ROCKWOOL FOUNDATION
BERLIN
INSTITUTE FOR THE ECONOMY AND THE FUTURE OF WORK

CReAM | Centre for Research & Analysis of Migration

DISCUSSION PAPER SERIES 30/24

Instrumental Variables with Unobserved Heterogeneity in Treatment Effects

Magne Mogstad, Alexander Torgovitsky

DECEMBER 2024

ROCKWOOL Foundation Berlin
www.rfberlin.com

Centre for Research & Analysis of Migration
www.cream-migration.org

Instrumental Variables with Unobserved Heterogeneity in Treatment Effects*

Magne Mogstad† Alexander Torgovitsky‡

August 27, 2024

Abstract

This chapter synthesizes and critically reviews the modern instrumental variables (IV) literature that allows for unobserved heterogeneity in treatment effects (UHTE). We start by discussing why UHTE is often an essential aspect of IV applications in economics and we explain the conceptual challenges raised by allowing for it. Then we review and survey two general strategies for incorporating UHTE. The first strategy is to continue to use linear IV estimators designed for classical constant (homogeneous) treatment effect models, acknowledge their likely misspecification, and attempt to reverse engineer an attractive interpretation in the presence of UHTE. This strategy commonly leads to interpretations of linear IV that involve local average treatment effects (LATEs). We review the various ways in which the use and justification of LATE interpretations have expanded and contracted since their introduction in the early 1990s. The second strategy is to forward engineer new estimators that explicitly allow for UHTE. This strategy has its roots in the Gronau-Heckman selection model of the 1970s, ideas from which have been revitalized through marginal treatment effect (MTE) analysis. We discuss implementation of MTE methods and draw connections with related control function and bounding methods that are scattered throughout the econometric and causal inference literature.

*Prepared for the *Handbook of Labor Economics*. We thank Deniz Dutz, Koichiro Ito, Pat Kline, Matt Masten, Vitor Possebom, Evan Rose, Henrik Sigstad, Tymon Słoczyński, Winnie van Dijk, and Thomas Wiemann for helpful discussions and comments. We thank Koichiro Ito and Evan Rose for providing data. Ian Xu provided excellent research assistance.

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

‡Kenneth C. Griffin Department of Economics, University of Chicago. Research supported by National Science Foundation grant SES-1846832.

Contents

1 Introduction

Instrumental variable (IV) methods are fundamental to causal inference in economics. They are now also widely used across the social and biological sciences. Their attraction lies in allowing for unobserved confounders, which arise generically in economic applications due to private information, preference heterogeneity, and simultaneity, among other reasons. This chapter synthesizes and critically reviews the literature on modern IV methods that allow for unobserved heterogeneity in treatment effects (UHTE).

In Section 2, we briefly review the basic ideas behind IV methods. We argue that UHTE is a generic feature of many economic applications, especially those in labor economics. The classical linear IV model found in textbooks does not allow for UHTE. We clarify the problems created by this misspecification and we outline the conceptual trade-offs associated with various ways of solving these problems.

The rest of the chapter is then organized into two parts, reflecting the two main approaches to incorporating UHTE into IV models.

The first approach, which was pioneered by Imbens and Angrist (1994), is to interpret linear IV estimators designed for the classical linear IV model through the lens of a nonparametric IV model that allows for UHTE. Based on their results, it has become increasingly common in the empirical literature to describe linear IV estimators, such as two-stage least squares (2SLS), as reflecting local average treatment effects (LATEs). This interpretation is derived from a baseline setup with a binary treatment, a binary instrument, and no covariates, a setup which does not characterize most empirical work in practice.

In Section 3, we provide a comprehensive survey of how the LATE interpretation is affected by moving away from the baseline setup. We find that it is remarkably specific to the baseline setup. Deviating from the baseline setup by having a multivalued treatment, multivalued instrument, or by linearly controlling for covariates complicates, qualifies, or breaks the widespread interpretation that "linear IV is LATE."

The interpretation problems we point to are orthogonal to the debate over whether LATEs are interesting objects, a debate which has been had many times before. Instead, the problems stem from the now-widespread methodological practice of trying to provide a misspecification-robust interpretation for a commonly-used estimator in the context of a less restrictive model for which it was not designed. We call this practice reverse engineering because it starts with an estimator rather than starting with a model. Reverse engineering arguments are increasingly fashionable in microeconometrics, having been applied to selection on observables (Angrist, 1998), difference-in-differences

and two-way fixed effects (e.g. Goodman-Bacon, 2021; Sun and Abraham, 2021), settings with multivalued treatments (Goldsmith-Pinkham et al., 2024), and regression discontinuity and kink designs (Lee, 2008; Card et al., 2015; Cattaneo et al., 2016). Our discussion in Section 3 shows that reverse engineering arguments for linear IV estimators are brittle.

The second approach, which has a longer and more diffuse history, is to forward engineer estimators of specific target parameters in models that explicitly allow for UHTE. This includes methods of directly estimating LATEs with estimators other than linear IV. It also includes the classical selection model developed by Gronau (1974) and Heckman (1974, 1976, 1979), and its nonparametric reincarnation in terms of the marginal treatment effect (Heckman and Vytlacil, 1999, 2005).

Section 4 is devoted to surveying these forward engineering approaches. We discuss methods for estimating unconditional LATEs that avoid some of the pitfalls of reverse engineering encountered with linear IV. We then discuss a practical linear regression framework for conducting marginal treatment effect (MTE) analysis with binary treatments. In doing so we emphasize the underappreciated point, formalized by Vytlacil (2002), that the separable threshold-crossing model used in both the Gronau-Heckman selection model and modern MTE analysis imposes exactly the same "monotonicity condition" about selection as the model used by Imbens and Angrist (1994), just with different notation. The benefit of the MTE analysis is that it provides a vehicle for both clearly stating the target parameter and for imposing additional assumptions to aid in estimating it. We show how the linear regression framework for MTE extends to ordered and unordered treatments, even as the equivalence with the monotonicity condition is lost. The key theme that emerges is the model of treatment selection and under what assumptions it is identified. We contrast these selection model methods to those that allow for UHTE but do not impose restrictions on how the instrument affects treatment choice.

In Section 5, we distill our discussion into a list of recommendations for researchers using IV methods. Section 6 provides some brief concluding remarks. Example Stata and `R` code for implementing some of the main methods we discuss in the chapter is available at `https://a-torgovitsky.github.io/ivhandbook/`.

2 Background

In this section we provide some brief background on IV methods with an emphasis on the motivation for incorporating unobserved heterogeneity in treatment effects (UHTE).

2.1 IV in a nutshell

IV methods are used to estimate the causal effect of one variable, the treatment, on another variable, the outcome. The motivating concern is that the treatment is endogenous in the sense that it covaries with other unmeasured factors that are associated with the outcome. The association between the treatment and outcome conflates the effect of the treatment with these unmeasured factors. An IV method instead focuses on the association between the outcome and a third variable, the instrumental variable, or instrument for short. If the instrument is associated with the treatment, but not with the unmeasured factors, and if it has no direct effect on the outcome itself, then the association between the instrument and the outcome should only reflect the causal effect of the treatment on the outcome.

This line of reasoning relies on three assumptions that all IV methods invoke to one extent or another: exclusion, exogeneity, and relevance. Exclusion means that the instrument itself has no direct effect on the outcome. Exogeneity means that it is not associated with any unmeasured factors that are associated with the outcome. Relevance means that the instrument is associated with the treatment.

The exclusion and exogeneity assumptions are often controversial in practice. The purpose of this chapter is not to litigate their merits either in general or in specific applications. The enormous body of published empirical work using IV methods suggests that at least some researchers find these assumptions reasonable in at least some applications. Our focus instead is on how these assumptions can be implemented while also allowing for the possibility of unobserved heterogeneity in treatment effects (UHTE), meaning systematic variation in the effect of the treatment on the outcome that persists even after controlling for other observable variables.

2.2 Why is there unobserved heterogeneity in treatment effects?

UHTE creates many complications in IV methods. These complications can be entirely avoided by assuming that treatment effects are either constant (homogeneous) or, slightly more generally, idiosyncratic in the sense of being unassociated with the treatment variable. So before diving in, we should take a moment to reflect on why such an assumption is often unpalatable in empirical economics.

A workhorse example from labor economics illustrates the issues clearly. Suppose that the treatment variable is college attendance and the outcome variable is a labor market outcome, such as subsequent earnings (e.g. Card, 1999). The classic endogeneity concern is that there are unmeasured factors, often described loosely as "ability," that are correlated with both educational attainment and labor market performance (e.g.

Becker, 1964; Griliches, 1977). This description is not particularly helpful because it obscures the role of choice; attending college is a choice and individuals make choices purposefully.

A more compelling explanation of the endogeneity problem is that individuals are heterogeneous in their anticipated returns to college due to unobserved private information about their skills, aptitudes, or outside options, and they use this information when making their attendance choices.[1] For example, some individuals have an aptitude for abstract reasoning that translates into strong labor market performance only with a college education. Other individuals have an aptitude for trade skills (welder, plumber, carpenter) that are equally remunerative with or without a college education. Individuals choose whether to attend college at least in part because of its anticipated effect on their future earnings. This explanation results in UHTE that is systematically related to the treatment variable itself: those who choose to attend college tend to be those who would benefit from it.

The essential ingredients of this story are common for causal inference questions involving human actors. Interesting treatment variables are often choices. Interesting outcome variables often reflect substantive consequences for the human beings under consideration. Human beings don't make choices randomly; they likely consider, at least in part, the effect that their choices may have on the outcome. These choices then become treatment variables that are associated with their effects on the outcome. Unless there's a compelling domain-specific reason to believe that the effect of the treatment cannot vary for some physical or institutional reason, then there will be UHTE that is systematically associated with the observed treatment choices.

2.3 From the classical linear IV model to potential outcomes

Classical textbook treatments of IV (e.g. Theil (1971), Wooldridge (2010)) start with an equation like

$$Y_i = \alpha_0 + \alpha_1 D_i + \epsilon_i, \tag{1}$$

where Y_i is the outcome variable, D_i is the treatment variable, and ϵ_i is an unobservable that collects all other unmeasured factors in Y_i. The instrument, Z_i, does not appear in this equation due to the exclusion assumption. The exogeneity assumption is that Z_i is uncorrelated with ϵ_i. The relevance assumption is that Z_i is correlated with D_i. All variables are indexed by a unit of observation, i, which we will think of as an individual

[1] Becker (1967), Willis and Rosen (1979), and Card (2001) develop models with this property. The following is adapted from Willis and Rosen (1979, pp. S10–S11).

for concreteness.

Equation (1) says that a one unit increase in D_i—holding all else ϵ_i fixed—causes a change of α_1 in Y_i for everyone (all i). This is restrictive in two ways: (i) it implies that the treatment effect of D_i on Y_i is linear, and (ii) it rules out heterogeneous treatment effects. In much of the literature and much of this chapter, it is assumed that D_i is binary (takes values 0 or 1), in which case (i) is not restrictive. Our focus is on relaxing (ii).

One way to relax (ii) is to allow for treatment effects to vary with some observable covariates, X_i. For example, (1) could be augmented to

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i + \alpha_3 X_i D_i + \epsilon_i, \tag{2}$$

so that the causal effect of D_i on Y_i is now $\alpha_1 + \alpha_3 X_i$, which can vary with i through $\alpha_3 X_i$. Equation (2) allows for observed heterogeneity in treatment effects, but not for unobserved heterogeneity (UHTE). So this relaxation of (ii) does not address self-selection created by factors such as private information and heterogeneity in skills or preferences. For many economic applications—especially those that appeal to an IV strategy—it is precisely the unobserved heterogeneity that is the concern.

Relaxing (ii) to allow for UHTE instead requires interacting the treatment variable with latent variables. One way to do this is to postulate a relationship like

$$Y_i = f(D_i, \epsilon_i), \tag{3}$$

for some function f. Such a relationship is called "nonseparable" because D_i and ϵ_i are not additively separable, as in (1). Unlike the classical model (1), a nonseparable relationship allows for unobserved heterogeneity in treatment effects because, $f(d', \epsilon_i) - f(d, \epsilon_i)$ still depends on ϵ_i.

Comparing $f(d, \epsilon_i)$ to $f(d', \epsilon_i)$ involves the mental exercise of considering the value that the outcome would have taken if the treatment variable had been fixed at a potentially counterfactual value d or d' while keeping all other factors ϵ_i the same. This exercise can be represented succinctly in the potential outcomes notation introduced by Neyman (republished as Splawa-Neyman et al., 1990) for experiments and ported to observational settings by Rubin (1974).[2] In potential outcomes notation, the function f and unobservable ϵ_i are replaced by a collection of potential outcomes $Y_i(d)$, one for each value that the treatment D_i can conceivably take.[3] Each potential outcome $Y_i(d)$

[2] Heckman and Vytlacil (2007a) detail the many other authors across numerous disciplines who have independently invented similar notation. This is perhaps a testament to its intuitive appeal.

[3] The exclusion restriction is already implicitly embedded in this form of potential outcomes notation. To

is a random variable that answers the same counterfactual question as $f(d, \epsilon_i)$: what would Y_i have been had D_i been fixed to d, keeping all other factors the same? The outcome actually observed, Y_i, corresponds to the potential outcome of the observed treatment state, $Y_i = Y_i(D_i)$, while all other potential outcomes are unobserved for individual i. When D_i is binary, this is often written as

$$Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1), \tag{4}$$

which is the potential outcomes analog of (3).

Some researchers have strong opinions about the merits of working with notation involving latent variables and nonseparable models versus working with potential outcomes notation. Sometimes these opinions seem to border on suggesting that the notation itself has some special powers. As we show in Appendix A, the difference between the two notations is indeed fully notational: every model written in form (3) implies one written in form (4), and conversely. Good notation is essential for clearly communicating arguments and assumptions. But at the end of the day, it is just notation.

In this chapter, we will use both types of notation. Our default is to use potential outcomes, which tends to be simpler for models that make fewer assumptions. As we will see, however, the challenges created by UHTE often demand more assumptions. Latent variable notation turns out to be useful for this purpose. A good example of the relationship between potential outcomes and latent variables arises when thinking about models of treatment selection or selection models.

2.4 Selection models

The reason that UHTE complicates IV methods is that it matters "who" takes treatment. Individuals with different treatment effects will also tend to make different treatment choices. And if the instrument indeed affects treatment choice, then the distribution of treatment effects conditional on the treatment will further vary conditional on the instrument. Modeling the selection process of how the instrument affects treatment provides a way to keep track of this relationship and restrict it through additional assumptions.

Modeling selection requires taking a stance on the dimensions of the treatment D_i and instrument Z_i. Consider the simplest setting in which both D_i and Z_i are binary (0 or 1).

be more explicit, one could start by postulating potential outcomes $Y_i(d, z)$, state the exclusion restriction as $Y_i(d, z) = Y_i(d, z')$ for all z and z', and then define $Y_i(d) \equiv Y_i(d, z)$.

A large body of research pioneered by Gronau (1974), Lewis (1974), Heckman (1974, 1976, 1979), Willis and Rosen (1979) and others, modeled selection with a threshold-crossing model like

$$D_i = \mathbb{1}[V_i \leq \gamma Z_i], \tag{5}$$

where γ is an unknown parameter, and V_i is a continuously distributed latent variable.[4] Empirical implementations of this model often incorporate additional control variables X_i, necessitating some functional form assumptions, notably on the distribution of V_i, which is often taken to be normally distributed. These parameterizations should be understood as specific practical implementation choices rather than an assumption inherent to (5).

Imbens and Angrist (1994) took an ostensibly different approach to modeling selection by applying the potential outcomes notation to potential *treatments* that vary with the instrument. We denote these $D_i(z)$ for $z = 0$ and $z = 1$, so that

$$D_i = (1 - Z_i)D_i(0) + Z_i D_i(1), \tag{6}$$

in analogy with (4). With a binary treatment there are four configurations of the pair $(D_i(0), D_i(1))$, which can be thought of as individual i's choice group. Angrist et al. (1996) later described these groups as never-takers, always-takers, compliers, and defiers, for $(D_i(0), D_i(1)) = (0, 0)$, $(1, 1)$, $(0, 1)$, and $(1, 0)$, respectively.

