrica* 75:757–79

Walters C. 2014. *The demand for effective charter schools*. NBER Work. Pap. 20640

Wing C, Cook TD. 2013. Strengthening the regression discontinuity design using additional design elements: a within-study comparison. *J. Policy Anal. Manag.* 32:853–77