Imbens and Angrist (1994) assumed that either $D_i(1) \geq D_i(0)$ for all i or $D_i(0) \geq D_i(1)$ for all i, a condition they described as monotonicity. An alternative way to state the condition, which ends up being easier to extend and modify, is in terms of probability: $\mathbb{P}[D_i(1) \geq D_i(0)] = 1$ or $\mathbb{P}[D_i(0) \geq D_i(1)] = 1$. There is no substantive difference between these two formulations. The monotonicity condition implies that there are either no defiers or no compliers. It is usually reasonable to simplify this to the assumption of no defiers, since there are few situations in which monotonicity is a compelling assumption but the direction of monotonicity is not known.

How does the Imbens and Angrist (1994) potential choices model with monotonicity differ from the classical threshold-crossing model (5)? It doesn't. A causal interpretation of (5), in which V_i represents "all other factors," implies that treatment choice

[4] In the early literature, the focus was typically on a one-sided selection problem where D_i indicated whether Y_i was observed for individual i (Heckman, 1976, is an exception). This problem is a simplified counterpart to evaluating the causal effect of a binary treatment, which can be seen as a two-sided selection problem. Models like (5) are now frequently described as Roy models after Roy (1951); see Heckman and Honoré (1990) for a discussion of the rationale.

would be given by $D_i(0) = \mathbb{1}[V_i \leq 0]$ if $Z_i = 0$, and by $D_i(1) = \mathbb{1}[V_i \leq \gamma]$ if $Z_i = 1$. Since the same latent variable V_i appears in both implied treatment choices, this implies that either $\mathbb{P}[D_i(1) \geq D_i(0)] = 1$, if $\gamma > 0$, or the reverse if $\gamma \leq 0$, which is exactly the monotonicity condition. Conversely, given potential choices, one can always construct a threshold-crossing model of form (5) that implies exactly the same potential choices by constructing V_i according to individual i's group. Assuming that $\gamma \geq 0$, make every always-taker have $V_i \leq 0$, every complier have V_i in $(0, \gamma]$, and every never-taker have $V_i > \gamma$ (see Figure 1). The threshold-crossing model (5) is therefore *equivalent* to the Imbens and Angrist (1994) potential choices model with the monotonicity condition.

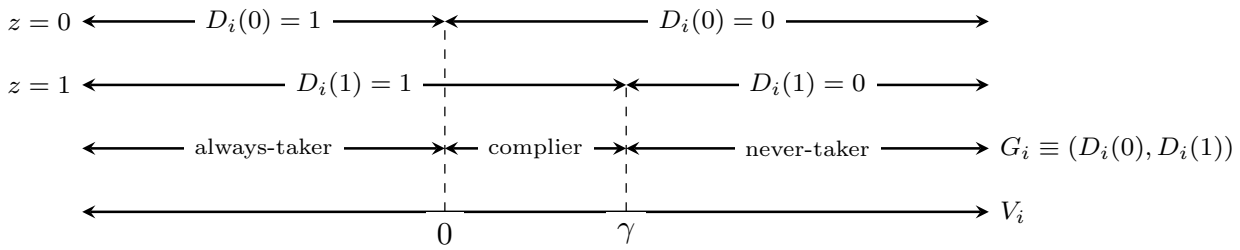
The argument just outlined is a special case of a more general equivalence theorem due to Vytlacil (2002). The key to the argument is the additive separability of Z_i and V_i in the threshold-crossing model. The equivalence continues to hold if (5) is replaced by $D_i = \mathbb{1}[V_i \leq \nu(Z_i)]$ for some unknown function ν. It breaks down in a more general model in which Z_i and V_i interact, such as $D_i = \mathbb{1}[\nu(Z_i, V_i) \geq 0]$, which no longer necessarily produces or is produced by potential treatments that satisfy monotonicity.

The implication of Vytlacil's equivalence theorem is that rather than presenting a new model, Imbens and Angrist (1994) were in fact continuing with the same selection model developed in econometrics in the 1970s–1980s, but using a different notation. Their contribution was not so much in the model itself, but in establishing an important nonparametric identification result, the local average treatment effect, which we discuss ahead in Section 3.2. Perhaps inadvertently, they also contributed to clarifying that the additive separability between Z_i and V_i in the threshold-crossing model has a behavioral interpretation as the monotonicity condition.

Vytlacil's equivalence theorem is quite specific to the case of a binary treatment, although within that context it extends naturally to multivalued instruments and covariates. A wider variety of selection models are used for non-binary treatments, many of which we discuss throughout this chapter. These models are more complicated, but they are motivated by a recognition that capturing the relevant treatment variation is important for an IV argument. While it can be tempting to turn a multi-valued treatment into a binary one by collapsing its values together, this can create violations of the exclusion condition.

When modeling with potential treatments, it is often clarifying to think of individuals i as being partitioned into latent groups depending on their potential treatments. Suppose that there are $K + 1$ instrument values $z_0, z_1, \ldots, z_K$ with potential treatments $D_i(z_0), D_i(z_1), \ldots, D_i(z_K)$. If the treatment can take say four values, then an individual can be in one of 4^{K+1} possible choice groups, which we write

Figure 1: The Vytlacil (2002) equivalence theorem



Notes: *The figure illustrates the case in which monotonicity is in the direction* $D_i(1) \geq D_i(0)$ *for all* i*. Individuals with potential no-instrument choice* $D_i(0) = 1$ *get mapped to* $V_i \leq 0$*. Because there are no defiers, these individuals are always-takers. Individuals with potential with-instrument choice* $D_i(1) = 0$ *get mapped to* $V_i > \gamma$*. Again because there are no defiers, these individuals are never-takers. Compliers get mapped to the remaining region of* $0 < V_i \leq \gamma$*.*

as $G_i = (D_i(z_0), D_i(z_1), \ldots, D_i(z_K))$.[5] Assumptions like the monotonicity condition can be viewed as requiring some choice groups to not exist (have zero probability). So, for example, with $K = 1$, $z_0 = 0, z_1 = 1$, and a binary treatment, there are $2^2 = 4$ choice groups—always-takers $G_i = (1, 1)$, never-takers $G_i = (0, 0)$, compliers $G_i = (0, 1)$, and defiers $G_i = (1, 0)$—and the monotonicity condition is the assumption that $\mathbb{P}[G_i = (1, 0)] = 0$, so there are no defiers. We use this group notation extensively ahead.

2.5 Full exogeneity

Both Imbens and Angrist (1994) and prior work using the threshold-crossing model (5) assumed the instrument to be exogenous with respect to both the outcome and the treatment. In the potential choices notation with a binary treatment and binary instrument the assumption is that Z_i is independent of $(Y_i(0), Y_i(1), D_i(0), D_i(1))$, while the equivalent assumption in latent variable notation is that Z_i is independent of (ϵ_i, V_i). In contrast, in the classical linear IV model, Z_i is only required to be exogeneous with respect to latent factors determining the outcome: $(Y_i(0), Y_i(1))$ or ϵ_i.[6] We call these contrasting assumptions outcome and full exogeneity for emphasis. Most approaches to incorporating UHTE impose something like full exogeneity.

[5] Choice groups are an example of what Frangakis and Rubin (2002) call a principal stratification (see also Robins and Greenland (1992)), and what Heckman and Pinto (2018) describe as response vectors or types. Manski (2007) used the same idea for discrete choice analysis.

[6] Exogeneity in the classical linear IV model is usually stated in terms of orthogonality, correlation or mean independence. Full independence is often necessary for analyzing UHTE. The substantive economic interpretation of the exogeneity of an instrument rarely depends on whether the mathematical formulation is independence or something weaker, like orthogonality.

Full exogeneity exposes an important distinction between selection models and statistical first stages of the sort that show up in discussions of the two stage least squares (2SLS) estimator. A statistical first stage satisfies

$$D_i = \pi Z_i + \eta_i \quad \text{where} \quad \mathbb{E}[Z_i \eta_i] = 0, \tag{7}$$

an equation often written in tandem with (1) for the classical IV model. Equation (7) can always be satisfied without any assumptions (beyond existence of moments) by taking π to be the population regression coefficient from regressing D_i onto Z_i, so that η_i are the population residuals. This is a statistical relationship, not a model of causality; the population residuals are simply the difference between D_i and its best linear predictor using Z_i. In contrast, the way in which selection models are typically used for IV models with UHTE presupposes a causal interpretation. This requires the unobservables to be viewed as "everything else," whether stated using a latent variable V_i, or potential choices $D_i(z)$. Assuming full exogeneity implies that the first stage coefficient π represents a causal effect of Z_i on D_i.[7]

The difference between outcome and full exogeneity has important and underappreciated practical implications for IV analysis. Under full exogeneity, different instruments for the same treatment variable cannot be considered in isolation. For example, Card (1995) used distance as an instrument for college, while Kane and Rouse (1995) used public college tuition. Suppose that the outcome is future earnings. Evaluating the outcome exogeneity of these instruments means considering whether they are correlated with any other omitted correlates of future earnings. This task is familiar from classical IV analysis with constant treatment effects. Evaluating the *full* exogeneity of these instruments means also considering whether they are correlated with any other omitted correlates of college attendance. In particular, if distance and tuition are correlated, then neither distance nor tuition will satisfy full exogeneity when used in isolation. Full exogeneity requires using both instruments together, or controlling for the omitted instrument as a covariate. For further discussion, see Heckman (2010, Section 3.6) and Mogstad et al. (2021, Section III.E).

Covariates are often used in IV analysis to try to weaken instrument exogeneity. An instrument that is not exogenous unconditionally might still be exogenous conditional on a vector of covariates X_i. This type of reasoning is routine in empirical work, see for example Gelbach (2002, pg. 309), Dinkelman (2011, pg. 3091) or Maestas et al.

[7] For example, if Z_i is binary, and if a constant term is included in (7), then $\pi = \mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$ as a matter of regression algebra. If full exogeneity is additionally maintained, then $\pi = \mathbb{E}[D_i(1) - D_i(0)]$ is the average causal effect of Z_i on D_i.

(2013, pp. 1811–1812), to name just a few. In the classical IV model, covariates are introduced by including them in the outcome equation (1) and then assuming that they are orthogonal to any remaining latent variation in Y_i:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i + \epsilon_i \quad \text{where} \quad \mathbb{E}[X_i \epsilon_i] = 0. \tag{8}$$

This outcome equation is the same as (2) with the interaction term removed, a common specification when covariates are used to support exogeneity rather than for estimating heterogeneity along observables.

Covariates are introduced more directly in modern IV analysis by making exogeneity conditional on covariates. With a binary treatment and binary instrument, conditional full exogeneity is the assumption that $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ is independent of Z_i, conditional on X_i. Conditional outcome exogeneity is that $(Y_i(0), Y_i(1))$ is independent of Z_i, conditional on X_i. Conditional independence is a nonparametric concept. This makes it easier to reason about than the orthogonality condition in the classical IV model (8), which also requires considering functional form: is a linear function of X_i enough to ensure orthogonality with ϵ_i, or are quadratic or more exotic terms needed as well? Separating the economic question of exogeneity from the statistical question of functional form is useful conceptually, even if (perhaps especially when) parametric functional forms end up being used for estimation in practice.

2.6 Target parameters

The classical IV model has a single homogeneous treatment effect, the coefficient α_1 on D_i. Allowing for treatment effect heterogeneity replaces this single effect with a distribution of effects across individuals. How do we want to summarize this distribution? The answer to this question inherently depends on the researcher's motivation for causal inference.

We see two broad and not necessarily exclusive motivations: policy and "science." Policy means inference with the intent to evaluate a change in the way the treatment variable is assigned. For example, Ito et al. (2023) evaluate the welfare impacts of different incentive policies designed to encourage users to adopt electricity plans with dynamic pricing. The "science" motivation is a bit of a residual category, but could perhaps be thought of as knowledge for the sake of knowledge, without necessarily being used to guide a concrete decision. Understanding the effect of education on labor market outcomes is important for understanding fundamental issues about human capital, something which has value independent of any policy implications.

Both motivations produce empirical questions. *What would the gain in welfare be*

if an $x incentive for dynamic pricing were provided? What is the average effect of a college degree on future earnings? Answering either type of empirical question requires estimating quantities that summarize the distribution of treatment effects. We call these quantities target parameters.

The choice of target parameter tends to be clearer for policy questions. One way to evaluate a policy change is to estimate the average outcome that would occur under the new policy and compare that to the status quo average observed in the data. Heckman and Vytlacil (2001a, 2005) call this the policy-relevant treatment effect (PRTE). If the conjectured policy change is not observed in the data, then estimating a PRTE requires extrapolation. Policy changes that involve mandating or forbidding a treatment do not require modeling selection because treatment choice is fully determined in the counterfactual. Policy changes that involve changing incentives to take treatment do require a model of how those changes affect selection into treatment.

Choosing a target parameter for the vaguer "scientific" motivation is more open-ended, but can be guided by two reasonable principles. Fix a population of interest. For example, the entire population in a representative survey, the population of individuals at risk of disability or unemployment, or the subset of females in the population covered by a given study. Target parameters that reflect larger subpopulations of the population of interest are more interesting than those that reflect smaller and more specific subpopulations. Target parameters that can be clearly interpreted in terms of basic statistical quantities, such as means and quantiles, should also be preferred. These two principles are consistent with the way randomized controlled trials are evaluated: with simple quantities such as means and quantiles, and with an understanding that the results are specific to the population being studied.

If different target parameters answer different questions, then it stands to reason that some target parameters will be harder to estimate than others. "Harder" here means, loosely, that less can be learned under the same assumptions, and that stronger assumptions are needed to learn the same amount; it could also involve various measures of statistical difficulty. Acknowledging the existence of this trade-off does not mean that there shouldn't be guiding principles to the choice of a target parameter. A reasonable approach to resolving this trade-off is to estimate a variety of target parameters that answer questions of different ambition under assumptions of different strengths.

An example of a target parameter that is often difficult to estimate is the average treatment effect (ATE), which is the overall average effect of a (binary) treatment on the population under study. The ATE may or may not answer an interesting policy question. In the Ito et al. (2023) study of dynamic electricity pricing, the

ATE compares a policy in which all consumers have static pricing to one in which all consumers are mandated to have dynamic pricing. In the context of active labor market programs, the ATE compares a policy that mandates training to one that prohibits it, a mental exercise that probably has little policy relevance (e.g. Heckman et al., 1999). However, on the two guiding principles for a scientific motivation, the ATE scores at the top. Averages are perhaps the easiest summary of a distribution to understand, and the population reflected in the ATE is the overall population under study, same as in a randomized controlled trial. The ATE is, however, usually difficult to estimate with an IV while allowing for UHTE; it is generally not identified without additional assumptions beyond full exogeneity and the sharp bounds on the ATE are often too wide to be of practical interest.

A target parameter that is easier to estimate is the average treatment effect for the compliers to a binary instrument, the so-called local average treatment effect (LATE), the mechanics of which we discuss extensively ahead. Whether the LATE answers an interesting policy question depends on what the instrument is. Ito et al. (2023) randomly assigned an incentive of $60 for adopting dynamic pricing relative to a baseline of no incentive. The LATE derived from this contrast provides a comparison of exactly this policy, which might be one potential policy of interest. The average effect for compliers is as easy to interpret as the ATE, but it only concerns the compliers, which are a smaller subset of the overall population. All things equal, LATEs that represent larger shares of compliers should be more interesting on a scientific basis.

Much ink has been spilt on the question of whether the LATE is an interesting target parameter compared to say, the ATE, or something else, such as the average treatment effect on the treated.[8] In our view, extreme positions on this question are indefensible. Whether a given target parameter is interesting or relevant depends on the context and the empirical question, which is itself necessarily driven by the researcher's motivation for pursuing causal inference. How interesting a target parameter is also cannot be divorced from the difficulty involved in estimating it; there are trade-offs involved and reasonable people can disagree on how these trade-offs are resolved. Instead, the important and hopefully less controversial point is that the target parameter should be clearly stated and correctly interpreted. Not doing so obscures the empirical question that the analysis is intended to answer.

[8] An exhausting but not exhaustive list is Angrist et al. (1996), Robins and Greenland (1996), Heckman (1997), Imbens (2010), Deaton (2010), Heckman and Urzua (2010), Pearl (2011), and Swanson and Hernán (2014).

2.7 Testability

The traditional route for testing the classical linear IV model is an overidentification test with multiple instruments (Sargan, 1958). The logic of an overidentification test can be viewed as comparing the equality of multiple possible IV estimates of the same constant treatment effect; see Windmeijer (2019) for a precise statement. Such a test might reject because of UHTE rather than because the instrument fails to be excluded or exogenous.

There's a well-developed literature that provides alternative tests for IV models that allow for UHTE. These tests do not require multiple instruments and instead are based on whether statistical quantities that should reflect well-defined treatment effects or potential outcome distributions actually have the properties of such objects. For example, if the outcome Y_i is known to lie in $[-1, 1]$, then does an estimator that should reflect an average causal effect for a subpopulation actually lie in $[-1, 1]$? If not, then model can be rejected. We do not discuss testability in this chapter out of length considerations, but see Balke and Pearl (1997), Imbens and Rubin (1997), and Heckman and Vytlacil (2005) for discussions of the testable implications, Bhattacharya et al. (2012), Huber and Mellace (2014), Kitagawa (2015), Mourifié and Wan (2016), and Kédagni and Mourifié (2020) for various ways of turning these implications into formal statistical tests, and Carr and Kitagawa (2023), Frandsen et al. (2023), and Sun (2023) for more recent developments and applications.

3 Reverse Engineering: Interpreting Linear Estimators

If there is UHTE then the classical linear IV model is misspecified, but a linear IV estimator can still be computed. Perhaps it has an interpretation that is robust to omitted UHTE? This line of reasoning has been popular in the recent microeconometrics literature. We describe it as reverse engineering because it starts with a tool—the estimator—and attempts to reverse engineer an interpretation for it. This section contains a comprehensive survey and synthesis of reverse engineering results for linear IV estimators.

In the next subsection, we begin by first introducing some concepts used in reverse engineering exercises. Then we use these concepts to review the well-known LATE interpretation that applies in the baseline case of a binary treatment, binary instrument, and no covariates. The remainder of the section then considers in turn what happens as one deviates from the baseline case by having either a non-binary instrument, a non-binary treatment, or by including covariates.

3.1 Estimators, estimands, and weak causality

Reverse engineering arguments start with an estimator and consider its associated estimand, meaning the population quantity to which the estimator can be expected to converge under a law of large numbers. For example, the estimand for the ordinary least squares estimator of the coefficient on D_i in a regression of Y_i on D_i and a constant is $\mathbb{C}[Y_i, D_i]/\mathbb{V}[D_i]$. The estimand is then decomposed into terms involving the underlying causal model, typically using potential outcome notation. The focus is on how assumptions about the underlying causal model affect the properties of the decomposition. In this way the estimator is taken as the starting point and its interpretation is then reverse engineered from an underlying causal model.

A minimal criterion for a successful interpretation is typically taken to be whether the estimand can be written in the form of a weighted average of mutually exclusive subgroup-specific average treatment effects with weights that are all non-negative. In an IV framework, an individual's group is defined by their unobserved potential choice behavior—the variable G_i introduced in Section 2.4—together with their observed covariates, X_i. Being able to write an estimand β as a non-negatively weighted average of group-specific treatment effects means that there exist weights $\omega(g, x) \geq 0$ such that

$$\beta = \sum_{g,x} \underbrace{\omega(g,x)}_{\text{weights}} \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) | G_i = g, X_i = x]}_{\text{subgroup-specific treatment effects}}. \tag{9}$$

The non-negativity of the weights is seen as important because it guarantees that the estimand cannot systematically have the "wrong" sign. Suppose that all underlying covariate- and group-specific treatment effects are non-negative. If β can be written as (9) with weights that are non-negative, then β must also be non-negative. Blandhol et al. (2022) generalize this reasoning to include estimands that do not have decompositions like (9) either because the treatment takes multiple values or because the weights applied to the treatment arms are asymmetric. They call an estimand "weakly causal" if β is non-negative whenever the causal effect of all covariate- and group-specific treatment contrasts are non-negative. Appendix B provides the formal definition and a generalization to unordered treatments.

Weak causality is, as the name suggests, an extremely minimal requirement for an estimand to be viewed as "causal." Any target parameter appropriate for the scientific motivation of estimating an average treatment effect for a subpopulation will be weakly causal. But a weakly causal estimand with non-constant weights need not reflect the average treatment effect for any single subpopulation.[9] If all that is known about an

[9] Poirier and Słoczyński (2024) show that a weakly causal estimand can, however, reflect the average

estimand is that it is weakly causal, then the scientific question it answers is an easy one based on a strong premise. *Assuming that everyone has either a positive or negative treatment effect, is the common sign positive or negative?*

Weak causality is not necessarily a useful property for policy purposes. A target parameter that answers a policy question might not be weakly causal if the policy shifts some individuals into treatment and others out of treatment. If such a target parameter had form (9) then it would give positive weight to groups induced by the policy change to increase treatment, but negative weights to those induced to decrease treatment. The opposite possibility can also arise: an estimand that is not weakly causal but *is* useful for policy. Kline and Walters (2016, Section V.C) provide an example of this in the context of unordered treatments, where the estimand conflates two different treatment contrasts but still reflects an important target parameter for a policy counterfactual.

Most weakly causal estimands also have weights that are convex in the sense of being both non-negative and summing to one across all subgroups: $\sum_{g,x} \omega(g,x) = 1$. The additional sum-to-one property ensures that if treatment effects are actually homogeneous, then β is equal to that common single effect. More generally, it ensures that β lies somewhere between the smallest and largest subgroup treatment effects, which seems like an intuitively attractive property.

The same estimand can have multiple different weakly causal interpretations. As a simple example, suppose that D_i is binary and randomly assigned, and let $\beta = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$ be the population difference in means. Then

$$\underbrace{\beta}_{\text{difference in means (estimand)}} = \overbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}^{\text{overall ATE (interpretation \#1)}} = \underbrace{\sum_x \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]\,\mathbb{P}[X_i = x]}_{\text{weighted average of conditional ATEs (interpretation \#2)}}. \tag{10}$$

The first equality is the usual interpretation of β as the overall ATE, which is a convex weighted average of one term. The second equality shows that β can alternatively be interpreted as a convex weighted average of the conditional ATEs across all covariate groups.

3.2 Binary treatment, binary instrument, no covariates

The leading example in which the linear IV estimator has a clear interpretation that is robust to misspecification occurs in the simplest possible setting. Suppose that

treatment effect for a new subpopulation formed by combining subsets of multiple covariate and/or choice groups.

$D_i \in \{0,1\}$ and $Z_i \in \{0,1\}$ are both binary, and there are no additional covariates X_i. Assume that the monotonicity condition (or threshold-crossing model) in Section 2.4 holds together with the full exogeneity condition discussed in Section 2.5. Imbens and Angrist (1994) showed that under these assumptions the average treatment effect for the compliers—what they called the local average treatment effect, or LATE—is identified:

$$\underbrace{\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}}_{\beta_{\text{WALD}} \equiv \text{ Wald estimand}} = \underbrace{\mathbb{E}[Y_i(1)-Y_i(0)|\overbrace{D_i(0)=0, D_i(1)=1}^{\text{subpopulation of compliers } (G_i=(0,1))}]}_{\text{average treatment effect for compliers}} \equiv \text{LATE}. \quad (11)$$

Angrist and Imbens (1995) later called the left-hand side the Wald (1940) estimand. The same result appeared in the biostatistics literature in lesser-known papers by Permutt and Hebel (1989) and Baker and Lindeman (1994).[10]

Equation (11) is a natural and intuitive nonparametric identification result. In the absence of additional assumptions, the only subpopulation whose treatment effects could possibly be identified are the individuals whose decisions are causally affected by the instrument. Treatment effects for always-takers cannot be identified without some sort of extrapolation because they are never observed in the untreated state; the instrument has no causal effect on their treatment choice behavior. The same is true for never-takers, who are never observed in the treated state.

In general, both compliers and defiers are affected by the instrument, and the numerator of the Wald estimand reflects the aggregate change in outcomes that results from shifting compliers into treatment and defiers out of treatment. The monotonicity condition eliminates the defiers, leaving only the impact on compliers. The numerator of the Wald estimand—the reduced form—then reflects the aggregate change in outcomes caused by the instrument, which reflects both the size of the complier group and the impact that treatment has on their outcomes. The denominator—the first

[10]The analysis of Permutt and Hebel (1989) is informal, but remarkably clear, elegant, and precise. See in particular the bottom of page 621, where the authors recognize the four choice groups created by a binary treatment and binary instrument, followed by the monotonicity condition, the treatment effect for the compliers, the implied identification of the shares of each choice group, and the attenuation result for multivalued treatments formalized by Angrist and Imbens (1995, Section 3.1). The analysis of Baker and Lindeman (1994) is more formal, but a bit obscured by its embedding inside a "paired availability design;" however see the beginning of Section 3, Section 5, and Appendix I.

stage—adjusts for the size of the complier group:

$$\underbrace{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}_{\text{denominator of the Wald estimand}} \overbrace{=\mathbb{E}[D_i(1)-D_i(0)]}^{\text{by full exogeneity}} \overbrace{=\mathbb{P}[\underbrace{D_i(0)=0, D_i(1)=1}_{\text{the complier choice group } G_i=(0,1)}]}^{\text{by monotonicity}}. \tag{12}$$

The ratio of the reduced form to the first stage—the Wald estimand—then reflects the average per-unit treatment effect among the compliers, which is the LATE defined in (11).

The misspecification-robust interpretation of the linear IV estimand as a LATE comes from the relationship

$$\underbrace{\text{LATE}=\beta_{\text{WALD}}}_{\text{from (11)}} \equiv \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} = \underbrace{\frac{\mathbb{C}[Y_i,Z_i]}{\mathbb{C}[D_i,Z_i]}}_{\text{simple IV estimand}} \equiv \beta_{\text{IV}}, \tag{13}$$

where the simple IV estimand β_{IV} refers to population coefficient on D_i for the linear IV estimator that instruments $[1, D_i]'$ with $[1, Z_i]'$. This equality between the Wald and IV estimands is specific to the case in which Z_i is binary and there are no covariates. It does not hold more generally. It is the source of the misspecification-robust interpretation that "linear IV is LATE."

3.3 Multivalued instruments

Suppose that we generalize the setting slightly to allow Z_i to take on multiple values $z_0, z_1, \ldots, z_K$, but that D_i is still binary.[11] Each instrument value now has an associated potential treatment $D_i(z_0), D_i(z_1), \ldots, D_i(z_K)$.

Imbens and Angrist (1994) generalized the monotonicity condition by assuming that the values of the instrument can be placed in order of their impact on treatment choice, with an ordering that does not vary across individuals i. As in the binary case, this ordering does not need to be known a priori. Suppose that the instrument is indexed in increasing order, so that the monotonicity condition becomes $\mathbb{P}[D_i(z_0) \leq D_i(z_1) \leq \cdots \leq D_i(z_K)] = 1$, meaning that larger instrument values make everyone more likely to take treatment. Vytlacil (2002) showed that the Imbens and Angrist (1994) assumptions are the same as modeling selection with the threshold-crossing

[11] We focus on a discrete number of instrument values for simplicity. The continuous instrument case is conceptually similar and intuitively requires replacing sums with integrals. See Alvarez and Toneto (2024) for details.

Table 1: Group definitions when D_i is binary and Z_i takes four values ($K = 3$)

$D_i(z_0)$	$D_i(z_1)$	$D_i(z_2)$	$D_i(z_3)$	G_i	**Group description**
1	1	1	1	AT	Always-takers
0	1	1	1	CP_1	z_1-compliers ($G_i = \text{CP}_1$)
0	0	1	1	CP_2	z_2-compliers ($G_i = \text{CP}_2$)
0	0	0	1	CP_3	z_3-compliers ($G_i = \text{CP}_3$)
0	0	0	0	NT	Never-takers
0	1	0	1	DF	Defier (one of $2^4 - (3+2) = 11$ types)

Notes: *When $K = 3$ the monotonicity condition allows for the six possible configurations of $G_i \equiv (D_i(z_0), D_i(z_1), D_i(z_2))$ shown here.*

model,

$$D_i = \mathbb{1}[V_i \leq \nu(Z_i)], \tag{14}$$

where ν is an unknown function. Full exogeneity is still required, and now means that Z_i is independent of $(Y_i(0), Y_i(1), D_i(z_0), D_i(z_1), \ldots, D_i(z_K))$ or, equivalently, that Z_i is independent of $(Y_i(0), Y_i(1), V_i)$.

With $K+1$ instrument values, there are 2^{K+1} treatment choice groups in general. The monotonicity condition implies that only $K+2$ of these can exist: the always-takers, the never-takers, and K different complier groups, one for each subsequent pair of values for the instrument. Table 1 illustrates. The same argument that produces (11) shows that the average treatment effect for each complier group is identified from the Wald estimand using these subsequent pairs:

$$\underbrace{\frac{\mathbb{E}[Y_i|Z_i = z_k] - \mathbb{E}[Y_i|Z_i = z_{k-1}]}{\mathbb{E}[D_i|Z_i = z_k] - \mathbb{E}[D_i|Z_i = z_{k-1}]}}_{\text{one of several possible Wald estimands}} = \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|\overbrace{D_i(z_{k-1}) = 0, D_i(z_k) = 1}^{k\text{-compliers, } (G_i = (0,0,\ldots,1,1\ldots,1) \equiv \text{CP}_k)}]}_{\text{the average treatment effect for } k\text{-compliers } (\text{LATE}_k)} \tag{15}$$

for each $k = 1, \ldots, K$. We use the short-hand LATE_k for the right-hand side of (15), a notation which clarifies that there are now multiple LATEs. Even in this simple extension from the previous section, the statement that "linear IV is LATE" already doesn't make sense: *which* LATE?

On top of that, *which* linear IV? With a single binary instrument, a binary treatment, and no covariates, the IV estimand given in (13) was the only one to consider. When the instrument takes multiple values, there are now many possible IV estimators, each producing a different IV estimand. A general formulation is to instrument

for $[1, D_i]'$ with $[1, \zeta(Z_i)]'$, where ζ is a scalar function of Z_i. This nests using Z_i directly as an instrument for D_i, in which case the IV estimand is the same as in (13). It also nests any 2SLS estimand, in which case $\zeta(Z_i)$ are the population fitted values from the first stage.

Imbens and Angrist (1994) showed that this IV estimand can be decomposed as

$$\underbrace{\frac{\mathbb{C}[Y_i, \zeta(Z_i)]}{\mathbb{C}[D_i, \zeta(Z_i)]}}_{\text{linear IV estimand}} = \sum_{k=1}^{K} \underbrace{\frac{\mathbb{P}[G_i = \text{CP}_k]\,\mathbb{C}[\zeta(Z_i), \mathbb{1}[Z_i \in \{z_\ell\}_{\ell \geq k}]]}{\mathbb{C}[D_i, \zeta(Z_i)]}}_{\text{weights}} \text{LATE}_k. \tag{16}$$

The weights sum to one and can be shown to be non-negative if $\zeta(z)$ is non-decreasing in z. This is satisfied if $\zeta(z) = z$, so that the estimand is again the simple IV estimand on the right-hand side of (13). It is also satisfied for the 2SLS specification whose first stage includes an indicator for each value of the instrument, in which case $\zeta(z) = \mathbb{P}[D_i = 1 | Z_i = z]$ becomes the propensity score. In either of these cases the IV/2SLS estimand on the left-hand side of (16) is weakly causal.

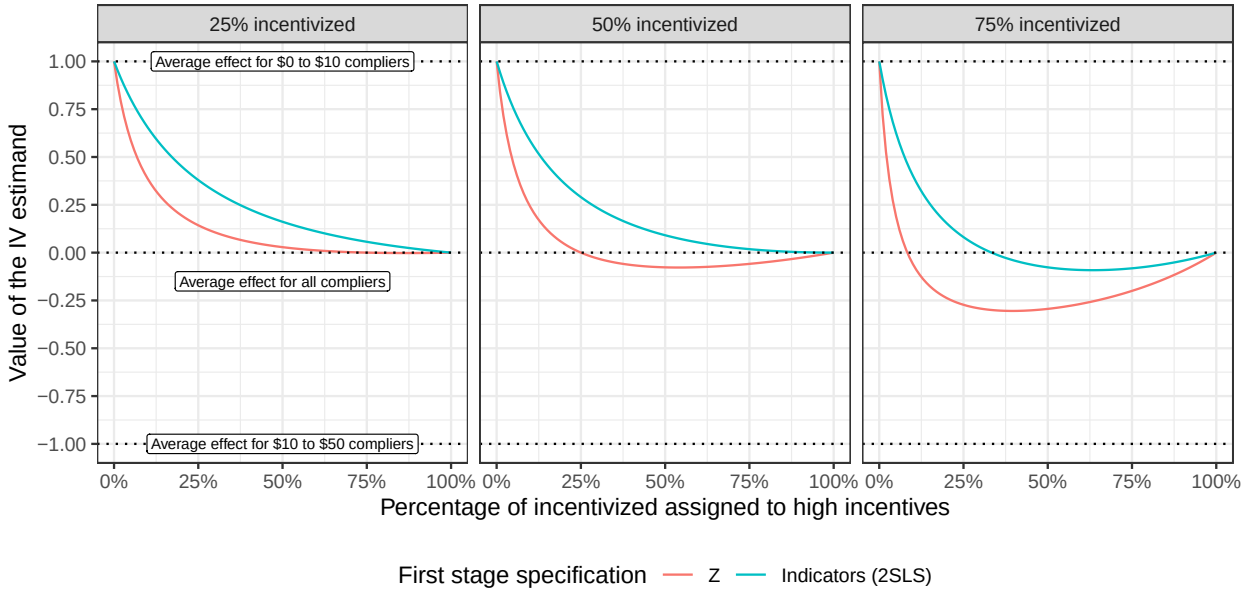
The weights in (16) are larger for larger complier groups, a feature that seems intuitive. However, the linear IV weights also vary with a second term that reflects how $\zeta(Z_i)$ and Z_i covary. One implication is that different choices of estimator—different choices of ζ—estimate different objects. A second implication is that the weights—and so also the estimand—depend on the marginal distribution of the instrument.

As an example, suppose that a researcher conducted a randomized experiment in which individuals were encouraged to take a treatment. Some individuals were given no additional incentive, while others were randomly assigned an incentive of \$10 or \$50 for taking treatment. Only 5% of subjects in the unincentivized arm took treatment, while 20% and 35% took treatment in the \$10 and \$50 arms.[12] There are two complier groups in this setting: those who wouldn't take treatment if unincentivized, but would if offered \$10, and those who would only take treatment if offered \$50. Suppose that the average treatment effect for the former group is 1, and for the latter group is -1, a difference which might reflect the second group's greater reluctance to participate. What will the IV estimand be?

Figure 2 shows that the answer depends delicately both on how the incentives were randomly assigned and on which IV estimand is being considered. The two lines indicate the values of two different IV estimands, one in which Z_i is used directly as an instrument as in (13), and one in which indicators for the different incentive arms are

[12] For example, Dutz et al. (2022, 2023a,b) implemented incentivized surveys with this design, Ito et al. (2023) randomly assigned incentives for switching to dynamic electricity pricing, and Lee et al. (2019) randomly assigned incentives to connect to the electrical grid.

Figure 2: The marginal distribution of the instrument affects the IV estimand



Notes: *Values of the linear IV estimand in a hypothetical experiment in which individuals were incentivized to take a binary treatment. There are three incentive arms: no incentive, low ($10) and high ($50). The panels show the proportion of individuals assigned to the low or high arms. The x-axis shows the proportion of individuals in the high arm relative to the low arm. Two different estimands are shown: one that uses the level of the incentive ($Z_i = 0, 10, 50$) and the 2SLS estimand that uses indicators for each incentive level.*

used in the first stage and combined through 2SLS. The left panel depicts a scenario in which the researcher assigns 75% of individuals to the no incentive arm and assigns some proportion of the remaining 25% to either the $10 or $50 incentive. When all of the incentivized individuals are in the low incentive arm, both estimands are the same and equal to the average treatment effect of 1 for the low incentive compliers. As the proportion assigned to the high incentive arm increases, the estimands begin to differ, and the value of the IV estimand starts to decrease as more high-incentive compliers are reflected in the estimand. The center and right panels of Figure 2 show the same comparisons when a larger share of individuals are assigned to receive any incentive. The difference between the estimands grows and in many cases they even have the opposite sign.

This scenario is one in which the premise of weak causality is not met because the low incentive compliers have positive treatment effects, while the high incentive compliers have negative treatment effects. There are three LATEs: one for each complier group separately, and a third when both complier groups are combined into a single

group. The latter can generally be written as

$$\text{LATE}_{0\to K} \equiv \mathbb{E}[Y_i(1) - Y_i(0) | \underbrace{D_i(z_0) = 0, D_i(z_K) = 1}_{\text{complies with any instrument}}] = \sum_{k=1}^{K} \mathbb{P}[G_i = \text{CP}_k]\text{LATE}_k, \quad (17)$$

which is like (16), but with weights that only depend on the complier proportions. As Figure 2 shows, neither of the two IV estimands is generally equal to any of these three LATEs when both incentives are assigned. If only low incentives are assigned, then it's as if we are back in the binary instrument case, and both IV estimands are equal to the low incentive LATE of one. If only high incentives are assigned, then we are again in a binary instrument case, and both IV estimands are equal to the combined LATE, which is zero in this example.[13] Outside of these two polar cases—that is, when the instrument actually takes multiple values—the IV estimand is not equal to any individual LATE.

3.4 Violations of monotonicity

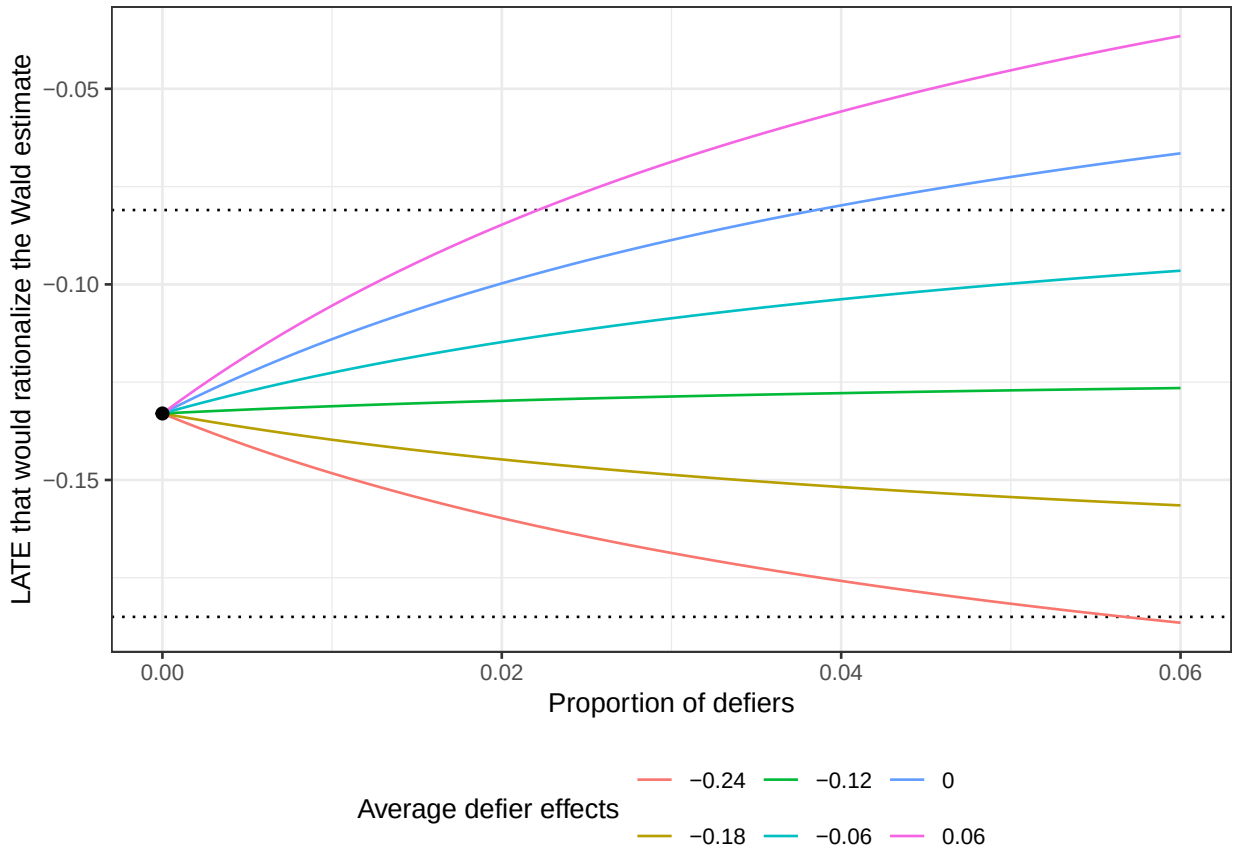
The monotonicity condition plays a central role in the LATE identification result (11). There are many settings in which it is usually uncontroversial, such as the type of incentivized experiment just discussed, where the instrument is a monetary incentive for treatment. In other contexts there can often be more scope for contention.

As one example, consider the Angrist and Evans (1998) study of the effect of fertility on labor supply. The authors used the sex composition of a family's existing children as an instrument for further childbearing. They found that among families that have at least two children, those in which the first two children had the same biological sex were more likely to go on to have a third child than those that had both a male and female child. They interpreted this as reflecting a preference for sex mix among children. For the monotonicity condition to hold requires *all* families to have this same preference for sex mix. Monotonicity would be violated if there are families whose fertility stopping rule is to have two male children.

Angrist et al. (1996, Section 5.2) show how to conduct a sensitivity analysis in the binary instrument case. They show that when the monotonicity condition does not

[13] The low and high compliers are $.20 - .05 = .15$ and $.35 - .20 = .15$ of the population, so the combined LATE is $\text{LATE}_{0\to 2} = 1 \times .15 + (-1) \times .15 = 0$.

Figure 3: Sensitivity of Angrist and Evans (1998) to violations of monotonicity



Notes: *The Wald point estimate is taken to be* $-.133$ *in the graph, matching the estimate in Table 5 of Angrist and Evans (1998) for the "worked for pay" binary outcome. Angrist and Evans (1998) report a standard error of* $-.026$*, and the associated 95% confidence intervals are indicated with dotted lines. We keep the denominator of the Wald estimate (the first stage) fixed at* $.060$*, consistent with the point estimate in Table 5 of Angrist and Evans (1998).*

hold,

$$\overbrace{\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]}}^{\text{difference between Wald and LATE}} - \text{LATE}$$
$$= \underbrace{\left(\frac{\mathbb{P}[G_i=\text{DF}]}{\mathbb{P}[G_i=\text{CP}]-\mathbb{P}[G_i=\text{DF}]}\right)}_{\text{relative size of defier group}} \underbrace{(\text{LATE}-\mathbb{E}[Y_i(1)-Y_i(0)|G_i=\text{DF}])}_{\text{difference in treatment effects}}. \quad (18)$$

The bias of the Wald estimand for the LATE has a product structure that is increasing in the size of the defier group and scaled by the difference between complier and defier

average treatment effects. Both terms in the product need to be large for the Wald estimand to differ substantially from the LATE.

Figure 3 illustrates this point using estimates from Angrist and Evans (1998). The authors report a Wald estimate of approximately $-.13$, the denominator of which is estimated to be .06. If the monotonicity condition holds, then this implies that 6% of the population are compliers and that the Wald estimate is the LATE. If the monotonicity condition doesn't hold because actually 2% of the population are defiers and 8% are compliers, then the Wald estimate would differ from the LATE, but not necessarily by much. For example, even if the 2% defiers have treatment effects that are $-.24$—nearly twice as negative as the Wald estimate of $-.13$—this would still imply a LATE of approximately $-.16$.[14]

Judge designs are a common example of an IV strategy in which the monotonicity condition can be suspect. These designs are based on institutionally-prescribed random assignment of a judge or other examiner to cases in which the judge chooses treatment. If certain judges are systematically more likely to assign treatment, then the judge identities serve as an instrument for treatment (e.g. Kling, 2006; Doyle Jr., 2007; Dahl et al., 2014; Bhuller et al., 2020).

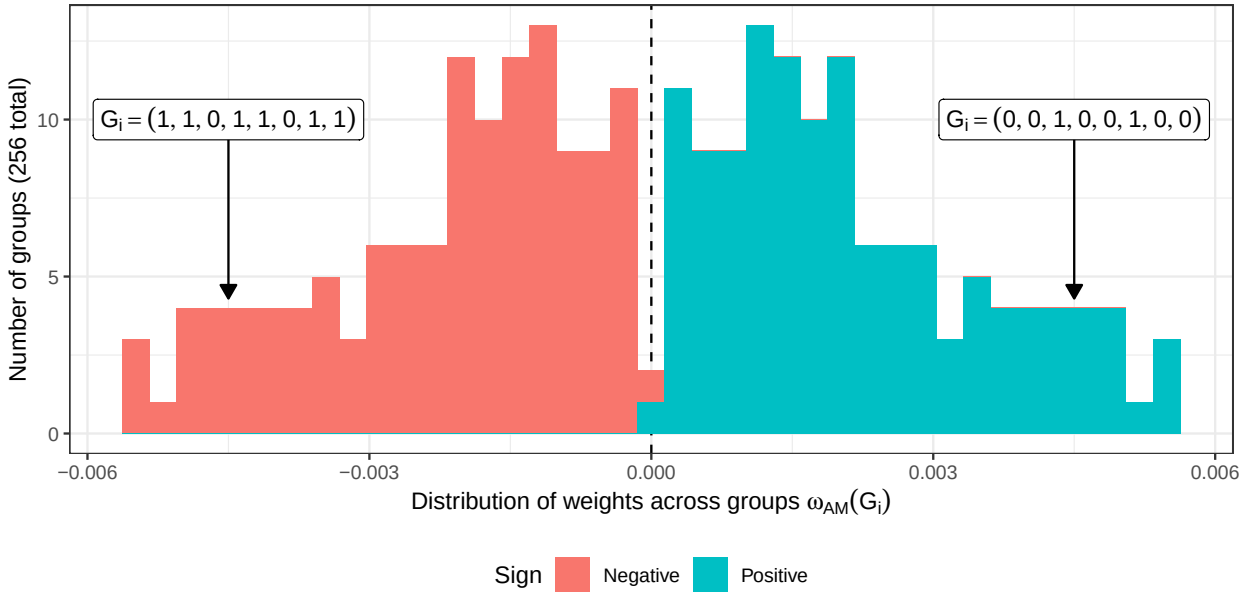
The monotonicity condition places strong restrictions on the behavior of judges. Suppose that judge A is stricter than judge B in the sense that judge A assigns treatment in a higher proportion of cases. Then the monotonicity condition requires judge A to *always* assign treatment to any case in which judge B would assign treatment. This effectively prevents judges from systematically disagreeing. These types of settings were actually discussed in the original work by Imbens and Angrist (1994, Example 2) as an example where monotonicity might be an unattractive assumption.

Frandsen et al. (2023) propose a weaker alternative to monotonicity that they describe as "average monotonicity." Their motivation is a judge design, although the assumption could be considered in other contexts as well. Suppose that the instrument Z_i denotes judge identity for $K+1$ judges labelled $z_0, z_1, \ldots, z_K$. Frandsen et al. (2023) observe that the 2SLS estimand produced by using an indicator for each judge as an excluded variable can be written as

$$\beta_{\text{2SLS}} = \frac{\mathbb{E}\left[w_{\text{AM}}(G_i)\,\mathbb{E}[Y_i(1) - Y_i(0)|G_i]\right]}{\mathbb{E}\left[w_{\text{AM}}(G_i)\right]}, \tag{19}$$

[14] See Noack (2021) for a formal analysis of this example in terms of the more general concepts of the breakdown frontier and falsification region used in the literature on sensitivity (e.g. Kline and Santos, 2013; Masten and Poirier, 2020, 2021).

Figure 4: Weights for the 2SLS estimand using the Stevenson (2018) data



Notes: *The eight judges in the Stevenson (2018) data imply* $2^8 = 256$ *treatment choice groups (values of* G_i*). The weight* $w_{\text{AM}}(G_i)$ *that each individual* i *contributes to* β_{2SLS} *is shown on the x-axis for each value that* G_i *can assume. Without any monotonicity assumption, half of the groups contribute to the 2SLS estimand with negative weight. Average monotonicity is satisfied if and only if the group shares for this half are zero.*

where w_{AM} are weights defined as

$$w_{\text{AM}}\Big(\underbrace{D_i(z_0), D_i(z_1), \ldots, D_i(z_K)}_{\equiv G_i}\Big) \equiv \sum_{k=0}^{K} \mathbb{P}[Z_i = z_k] \left(\mathbb{P}[D_i = 1 | Z_i = z_k] - \mathbb{P}[D_i = 1]\right) \times \left(D_i(z_k) - \sum_{\ell=0}^{K} \mathbb{P}[Z_i = z_\ell] D_i(z_\ell)\right). \tag{20}$$

In particular, (19) is true whether or not the monotonicity condition holds.

Frandsen et al. (2023) define average monotonicity as the assumption that the weights w_{AM} are all non-negative, meaning that $\mathbb{P}[w_{\text{AM}}(G_i) \geq 0] = 1$, or equivalently that $\mathbb{P}[G_i = g] = 0$ for any choice group g such that $w_{\text{AM}}(g) < 0$. As can be seen from (19), average monotonicity is equivalent to the assumption that β_{2SLS} is weakly causal. In this sense it assumes away the problems raised by failures of the usual monotonicity condition.

Figure 4 illustrates the content of the average monotonicity assumption using the data from Stevenson (2018) (provided by Cunningham, 2021), which has eight judges. Without any assumptions on potential treatment choice behavior there are

$2^{K+1} = 2^8 = 256$ treatment choice groups, one for each configuration of potential judge decisions. The weight that each individual i contributes to β_{2SLS} is given by $w_{\text{AM}}(G_i)$, which depends only on their choice group, G_i. Although the share of each choice group is not identified, the weights $w_{\text{AM}}(g)$ are identified for each value of g. Figure 4 plots estimates of $w_{\text{AM}}(g)$ as a histogram taken over all 256 treatment choice groups.

The weights are symmetric around zero. If all groups occur, then exactly half will have non-negative weights while the other half will have non-positive weights.[15] To see why, consider a pair of individuals i and i' who satisfy $D_i(z_k) = 1 - D_{i'}(z_k)$ for all k. For these individuals,

$$D_{i'}(z_k) - \sum_{\ell=0}^{K} \mathbb{P}[Z_i = z_\ell] D_{i'}(z_\ell) = -1 \times \left(D_i(z_k) - \sum_{\ell=0}^{K} \mathbb{P}[Z_i = z_\ell] D_i(z_\ell) \right), \tag{21}$$

which implies that $w_{\text{AM}}(G_{i'}) = -w_{\text{AM}}(G_i)$. An example of this symmetry is highlighted in Figure 4. One choice group is only assigned treatment for the third and sixth judge; their weights are positive. The other choice group is assigned treatment for all judges except the third and sixth; their weights are negative.[16]

Justifying average monotonicity requires explaning why all the negatively-weighted groups in Figure 4 do not exist. This seems challenging to do here, even with only eight judges. A judge design with one hundred judges has 2^{100} choice groups to consider, so that justifying average monotonicity requires arguing that half (2^{99}) of these groups do not occur, presumably an even more difficult feat. Compounding this difficulty, the identity of the groups that must not occur depends on the observable data through the propensity scores (leniencies) of all judges and the frequency with which each judge is observed. This makes any attempt at justifying average monotonicity inherently application-specific.[17] While average monotonicity is mathematically weaker than the usual monotonicity condition, it is not clear that it is any easier to justify on substantive grounds.[18]

[15] Two groups always have zero weight: the always-takers and the never-takers.

[16] The third and sixth judges are not particularly remarkable: the third one has the sixth highest propensity score and has the fifth most cases, while the sixth one has the seventh highest propensity score and has the sixth most cases.

[17] Chyn et al. (2024, Appendix A) consider the content of average monotonicity in a hypothetical context with four judges. They show that average monotonicity can be satisfied under an assumption that reduces individuals to four latent types: minority/majority and with/without a criminal background. Their justification involves stylized assumptions on the frequency of these types in the population and the way in which the four judges respond to these types.

[18] Sigstad (2024b) estimates treatment group shares using judge panels under the assumption that judges rule in panels the same way that they would individually. He finds evidence against the usual monotonicity

Reverse engineering arguments grind to a halt without some sort of monotonicity condition. de Chaisemartin (2017) shows that the IV estimand can still be interpreted as representing average treatment effects for some *subset* of the compliers, even if the monotonicity condition is not satisfied, as long as the distribution of treatment effects between the compliers and defiers is not too different. The result effectively provides conditions under which a single weighted average with some negative weights—the IV estimand without a monotonicity condition—is equal to a weighted average of some subset of its non-negatively weighted components. Whether an average treatment effect for the group represented by such a subset is interesting (or indeed, even uniquely defined) is another question, but the result provides a thought-provoking examination of the logical limits of reverse engineering. If all we want to know is whether a given estimate represents a treatment effect for *someone*, then we can do that under general conditions. Is that robustness or superficiality?

3.5 Multiple instruments

The credibility of the monotonicity condition is not about the number of values that the instrument takes. Compare monetary incentives to judges. If it's reasonable to assume that everyone prefers treatment with $10 to treatment without compensation, then it's also reasonable to assume that everyone prefers treatment with $50 to treatment without compensation. Adding another $100 incentive arm would not jeopardize this argument. On the other hand, if we're worried that any two of eight judges may disagree on some cases, then we would probably also have that concern in a setting with only two judges. The issue instead is whether the instrument has a natural ordering: ordering monetary incentives from small to large is natural, but ordering judges, say in terms of their leniency, attempts to reduce something complicated (a judge) down to one dimension.

This ordering issue arises when Z_i is a vector containing multiple distinct components, a case we describe as multiple instruments. Multiple instruments are always multivalued. For example, if Z_i is a vector of two binary instruments, then Z_i takes four values. If the monotonicity condition holds for these four values, then all of the discussion in the previous section for multivalued instruments continues to apply with multiple instruments.[19] But the fact that the four values represent a combination of two distinct components usually makes the monotonicity condition an extremely unattractive assumption.

condition and somewhat weaker evidence against the average monotonicity condition.

[19] Imbens and Angrist (1994, pg. 470) explicitly mention the case in which Z_i is a vector in deriving the decomposition (16).

To see why, suppose that incentives for treatment vary along two dimensions: a monetary incentive, and the distance to the location where treatment is administered. An experimental design like this was used by Thornton (2008) to study the demand for learning about HIV status in Malawi. For simplicity, suppose that both instruments are binary, with $Z_{i1} \in \{0,1\}$ denoting a monetary incentive (no or yes), and $Z_{i2} \in \{0,1\}$ denoting distance (far or near), so that there are four potential treatments $D_i(z_1, z_2)$. The monotonicity condition requires these potential treatments to be ordered for all individuals; in particular, it requires either $\mathbb{P}[D_i(0,1) \geq D_i(1,0)] = 1$ or that $\mathbb{P}[D_i(1,0) \geq D_i(0,1)] = 1$. The first condition says that there is no one who would take treatment if they were given a monetary incentive but not take treatment if they were assigned to a close location. The second condition says the opposite: there is no one who would take treatment if they were assigned to a close location but not if they were given a monetary incentive. Either condition assumes that there is no meaningful heterogeneity in the opportunity cost of time (responsiveness to distance).[20]

If the monotonicity condition is dropped entirely, then a linear IV estimand will include negatively-weighted average treatment effects for some groups, and so fail to be weakly causal. The problem here is the model of treatment choice: some model is needed for a weakly causal interpretation, but the monotonicity (threshold-crossing) model is too strong to be credible. Mogstad et al. (2021) consider an intermediate model of treatment choice called partial monotonicity. Partial monotonicity requires the usual monotonicity condition to be satisfied for each instrument separately, while holding all other instruments fixed. For example, it is satisfied if all individuals are more likely to take treatment when given a monetary incentive and when closer to a treatment center, but it does not require one or the other to be a uniformly more effective inducement to treatment. Mogstad et al. (2021) show that partial monotonicity can be sufficient for a weakly causal interpretation: with two binary instruments, a 2SLS estimand with a saturated first stage will be weakly causal as long as the instruments are not negatively correlated.

3.6 Ordered, cardinal treatments

The monotonicity condition is a model of treatment choice, so it must be reconsidered when the treatment has more than two values. Suppose that the treatment takes $J+1$

[20] This point was first made by Heckman and Vytlacil (2005, Section 6) and Heckman et al. (2006, Section III.D). As those authors observed, the "monotonicity" condition is not really about monotonicity, but about uniformity in treatment choice behavior. The two descriptions often coincide when the instrument is scalar, but they become meaningfully different with multiple instruments.

values labeled in increasing order as $d_0, d_1, \ldots, d_J$.[21] If the treatment is ordered, then a natural generalization of the monotonicity condition is $\mathbb{P}[D_i(1) \geq D_i(0)] = 1$. Both the notation and content are the same as in the binary case: receiving the instrument causes treatment to weakly increase for everyone.

The natural generalization of the binary threshold-crossing model (5) to an ordinal treatment would be to an ordered response model (e.g. Greene and Hensher, 2009). Depending on how the ordered response model is specified, it may or may not entail the same restrictions on treatment choice behavior as the monotonicity condition (Vytlacil, 2006). We return to this point in Section 4.5, where it becomes particularly salient. In this section, we consider reverse engineering under the Angrist and Imbens (1995) monotonicity condition.

Let $Y_i(d_0), Y_i(d_1), \ldots, Y_i(d_J)$ be potential outcomes for each treatment state. Even for a given individual i there is no longer a single treatment effect because $Y_i(d_2) - Y_i(d_1)$ could be different than $Y_i(d_1) - Y_i(d_0)$ if the size of the treatment increment differs and/or treatment effects are nonlinear. Angrist and Imbens (1995) showed that if Z_i is binary then the Wald estimand, which is still equal to the simple IV estimand, has the following decomposition:

$$\beta_{\text{IV}} \equiv \frac{\mathbb{C}[Y_i, Z_i]}{\mathbb{C}[D_i, Z_i]} = \sum_{j=1}^{J} \omega_{\text{ACR}}(j)\, \mathbb{E}[Y_i(d_j) - Y_i(d_{j-1}) | D_i(1) \geq d_j > D_i(0)],$$

$$\text{where} \qquad \omega_{\text{ACR}}(j) \equiv \frac{\mathbb{P}[D_i(1) \geq d_j > D_i(0)]}{\sum_{\ell=1}^{J} \mathbb{P}[D_i(1) \geq d_\ell > D_i(0)]\,(d_\ell - d_{\ell-1})}. \tag{22}$$

The decomposition allows for treatment effects that are both nonlinear and heterogeneous.[22] The treatment variable D_i should have a cardinal interpretation to consider β_{IV}. If it is ordered but not cardinal (e.g. low, medium, high), then β_{IV} will be sensitive to the arbitrary coding of the values d_j, making estimators that use different treatment indicators more appropriate. This case is discussed in the next section.

Angrist and Imbens (1995) described the right-hand side of (22) as the average causal response (ACR). The unit "causal response" is the effect of increasing treatment from d_{j-1} to d_j, and the "average" is a weighted one taken across all treatment indices j. The weights in the average are proportional to the probability of the event that $D_i(1) \geq d_j > D_i(0)$, and the causal effects are conditioned on this event. This event includes all

[21] All of the conceptual issues in the following discussion extend to the case of a continuous treatment, essentially by replacing sums with integrals and finite differences with derivatives; see (Angrist et al., 2000).

[22] Expression (22) is slightly more general than the one in Angrist and Imbens (1995) because they assume that the treatment is coded as $d_j = j$, so that each treatment value is one increment apart. A derivation of (22) is in Appendix C.

individuals i that would have treatment value larger than d_j with the instrument, but strictly smaller than d without it. Individuals unaffected by the instrument ($D_i(1) = D_i(0)$) do not contribute to the ACR. The effect of increasing treatment from d_{j-1} to d_j for individuals who would never have either treatment level (i.e. $D_i(0) \geq d_j$ or $D_i(1) \leq d_{j-1}$) does not enter into the ACR. These properties are sensible analogs of the binary treatment LATE identification result. The weights are non-negative, so the ACR is weakly causal. They do not sum to one unless $d_j - d_{j-1} = 1$ for all j, which is the case discussed in Angrist and Imbens (1995) and Angrist and Pischke (2009).[23]

The ACR has been criticized on the grounds that the conditioning events it represents are not mutually exclusive. That is, an individual with $D_i(1) = d_2$ and $D_i(0) = d_0$ gets "double counted" in both of the events $D_i(1) \geq d_1 > D_i(0)$ and $D_i(1) \geq d_2 > D_i(0)$. Angrist and Imbens (1995, pg. 435–436) and Heckman et al. (2006) both discuss this criticism, although the former authors downplay it on the grounds that they do not expect the instrument in their example to have more than a one unit effect. Heckman et al. (2006, Section VI) provide comparable reverse engineering results under an ordered response model and observe that the same criticism does not arise; see Section 4.5 for more discussion.

One way to address this criticism while sticking with the Angrist and Imbens (1995) monotonicity condition is to write the ACR as a different weighted average in which individuals only appear once.[24] Using the group notation $G_i \equiv (D_i(0), D_i(1))$, let $\mathcal{G} \equiv \{(g(0), g(1)) : g(1) \geq g(0)\}$ denote the subset of the $(J+1)^2$ possible groups that can have non-zero probability under the monotonicity condition. In Appendix C, we show that

$$\beta_{\text{IV}} = \sum_{g \in \mathcal{G}} \overbrace{\frac{\mathbb{P}[G_i = g](g(1) - g(0))}{\sum_{g' \in \mathcal{G}} \mathbb{P}[G_i = g'](g'(1) - g'(0))}}^{\text{weights reflect group size and instrument effect on treatment}} \underbrace{\mathbb{E}\left[\frac{Y_i(g(1)) - Y_i(g(0))}{g(1) - g(0)} \middle| G_i = g\right]}_{\text{average per-unit treatment effect for group } g}. \tag{23}$$

The conditioning events in this decomposition are mutually exclusive because each individual belongs to exactly one choice group. The treatment effects being weighted are expressed in per-unit averages across the range of values that the instrument shifts

[23] The fact that the weights do not sum to one is not necessarily a concern. Compare one IV estimand with $J = 2$ in which d_0, d_1, and d_2 are coded as $0, 1$ and 2, to another in which they are coded as $0, 2, 4$. The latter estimand will be half the size of the former, and its weights will also sum to one half. This is because the unit causal response does not depend on the coding of the treatment, implying that the weights must depend on it.

[24] This point was made by Frölich (2007, pg. 50). Equation (23) is an alternate phrasing of his equation (19).

a group's treatment choice. Researchers who find the ACR hard to appreciate due to overlapping conditioning events may find (23) more attractive.[25]

Even so, multivalued treatments are undoubtedly more complicated than binary treatments. Researchers are often tempted to binarize a treatment in order to avoid this complication. Angrist and Imbens (1995) show that this practice leads to a Wald estimand that is larger in magnitude than the ACR. Marshall (2016), Andresen and Huber (2021), and Rose and Shem-Tov (2023) consider additional assumptions under which the binarized Wald estimand has a more attractive interpretation.

3.7 Unordered or non-cardinal treatments

The previous reverse engineering results all consider linear IV specifications with the treatment as the sole endogenous variable. These specifications only make sense if the treatment has a natural cardinal ordering. If it doesn't, either because it's ordered but not cardinal, or because it's unordered, then the natural specification to consider is one with multiple endogenous variables that are indicators for different treatment states or sets of states.

The simplest setting is when D_i takes one of three treatment states, d_0, d_1, or d_2, which are coded up using two binary endogenous variables, $D_{i1} \equiv \mathbb{1}[D_i = d_1]$ and $D_{i2} \equiv \mathbb{1}[D_i = d_2]$. The order condition requires at least two excluded variables for a linear IV estimand to be defined. Suppose that we have access to an instrument Z_i that takes three values, $0, 1$, and 2, which have been similarly coded into two binary variables, Z_{i1} and Z_{i2}. Potential outcomes $Y_i(d_j)$ and potential treatments $D_i(z)$ are defined as before, with $D_{ij}(z) \equiv \mathbb{1}[D_i(z) = d_j]$ giving the implied potential binary treatment indicators. Full exogeneity is assumed, as usual.

With three treatment states and three instrument values there are $3^3 = 27$ a priori possible choice groups G_i reflecting different combinations of $(D_i(0), D_i(1), D_i(2))$. Suppose that we generalize the monotonicity condition to the assumption that instrument values one and two are targeted towards the corresponding first and second treatment states, and that receipt of these instrument values weakly pushes all individuals towards those states. The formal assumption is that $D_{i1}(1) \geq D_{i1}(0)$ and $D_{i2}(2) \geq D_{i2}(0)$, so that receiving $Z_i = j$ makes choosing $D_i = d_j$ more likely than when $Z_i = 0$. This eliminates 17 of the 27 choice groups, leaving the first ten shown in Table 2. The final row of Table 2 gives an example of a group that doesn't satisfy this monotonicity condition: $G_i = (d_1, d_0, d_2)$ would choose $D_i = d_0$ when $Z_i = 1$, but

[25] For the ACR decomposition (22), one could also multiply the weights by $d_j - d_{j-1}$ and divide the unit causal response by $d_j - d_{j-1}$ to get a third expression in which the weights sum to one.

would choose $D_i = d_1$ when $Z_i = 0$, violating the assumption that $Z_i = 1$ encourages takeup of state d_1 for everyone.

The linear IV estimand for this case is the one with outcome equation linear in a constant, D_{i1}, and D_{i2}, and first stage variables mirrored with a constant, Z_{i1}, and Z_{i2}. The coefficients on D_{i1} and D_{i2} can be written as a vector by partialling out the constant (demeaning) from the excluded variables:

$$\beta_{\text{IV}} \equiv \begin{bmatrix} \beta_{\text{IV},1} \\ \beta_{\text{IV},2} \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} \tilde{Z}_{i1} \\ \tilde{Z}_{i2} \end{bmatrix}\begin{bmatrix} D_{i1} \\ D_{i2} \end{bmatrix}'\right]^{-1} \mathbb{E}\left[\begin{bmatrix} \tilde{Z}_{i1} \\ \tilde{Z}_{i2} \end{bmatrix} Y_i\right] \quad \text{where} \quad \tilde{Z}_{ij} \equiv Z_{ij} - \mathbb{E}[Z_{ij}]. \quad (24)$$

For this unordered case, $\beta_{\text{IV},j}$ is weakly causal if it is non-negative whenever $\mathbb{E}[Y_i(d_j) - Y_i(d_0)|G_i = g]$ is non-negative for all g (Appendix B).

Kirkeboen et al. (2016) and Heinesen et al. (2022) show that neither of the components of β_{IV} are even close to weakly causal under the given monotonicity condition. Instead, $\beta_{\text{IV},1}$ captures a complicated weighted average of $Y_i(d_1) - Y_i(d_0)$ *and* $Y_i(d_2) - Y_i(d_0)$ involving all of the seven non-always-taker groups not ruled out by monotonicity.[26] Many of the weights will be negative. So $\beta_{\text{IV},1}$ fails to reflect the sign of $\mathbb{E}[Y_i(d_1) - Y_i(d_0)|G_i = g]$ both because it negatively weights some groups and because it also reflects potential outcomes for treatment state d_2. Symmetric conclusions apply to $\beta_{\text{IV},2}$.

The reason this happens is that the monotonicity condition does not sufficiently restrict treatment choice behavior. Groups with $D_i(1) = d_1$ who choose the first state when its instrument is switched on ($Z_i = 1$) could switch to choosing either d_0 (groups g_2, g_5, g_9) or d_2 (group g_8) when switched off. A contrast between $Z_{i1} = 1$ and $Z_{i1} = 0$ would not isolate a single treatment contrast even if it were able to keep Z_{i2} fixed, which the linear IV estimand (24) does not. More assumptions on treatment choice behavior are needed.

An early proposal by Behaghel et al. (2013) is to impose an "extended monotonicity" (EM) condition that eliminates groups g_7 through g_{10}. Their motivation was a multi-armed encouragement design in which individuals are given a specific encouragement to take the first or second treatment. The EM restriction rules out groups like g_7 who would enter into treatment d_2 when encouraged to, but would switch to treatment d_1 when encouraged to or when not encouraged at all. Under EM, the authors show that $\beta_{\text{IV},1}$ is equal to $\mathbb{E}[Y_i(d_1) - Y_i(d_0)|D_{i1}(1) = 1, D_{i1}(0) = 0]$, which is the average treatment effect of state d_1 relative to d_0 for the combined complier group $G_i \in \{g_2, g_5\}$ that responds to $Z_i = 1$. Similarly, $\beta_{\text{IV},2}$ is equal to

[26] The full expression can be found in Proposition 3 of Heinesen et al. (2022).

Table 2: Choice groups for an unordered treatment with three states

G_i	Mon.	EM	IR	NB	KLM
$(d_0, d_0, d_0) \equiv g_1$	✓	✓	✓	✓	✓
$(d_0, d_1, d_0) \equiv g_2$	✓	✓	✓	✓	✓
$(d_0, d_0, d_2) \equiv g_3$	✓	✓	✓	✓	✓
$(d_0, d_1, d_2) \equiv g_4$	✓	✓	✓	✓	✓
$(d_1, d_1, d_1) \equiv g_5$	✓	✓	✓		
$(d_2, d_2, d_2) \equiv g_6$	✓	✓	✓		
$(d_1, d_1, d_2) \equiv g_7$	✓		✓		
$(d_2, d_1, d_2) \equiv g_8$	✓		✓		
$(d_0, d_1, d_1) \equiv g_9$	✓			✓	
$(d_0, d_2, d_2) \equiv g_{10}$	✓			✓	
$\vdots$					
(d_1, d_0, d_2)					
$\vdots$					

Notes: *The groups shown are for a treatment that takes three states d_0, d_1, d_2 and an instrument that takes three values $0, 1, 2$ so that choice groups are determined by a combination of $G_i \equiv (D_i(0), D_i(1), D_i(2))$. There are 27 groups possible a priori. Only those satisfying the natural extension of the monotonicity condition (Mon.) are shown, together with one example of a group that violates that condition. Acronyms: EM (extended monotonicity), IR (irrelevance), NB (next-best), and KLM (Kirkeboen et al., 2016).*

$\mathbb{E}[Y_i(d_2) - Y_i(d_0)|D_{i2}(2) = 1, D_{i2}(0) = 0] = \mathbb{E}[Y_i(d_2) - Y_i(d_0)|G_i = g_4 \text{ or } g_5]$. These quantities have the hoped-for interpretation of average treatment effects for their respective treatment states relative to the omitted state d_0 for their respective subpopulation of compliers.

Kirkeboen et al. (2016) point out that EM is a strong restriction on preferences. Under EM, an individual who chooses state d_2 when its cost is low cannot switch to state d_1 when the cost of d_2 becomes high. Even in an experimental setting, this may not be an attractive assumption unless there is one-sided noncompliance. Behaghel et al. (2013) note that EM has the testable implication that $\mathbb{P}[D_i = d_2|Z_i = 1] = \mathbb{P}[D_i = d_2|Z_i = 0]$, meaning that the share of those choosing d_2 should be the same among those encouraged to choose d_1 and those not encouraged to choose either d_1 or d_2, with a similar implication for the probability that $D_i = d_1$. They reject EM for part of their experiment, as do Kirkeboen et al. (2016) in a different context.

Kirkeboen et al. (2016) consider two alternative assumptions. The first, which they call "irrelevance" (IR), is that if the instrument value that encourages d_2 doesn't actually induce d_2, then it doesn't instead induce d_1. This rules out group g_9, who

has $D_i(2) = d_1$ but $D_i(0) = d_0$. It also rules out group g_{10} symmetrically. The second assumption, which they call "next best" (NB), is that they can observe in the data who would choose d_0 if assigned $Z_i = 0$. In their application to estimating the returns to field of study, they justify the NB assumption by restricting their sample to individuals who list the same second-preferred field of study in a centralized college admissions process. Conditioning on $D_i(0) = d_0$ excludes members of groups g_7 and g_8.[27] Together, IR and NB leave the same choice groups as under EM, and so also imply that $\beta_{\text{IV},1}$ and $\beta_{\text{IV},2}$ reflect average treatment effects among their respective complier groups (g_2, g_5 and g_4, g_5), although now conditioned on the sample selection rule of having a particular next best alternative.

Bhuller and Sigstad (2024) show that the conditions used by Behaghel et al. (2013) and Kirkeboen et al. (2016) are necessary for β_{IV} to be weakly causal interpretation.[28] In particular, the necessary (and sufficient) condition is that the each instrument essentially only affects the indirect utility of one treatment state, and that the excluded treatment state d_0 is always either the preferred or next best alternative.[29] The only groups in Table 2 that satisfy this description are g_1 through g_6. The implication of the Bhuller and Sigstad (2024) result is that the linear IV estimand will only have a sensible interpretation if heterogeneity in choice behavior is heavily restricted, as in Behaghel et al. (2013), or if one has data on next-best alternatives, as in Kirkeboen et al. (2016). Heinesen et al. (2022) show how to conduct a sensitivity analysis similar to the one in Section 3.4, which suggests that given the monotonicity condition, the interpretation of the 2SLS estimand given in Kirkeboen et al. (2016) will be relatively insensitive to modest violations of either IR or NB (but not both) if the amount of treatment effect heterogeneity is also modest.

In some cases, the values d_0, d_1, and d_2 might have a natural ordering even if not a cardinal interpretation. An example studied by Arteaga (2023), Humphries et al. (2023b), and Kamat et al. (2024) is criminal case outcomes, where d_2 is incarcerate, d_1 is convict with no incarceration, and d_0 is do not convict. Bhuller and Sigstad (2024) characterize sufficient and necessary conditions for weak causality of the linear IV estimand in (24) coded up as $D_{i1} = \mathbb{1}[D_i \geq d_1]$ and $D_{i2} = \mathbb{1}[D_i = d_2]$. With constant effects, $\beta_{\text{IV},1}$ would be interpreted as the causal effect of d_1 relative to d_0 and the coefficient on $\beta_{\text{IV},2}$ would be interpreted as the incremental causal effect of d_2 relative

[27] It also eliminates the always-taker groups g_3 and g_6, however these groups get differenced out regardless.

[28] The Bhuller and Sigstad (2024) results apply to any number of J discrete treatment states with β_{IV} defined as the coefficients on $J-1$ indicators that are instrumented for by $J-1$ binary instruments.

[29] As usual in discrete response, the minor caveat suggested by "essentially" is the possibility that an instrument shifts the indirect utility of choosing a different state as long as the shift is inframarginal and does not result in a discrete change in choice behavior.

to d_1. Bhuller and Sigstad (2024) show that if treatment follows an ordered response model, then an additional parametric assumption of linearity between the first stage fitted values for the two treatments is necessary for a weakly causal interpretation. Humphries et al. (2023b) and Kamat et al. (2024) both conclude that the implied restrictions on choice behavior are unattractive in their setting.

Given these difficulties, one might consider using a linear IV estimand with only a single treatment indicator, such as D_{i2}, and instrumenting for it with Z_{i2} alone. The concern here is substitution bias (e.g. Heckman et al., 2000). As several authors have noted, the resulting IV estimand will be equal to a weighted average of treatment effects relative to both d_0 and d_1 depending on which treatment state would have been chosen if $Z_{i2} = 0$ (e.g. Kirkeboen et al., 2016; Kline and Walters, 2016; Mountjoy, 2022). The conflation of two different treatment contrasts makes such a quantity not weakly causal and generally makes it difficult to interpret even under restrictive assumptions on choice behavior; see Mountjoy (2022) for one clear example. Kline and Walters (2016) show that a quantity that conflates two different treatment contrasts can, however, still be useful for evaluating policy questions that do not require accounting for substitution bias. Humphries et al. (2023b) show how an interpretation in terms of a single treatment contrast can be restored under restrictions on choice behavior if the instruments are probabilities of assignment to the treatment states, as in a judge design, and one of the instruments is conditioned on as a covariate.[30]

What about multiple different treatment variables collected into a vector? This would be an unordered treatment even if the individual components are all ordered and cardinal. We are not aware of any work for this case on reverse engineering linear IV with UHTE. The difficulties encountered with the scalar unordered case suggest that the conclusion is unlikely to be inspiring.

3.8 Covariates

Covariates X_i have so far been absent from our discussion of reverse engineering. Yet they often play an important role in bolstering the credibility of the exogeneity assumption. Their inclusion complicates reverse engineering interpretations considerably.

[30] In particular, the authors maintain the unordered partial monotonicity condition used by Mountjoy (2022). They point out that this restriction has some unattractive implications for behavior in a judge design, which they find some evidence against in their data.

3.8.1 Controlling for covariates nonparametrically

Conditioning on covariates nonparametrically doesn't have any substantive implications for reverse engineering, since it only changes the subpopulation to which the arguments and assumptions must apply. For example, dropping observations that don't fall into the conditioning set leads to conditional versions of any of the previous interpretations. These can be aggregated or compared across conditioning sets as desired. Estimates from a crude binning approach like this may be too noisy to be useful. Nonparametric machine learning estimators are an appealing alternative, but they are not linear IV (see Section 4.2). The current dominant practice is instead to condition on covariates by controlling for them linearly.

3.8.2 Controlling for covariates linearly

Linearly controlling for a vector of covariates X_i changes the IV estimand to

$$\beta_{\text{IV}} \equiv \frac{\mathbb{E}[Y_i \tilde{Z}_i]}{\mathbb{E}[D_i \tilde{Z}_i]} \quad \text{where} \quad \tilde{Z}_i \equiv Z_i - \underbrace{X_i' \overbrace{\mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Z_i]}^{\text{population coefficients } \delta \text{ from regressing } Z_i \text{ onto } X_i}}_{\text{population fitted values from linear regression of } Z_i \text{ onto } X_i} \equiv Z_i - X_i'\delta. \tag{25}$$

Equation (25) has the same reduced-form-to-first-stage structure as the simple IV estimand without covariates (the right-hand side of equation (13)), except that the instrument is residualized against the covariates first, leaving an effective instrument, $\tilde{Z}_i$. This effective instrument contains variation due to both the instrument *and* the covariates. The two sources of variation can be separated in the reduced form (the numerator):

$$\overbrace{\mathbb{E}[Y_i \tilde{Z}_i]}^{\text{numerator of IV estimand}} = \mathbb{E}\left[\mathbb{E}[Y_i \tilde{Z}_i | X_i]\right] = \overbrace{\mathbb{E}\left[\mathbb{C}[Y_i, Z_i | X_i]\right]}^{\text{variation in } Y_i \text{ caused by } Z_i} + \overbrace{\mathbb{E}\left[Y_i \, \mathbb{E}[\tilde{Z}_i | X_i]\right]}^{\text{covariation between } Y_i \text{ and } X_i}. \tag{26}$$

The first term contains the type of variation we hope to extract with an IV estimator: the relationship between Y_i and Z_i, *conditional* on X_i. In contrast, the second term reflects variation between Y_i and a function of X_i, something that isn't part of the nonparametric rationale of an IV strategy.

The variation in the second term *is* part of the rationale of controlling for covariates in the classical model (8). This is because the random variable $\mathbb{E}[\tilde{Z}_i|X_i]$ contained in the second term is also orthogonal to X_i.[31] As a consequence, the second term—like

[31] Because $\tilde{Z}_i$ is the residual from a linear regression of Z_i onto X_i, $0 = \mathbb{E}[X_i \tilde{Z}_i] = \mathbb{E}[X_i \, \mathbb{E}[\tilde{Z}_i|X_i]]$.

the first term—still reflects the constant treatment effect, β_1:

$$\overbrace{\mathbb{E}\left[Y_i\,\mathbb{E}[\tilde{Z}_i|X_i]\right]}^{\text{substitute (8), } Y_i=\alpha_0+\alpha_1 D_i+\alpha_2 X_i+\epsilon_i} = \underbrace{\beta_1\,\mathbb{E}\left[D_i\,\mathbb{E}[\tilde{Z}_i|X_i]\right]}_{\text{treatment effect (scaled)}} + \beta_2\underbrace{\mathbb{E}\left[X_i\,\mathbb{E}[\tilde{Z}_i|X_i]\right]}_{=\mathbb{E}[X_i\tilde{Z}_i]=0}. \tag{27}$$

This happy simplification only occurs because of the structure of (8): a constant coefficient on D_i, and a linear adjustment for X_i. The very point of reverse engineering an interpretation for the IV estimand is to avoid these assumptions.

3.8.3 Level-dependence caused by covariates

Heterogeneous treatment effects in particular imply that the second term of (26) will generally reflect problematic variation. Consider the binary treatment, binary instrument case. The intuition of the baseline LATE argument is not that the always-takers and never-takers disappear, but rather that their contribution is differenced out in the instrument contrast represented by the numerator of the Wald estimand. This differencing occurs in the first term of (26), but not in the second. As a consequence, the second term reflects outcomes for the always- and never-takers, outcomes that inherently do not involve the causal effect of the treatment, because treatment does not vary for these groups. The implication is that the IV estimand is picking up not just the effect of the treatment on outcomes, but also the level of the outcome itself. Blandhol et al. (2022) call this phenomenon level-dependence and show that an estimand that is level-dependent cannot be weakly causal.[32]

The size of the second term of (26) is mediated by the instrument residual, $\tilde{Z}_i$, and in particular by its conditional mean

$$\mathbb{E}[\tilde{Z}_i|X_i] = \mathbb{E}[Z_i|X_i] - X_i'\delta. \tag{28}$$

This quantity reflects the difference between the conditional mean of Z_i given X_i and the linear regression of Z_i onto X_i. The latter provides the best linear approximation to the conditional mean in terms of mean squared error, so it is only zero if the covariate specification is sufficiently flexible to exactly reproduce the conditional mean of the instrument. Blandhol et al. (2022) say that IV specifications with this property have

[32]The intuition is that an estimand that depends on potential outcome levels could have any value—and any sign—even if all of the underlying treatment effects are positive. By contrast, for an estimand that only depends on treatment effects, weak causality is only a matter of whether the weights on these treatment effects are non-negative.

"rich covariates." They show that IV estimands for specifications that do not have rich covariates will necessarily be level-dependent and will therefore not be weakly causal.

There are two cases in which rich covariates is not a controversial assumption.

The first is when the instrument is independent of the covariates, so that $\mathbb{E}[Z_i|X_i] = \mathbb{E}[Z_i]$ is fit exactly as long as X_i contains a constant. This can happen when the instrument is experimentally assigned, but also in some natural experiments. Fuzzy regression discontinuity designs implemented with local IV estimators (e.g. Hahn et al., 2001; Calonico et al., 2014) are another example where this occurs, albeit in a limiting sense. While controlling for covariates is not necessary for exogeneity in these settings, researchers still often do so to reduce residual variation in the outcome and/or treatment variables, which can improve statistical precision.

The second case is when the covariate specification is so flexible that it cannot fail to be rich. When discussing linear IV estimands with covariates, Angrist and Pischke (2009) analyze a specification they call "saturate and weight," which controls for covariates nonparametrically by including an indicator for each covariate value.[33] Controlling for covariates in this way ensures that $\mathbb{E}[\tilde{Z}_i|X_i] = 0$, so that the second term of (26) disappears, and the IV estimand is not level-dependent. However, saturate and weight is a ravenously data-hungry specification, and it tends to produce noisy and poorly-behaved IV estimators. Blandhol et al. (2022) report a survey of IV papers which indicates that saturating in covariates is uncommon in practice.

Outside of these two cases, assuming that a specification has rich covariates is a parametric assumption. The Blandhol et al. (2022) analysis shows that having rich covariates is necessary for an IV estimand to have a weakly causal interpretation, and therefore also necessary for the IV estimand to have some sort of interpretation as a convex average of LATEs when the treatment is binary. Interpreting IV estimands with covariates as reflecting LATEs therefore implicitly requires the parametric assumption that $\mathbb{E}[Z_i|X_i]$ is linear in X_i, a conclusion that is uncomfortably at odds with the motivation of reverse engineering as providing an interpretation that is robust to misspecification. Blandhol et al. (2022) point out that that rich covariates can be tested, for example with Ramsey's (1969) RESET test.

[33] This requires assuming that the covariates are discrete or have been adequately discretized. The saturate and weight specification was originally Theorem 3 in Angrist and Imbens (1995).

3.8.4 Weighting expression for linear IV under rich covariates

If rich covariates holds, then (25) can be written as

$$\beta_{\text{IV}} = \mathbb{E}\left[\frac{\mathbb{C}[D_i, Z_i|X_i]}{\mathbb{E}[\mathbb{C}[D_i, Z_i|X_i]]}\frac{\mathbb{C}[Y_i, Z_i|X_i]}{\mathbb{C}[D_i, Z_i|X_i]}\right] = \mathbb{E}\left[\frac{\mathbb{C}[D_i, Z_i|X_i]}{\mathbb{E}[\mathbb{C}[D_i, Z_i|X_i]]}\beta_{\text{IV}}(X_i)\right], \tag{29}$$

where $\beta_{\text{IV}}(x)$ is the linear IV estimand with no covariates (other than a constant), but now conditional on the subpopulation with $X_i = x$.[34] The interpretation of $\beta_{\text{IV}}(x)$ can be considered for each $X_i = x$ subpopulation without concern about linear extrapolation across these subpopulations. If $\beta_{\text{IV}}(x)$ is weakly causal for each x, then (29) shows that β_{IV} will be weakly causal if and only if $\mathbb{C}[D_i, Z_i|X_i = x] \geq 0$ for all x. This latter condition says that the sign of the first stage relationship is the same for all $X_i = x$. It has a close connection to the monotonicity condition.

3.8.5 Monotonicity-correct first stage specifications

The monotonicity condition we have been considering so far requires the instrument to operate in the same direction for all individuals. One can weaken this to allow the direction of the effect to depend on each individual's covariates, so that

$$\mathbb{P}[D_i(1) \geq D_i(0)|X_i = x] = 1 \quad \text{or} \quad \mathbb{P}[D_i(0) \geq D_i(1)|X_i = x] = 1 \quad \text{for all } x. \tag{30}$$

Słoczyński (2020) describes (30) as weak monotonicity. Strong monotonicity by contrast is the assumption that we have previously been working with, where the directional effect of the instrument does not change when conditioning on covariates. Vytlacil's (2002) equivalence theorem continues to hold under weak monotonicity if the instrument and covariates are interacted in the threshold-crossing model.

Is weak monotonicity appreciably weaker than strong monotonicity? Many objections to the monotonicity condition are about ruling out heterogeneity in treatment choice due to unobservables, such as preferences or costs. Weak monotonicity doesn't address those concerns. In fact, in judge designs researchers commonly test mono-

[34] Multiplying and dividing by $\mathbb{C}[D_i, Z_i|X_i]$ in (29) raises the potential concern that this term may be zero with positive probability. This turns out to not be a concern in the case considered here because $\mathbb{C}[Y_i, Z_i|X_i = x]$ will be zero whenever $\mathbb{C}[D_i, Z_i|X_i = x]$ is zero. To see why, suppose that $\mathbb{C}[D_i, Z_i|X_i = x] = \mathbb{E}[D_i(1) - D_i(0)|X_i = x]\,\mathbb{V}[Z_i|X_i = x] = 0$. If this is because the first term is zero, then the monotonicity condition implies that $\mathbb{P}[D_i(1) - D_i(0) = 0|X_i = x] = 1$. In this case, full exogeneity implies that $\mathbb{C}[Y_i, Z_i|X_i = x] = \mathbb{E}[(D_i(1) - D_i(0))(Y_i(1) - Y_i(0))|X_i = x]\,\mathbb{V}[Z_i|X_i = x] = 0$ as well. Alternatively, if $\mathbb{V}[Z_i|X_i = x] = 0$, then $\mathbb{C}[Y_i, Z_i|X_i = x] = \mathbb{C}[D_i, Z_i|X_i = x] = 0$. Instead of indicating for the event that $\mathbb{C}[D_i, Z_i|X_i] \neq 0$ in (29), we just define $\mathbb{C}[D_i, Z_i|X_i = x]\beta_{\text{IV}}(x) = 0$ whenever $\mathbb{C}[D_i, Z_i|X_i = x] = 0$ to keep the expression cleaner.

tonicity by examining the sign of the first stage relationship across cases with different covariates (see e.g. Dobbie and Song, 2015; Dobbie et al., 2018; Bhuller et al., 2020; Norris et al., 2021). This exercise can be interpreted as a suggestive test against strong monotonicity, but it wouldn't be appropriate as a test of weak monotonicity, which would allow for these sign changes. This leads us to believe that researchers are concerned with failures of strong monotonicity, not weak monotonicity, at least in these contexts. An exception is given by Mueller-Smith (2015), who is explicit in preferring weak monotonicity to strong monotonicity.

Słoczyński (2020) considers the implications of weak and strong monotonicity for interpreting linear IV estimands in a setting with a binary treatment and binary instrument, taking rich covariates as given. He shows that if weak monotonicity holds but strong monotonicity does not, then the linear IV estimand (29) that uses only Z_i as an excluded instrument will reflect negatively-weighted complier treatment effects, and so will fail to be weakly causal. The reason is that when D_i and Z_i are both binary, $\beta_{\text{IV}}(x)$ is a LATE for the subgroup with $X_i = x$, while $\mathbb{C}[D_i, Z_i|X_i = x]$ is positive if the direction of monotonicity is the first case of (30) and negative if it is the second case. If weak monotonicity holds but strong monotonicity does not, then the sign will be different for some values of x, so compliers for some subpopulations will necessarily be negatively-weighted.

These sign reversals arise because the first stage for β_{IV} has a single excluded instrument Z_i and so is not flexible enough to capture the changes in the direction of monotonicity allowed under weak monotonicity. A 2SLS specification that includes interactions between X_i and Z_i as excluded variables is more flexible in this regard and can restore non-negativity in the weights (Słoczyński, 2020). Blandhol et al. (2022) extend this concept to 2SLS specifications with non-binary endogenous variables and instruments, describing a specification as "monotonicity-correct" if it is sufficiently flexible to capture covariate-mediated changes in monotonicity. They show that given rich covariates, monotonicity-correctness is the additional sufficient and necessary condition for a 2SLS estimand to be weakly causal.

3.8.6 Specification considerations with covariates

There are two specification considerations when using covariates with linear IV. First, are the covariates rich? If not, then the linear IV estimand reflects potential outcome levels, not just treatment effects, and so is not weakly causal. Second, assuming that covariates are rich, does strong or weak monotonicity hold and, if only weak monotonicity holds, have enough instrument-covariate interactions been included in the first

stage to capture changes in monotonicity? If not, then the linear IV estimand reflects some negatively-weighted treatment effects, and so is also not weakly causal.

The saturate and weight specification in Angrist and Pischke (2009) addresses both considerations by including a full set of instrument-covariate interactions in the first stage. Doing so is a dangerous invitation to many instruments bias, the phenomenon whereby an overfit first stage leads to a IV estimate that is similar to an OLS estimate (e.g. Bekker, 1994). Blandhol et al. (2022) provide both simulation and empirical evidence that the saturate and weight specification is irredeemably contaminated by many instruments bias in realistic use cases.[35]

On the other hand, (29) shows that if one is willing to maintain strong monotonicity, then including instrument-covariate interactions in the first stage is usually not necessary for securing a weakly causal interpretation. The sole requirement for the IV estimand to be weakly causal in this case is that it has rich covariates, which is a matter of the flexibility in which covariates are controlled for, not their interactions with the instrument in the first stage. In particular, a specification that is saturated in covariates, but still only uses a single excluded variable Z_i, will generally produce a weakly causal linear IV estimand under strong monotonicity.[36] This is the natural counterpart to the Angrist and Pischke (2009) saturate and weight specification, but does not suffer from the pitfalls of many instruments.

Even so, saturating in X_i will often still be too statistically demanding. For these cases, rich covariates must be maintained as a substantive parametric assumption to ensure that linear IV is a weakly causal estimand. A natural alternative is to use flexible machine learning methods to help select the functional form in which the covariates enter, but this takes one outside of the realm of linear IV; see Section 4.2 for more detail.

3.9 Summary of reverse engineering

Table 3 summarizes reverse engineering interpretations for linear IV estimands across the different cases we have considered: the baseline case with a binary treatment and binary instrument, multivalued (or multiple) instruments, multivalued ordered treatments, multivalued unordered treatments, and specifications that control for covariates.

[35] Blandhol et al. (2022) provide simulation evidence that jackknife IV estimators (Phillips and Hale, 1977; Angrist et al., 1999; Kolesár, 2013) can struggle with the herculean task of correcting the massive many instruments bias introduced by saturate and weight. The limited information maximum likelihood (LIML) estimator, which is also sometimes suggested as a solution for many instruments (e.g. Bekker, 1994; Hansen et al., 2008), was shown by Kolesár (2013) to generally not be weakly causal.

[36] There are some minor caveats here if the instrument is multivalued; see Blandhol et al. (2022) for a precise statement.

The only case in which the linear IV estimand has an unqualified interpretation as a LATE is the first one, the baseline case. The baseline case is commonly the exclusive focus of discussions of the LATE idea, see e.g. textbook discussions by Wooldridge (2010, Section 21.4.3) and Hansen (2022b, Section 12.34), the Nobel lectures by Angrist (2022) and Imbens (2022), or the scientific background provided by the Nobel committee itself (Nobel Committee, 2021). But it is rarely the setting in which empirical work using linear IV is actually conducted.[37]

Blandhol et al. (2022) find that empirical researchers routinely describe their linear IV estimates using LATE language as if they were in the baseline case. The source of this confusion may be due to the way LATE interpretations have been discussed in some influential texts. An early example is the *Handbook of Labor Economics* chapter by Angrist and Krueger (1999, pg. 1326), who wrote

> *Finally, we note that the discussion of IV in heterogeneous and non-linear models so far has ignored covariates . . . IV estimates in models with covariates can be thought of as producing a weighted average of covariate-specific Wald estimates* [conditional LATEs] *as long as the model for covariates* [satisfies "saturate and weight"]. *In other cases it seems reasonable to assume that some sort of approximate weighted average is being generated, but we are unaware of a precise causal interpretation that fits all cases.*

The precise (sufficient and necessary) causal interpretation that Angrist and Krueger conjecture about was only recently provided by Blandhol et al. (2022), whose results show that their reasonable assumption is not true (see Section 3.8). In their hugely influential monograph, Angrist and Pischke (2009, pg. 173) make a similar but less circumspect assertion:

> *The econometric tool remains 2SLS and the interpretation remains fundamen-*

[37] Blandhol et al. (2022) show that the vast majority of empirical studies using IV control for covariates in a way that suggests they are not just being used to improve precision, while Mogstad et al. (2021) find that over 40% use multiple instruments.

We have not included a discussion of longitudinal settings out of space considerations. The few reverse engineering results on this topic are quite negative, even in simple settings. Blundell and Dias (2009, pp. 589-591) considered difference-in-differences IV strategies in which differential changes over time in the treatment between two groups serves as an instrument. They showed that the corresponding Wald estimand can be interpreted as a LATE for the exposed group if the treatment effect is constant over time and treatment rates do not change in the unexposed group. If they do change in the unexposed group, then the estimand is a weighted average of LATEs between the two groups, with weights that can be negative. This result later appeared in de Chaisemartin and D'Haultfoeuille (2018, Theorem 1), who emphasized the importance and restrictiveness of the time-constant treatment effect assumption. Miyaji (2024a,b) has recently considered the implementation of IV-DID with staggered events through two-way fixed effects, complications that certainly don't make reverse engineering any more attractive (Bailey and Goodman-Bacon, 2015; Sun and Abraham, 2021). de Chaisemartin and Lei (2023) and Hahn et al. (2024) have shown that it is difficult to ensure a weakly causal interpreted for linear IV estimands used with Bartik instruments.

> *tally similar to the basic LATE result, with a few bells and whistles . . . These results provide a simple casual* [sic] *interpretation for 2SLS in most empirically relevant settings.*

Sweeping descriptions like these have been repeated more recently, for example in Cunningham's (2021, pg. 351) popular monograph:

> *The intuition of LATE generalizes to most cases where we have continuous endogenous variables and instruments, and additional control variables, as well.*

Statements like these are, at best, only true subject to many unstated caveats.

One reaction to the many caveats of reverse engineered LATE interpretations is to downplay the theory. Does any of this actually matter "in practice?" It's an interesting question because by their nature reverse engineering arguments are creatures of theory: they do not change practice, they only change interpretation. This probably explains their seductive appeal to busy empirical researchers, and the understandable desire to stretch the theory to fit cases that it does not.

In our view, focusing such intense attention on the reverse-engineered interpretation of a single number like the linear IV estimand makes it more—not less—important for the theory to accurately reflect practice. With forward engineering, the consequences of changing the estimation procedure can be seen directly in the results; the estimates change, but the estimand stays fixed. Reverse engineering, by contrast, cannot be "seen" in the results, but is instead a matter of theoretical justification, the subtle assumptions of which are easy to sweep under the rug in practice.

Reverse engineering can also create problems in other, unexpected places. A clear if mundane example of this was emphasized by Lee (2018), who points out that the usual standard errors for overidentified 2SLS estimators, such as those commonly used with multiple or multivalued instruments, are not correct if there is treatment effect heterogeneity and heteroskedasticity. The classical derivation of these standard error formulas makes use of the assumption that the linear IV model is correctly specified as having constant treatment effects. The derivation omits a term that shows up if the model is misspecified due to UHTE. Imbens and Angrist (1994, Theorem 3) recognized this point, but it seems to have been forgotten in the subsequent empirically-oriented literature, including Angrist and Pischke (2009). Whether this point is substantively important is debatable, and Lee's empirical illustrations suggest that it may not be. Nevertheless, it provides a clear example of how reverse engineering makes it easy for practitioners to ignore the theory that justifies their practice.

There's a separate issue of whether these reverse-engineered interpretations—when properly invoked—actually answer useful questions. The baseline LATE result in Sec-

Table 3: Reverse engineering linear IV estimands

Treatment (D_i)	Instruments (Z_i)	Covariates (X_i)	Summary
Bin.	**Bin.**	**No**	The Wald and simple linear IV estimands are equal to each other and equal to the LATE under monotonicity and full exogeneity.
Bin.	**Mul.**	**No**	Each pair of instrument values defines a different complier group with an associated LATE. Different linear IV estimands produce different weighted averages of LATEs. 2SLS with a saturated instrument specification leads to non-negative weights. The weights can be negative with non-saturated specifications, but will be non-negative if the specification reproduces the monotonicity order of the instruments. The monotonicity condition can be especially unattractive if the multivalued instrument is not ordered, for example in judge designs, or when there are multiple instruments.
Ord.	**Any**	**No**	If the instrument is binary, and the treatment is a single scalar cardinal variable, then the linear IV estimand can be interpreted as the average causal response (ACR). The ACR can in turn be interpreted either as an average treatment effect among overlapping groups whose treatment choice is shifted by the instrument, or an average per-unit treatment effect across all (disjoint) complier groups. The second interpretation is a natural generalization of the LATE from the binary treatment case. If the instrument is multivalued, then these generalized LATEs get averaged according to different instrument contrasts, the same way as in the binary treatment case, and with the same caveats. Ordered treatments that are not cardinal are better analyzed through the unordered treatment case.
Uno.	**Any**	**No**	The linear IV estimand in this case has indicators for each treatment state, except for the excluded state, which is captured by a constant. If there are instruments that affect each treatment state, then the two linear IV estimands will be weakly causal if and only if each instrument affects choices only in its targeted treatment state and the excluded state is always the preferred or next best choice. Achieving this requires strong behavioral restrictions or data on next best choices. With ordered treatments that are not cardinal there are possibilities for restoring a weakly causal interpretation, but they are complicated; see main text.
Any	**Any**	**Yes**	Two assumptions are required for a linear IV estimand to be interpretable as a convex weighted average of LATEs: rich covariates and a monotonicity-correct first stage. Rich covariates is often satisfied in randomized experiments, but may not be satisfied when the instrument is not independent of covariates. The first stage will usually be monotonicity-correct under strong monotonicity, but under weak monotonicity it will only be monotonicity-correct if it includes covariates in a way that is flexible enough to account for changes in the direction of monotonicity across covariates.

Notes: *This table summarizes the discussion in Section 3.*

tion 3.2 is clearly useful in some settings, but this is arguably also not really an example of reverse engineering, as it is based on a nonparametric estimand (the Wald estimand) that simply happens to be equal to a linear IV estimand in the baseline setting. The interpretations with a multivalued ordered treatment and binary instrument in Section

3.6 also seem useful: with a single binary instrument it is difficult to explore nonlinearity, so a sensibly-averaged summary of the nonlinearity seems like the best one can hope to identify nonparametrically.

For the other cases, it is less clear that these reverse engineering interpretations provide much useful information about causality. An ideal interpretation with convex weights allows one to conclude that the estimand lies somewhere between the smallest and largest average subgroup-specific treatment effects. This conclusion is more informative the less treatment effect heterogeneity there is. With substantial treatment effect heterogeneity—perhaps even effects of different signs—knowing that the weights are convex is not particularly conclusive. The implication is that reverse engineering interpretations work best at providing an interpretation robust to omitted UHTE exactly when this form of misspecification is a lesser concern.

The underlying problem is the form of the weights. For many reverse engineering interpretations of linear IV, the weights reflect statistical features rather than substantive concerns. The clearest example of this is with multivalued instruments, where linear IV estimands were seen to be interpretable as a weighted average of different LATEs, but the weights depended on the marginal distribution of the instrument, as well as on the choice of linear IV estimator. Different experimenters operating under different budgets or making different but sensible choices of evaluation could reach different conclusions in the same economic environment. This is clearly unappealing.[38]

Changing the weights would lead to more interpretable estimands without these drawbacks. In the multivalued instrument case, weights that were equal or given by the respective complier shares would produce a quantity with a clear interpretation. It would be invariant to the distribution of the instrument and could be defined without reference to a choice of estimator. But for the same reason, such a quantity is unlikely to arise from a reverse engineering mindset as the estimand to some commonly-used estimator: reverse engineering *starts* with the estimator. Estimating target parameters with purposefully-chosen weights requires forward engineering.

[38]Statistically-driven weights that appear in reverse engineering expressions have also been argued to be objectionable on other grounds. Słoczyński (2022) argues that the weights that appear in reverse-engineered interpretations of the OLS estimand under selection on observables have counterintuitive properties. Słoczyński (2020) argues that the weights that appear in linear IV estimands can be difficult to interpret in some cases. Balla-Elliott (2023) argues that the weights that appear in linear IV estimands are likely to systematically understate the causal effects of beliefs in information provision experiments.

4 Forward Engineering: Estimating Target Parameters

A forward engineering approach starts with a model and then constructs estimators under the assumption that the model is correctly specified. This is arguably traditional approach taken by economists, with the earliest examples being the Gronau-Heckman selection model (Gronau, 1974; Heckman, 1974, 1976). In this section we survey approaches to forward engineering with a eye towards more recent practice.

4.1 Assuming away the problem

The simplest "solution" to the difficulties raised by UHTE is to assume that, in fact, there is no such unobserved heterogeneity. Angrist and Fernández-Val (2013) reintroduce this assumption under the description of "conditional effect ignorability" (CEI). Stated in our notation for the binary treatment, binary instrument case with covariates, their Assumption 3 is that

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|G_i = g, X_i = x]}_{\text{LATE}(x) \text{ when } g=(0,1)} = \overbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]}^{\text{ATE}(x) \text{ — the average treatment effect given } X_i = x} \quad \text{for all groups } g. \tag{31}$$

The CEI assumption allows for treatment effect heterogeneity across observable covariates X_i but assumes that the unobservably-different choice groups of always-takers, never-takers, and compliers have the same average treatment effects.

Under the CEI assumption, the ATE is equal to the average of covariate-specific LATEs. This can be seen by applying the law of iterated expectations with (31):

$$\mathbb{E}\left[Y_i(1) - Y_i(0)\right] = \mathbb{E}\left[\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i]}_{\text{ATE}(X_i)}\right] \overbrace{= \mathbb{E}\left[\text{LATE}(X_i)\right]}^{\text{by force of CEI assumption (31)}}. \tag{32}$$

Angrist and Fernández-Val (2013) propose estimating covariate-specific LATEs and then averaging them into the ATE, or into the average treatment on the treated/untreated (ATT/ATU), to which similar arguments apply. They call this argument "LATE-Reweight," but given the strength of the CEI assumption, one could equally well describe it as "ATE-Reweight." Such a reweighting scheme can be implemented in the classical linear model by including treatment-covariate interaction terms and then summarizing the resulting observable treatment effect heterogeneity however one sees fit. Angrist and Fernández-Val (2013) maintain both monotonicity and full exogeneity, but neither are required given the strong CEI assumption.

Assuming away unobserved heterogeneity in treatment effects—whether phrased in

the classical model or with CEI—is a strong assumption. Assumptions are necessary for causal inference and strong assumptions may be justified in difficult empirical problems. But it's important to remember the economic considerations drove the concern about UHTE to begin with (Section 2.2).

In describing why fertility and female labor supply are endogenous—but before introducing the CEI— Angrist and Fernández-Val (2013, pg. 406) write:

> *Mothers with weak labor force attachment or low earnings potential may be more likely to have children than mothers with strong labor force attachment or high earnings potential.*

This description nearly precludes the CEI: mothers with weak labor force attachment may be more likely to have children because the labor supply impacts are different than for mothers with strong labor force attachment. The only way in which the description of endogeneity can coexist with the CEI is if less fertile mothers would still work more than more fertile mothers even in the counterfactual world in which both types of mothers have the same number of children. A story like this rules out UHTE by disconnecting the labor supply and fertility decisions, thereby weakening the very motivation for using an IV to begin with.

4.2 Estimating LATEs and ACRs in the presence of covariates

The difficulties encountered when reverse engineering linear IV estimands with covariates (Section 3.8) were more mechanical than conceptual. Conceptually, nothing about the LATE identification argument changed, at least for the binary treatment and binary instrument case; conditional-on-covariate LATEs were identified and could be estimated cell-by-cell and then aggregated in whatever way desired. The mechanical problem was with the linear IV estimand, which was not guaranteed to implement a convex weighting without additional implicit (and typically unstated) assumptions. Even if these assumptions were met, the weighting implemented by the linear IV estimand was statistical, making it difficult to interpret the resulting quantity, and difficult to transfer it across settings.

A well-developed econometric literature solves these problems by forward engineering direct estimators of the unconditional LATE. The estimators are primarily designed for the case of a binary treatment and a binary instrument, although they can also be applied with a multivalued ordered treatment, in which case the target parameter is the unconditional ACR. In this section, we discuss two types of estimators and then illustrate their implementation in an empirical example.

4.2.1 Propensity score weighting

A binary instrument allows for a fruitful connection with the large literature on estimation under selection on observables (e.g. Imbens, 2015). Let

$$W_i(z) \equiv Y_i(D_i(z)) = Y_i(0) + D_i(z)(Y_i(1) - Y_i(0))$$

denote the potential outcome for Y_i associated with a manipulation of the instrument, Z_i, with $Y_i = (1-Z_i)W_i(0)+Z_iW_i(1)$. Full exogeneity implies that $W_i(z)$ is independent of Z_i, conditional on X_i. The average treatment effect of Z_i on Y_i—often called the intent to treat (ITT)—is then identified by averaging covariate-conditioned contrasts, assuming an appropriate overlap condition:

$$\underbrace{\mathbb{E}\left[\,\mathbb{E}[Y_i|Z_i=1,X_i] - \mathbb{E}[Y_i|Z_i=0,X_i]\right]}_{\text{assuming } 0<\mathbb{P}[Z_i=1|X_i]<1 \text{ (instrument overlap)}} = \overbrace{\mathbb{E}[W_i(1)-W_i(0)]}^{\text{the ATE of } Z_i \text{ on } Y_i \text{ (the ITT)}}. \tag{33}$$

If D_i is binary and exclusion and strong monotonicity are satisfied, then the ITT only reflects treatment effects for the compliers:

$$\mathbb{E}[W_i(1) - W_i(0)] = \mathbb{E}[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0))] = \textsc{late} \times \mathbb{P}[G_i = \textsc{cp}], \tag{34}$$

where $\textsc{late} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|G_i = \textsc{cp}]$ is the unconditional LATE. The proportion of compliers is itself identified as the average treatment effect of Z_i on D_i, so can be written in an analogous form:

$$\mathbb{E}\left[\mathbb{E}[D_i|Z_i=1,X_i] - \mathbb{E}[D_i|Z_i=0,X_i]\right] = \mathbb{E}[D_i(1) - D_i(0)] = \mathbb{P}[G_i = \textsc{cp}]. \tag{35}$$

Putting together (33)–(35) gives

$$\textsc{late} = \frac{\mathbb{E}[W_i(1) - W_i(0)]}{\mathbb{E}[D_i(1) - D_i(0)]} = \frac{\mathbb{E}\left[\mathbb{E}[Y_i|Z_i=1,X_i] - \mathbb{E}[Y_i|Z_i=0,X_i]\right]}{\mathbb{E}\left[\mathbb{E}[D_i|Z_i=1,X_i] - \mathbb{E}[D_i|Z_i=0,X_i]\right]}, \tag{36}$$

which is an X_i-averaged reduced form divided by an X_i-averaged first stage.[39] A similar argument can be used when D_i is multivalued and ordered, with the change being that the right-hand side of (36) identifies the ACR in (22), rather than the LATE (see Appendix D).

[39] Similar arguments can be used to construct an expression the average treatment effect for treated or untreated compliers (Hong and Nekipelov, 2010). A treated complier has $Z_i = 1$ while an untreated complier has $Z_i = 0$. They are probabilistically identical groups without covariates, but with covariates they can differ because the distribution of X_i varies conditional on Z_i.

Equation (36) was derived by Tan (2006) and Frölich (2007), who also explicitly made the connection to the problem of estimating the ATE under selection on observables. This connection is helpful because it suggests applying one of the roughly three types of approaches used for that problem: imputation, propensity score weighting, and doubly-robust estimators that combine imputation and weighting. The propensity score referred to here is for the instrument, denoted as $q(x) \equiv \mathbb{P}[Z_i = 1|X_i = x] = \mathbb{E}[Z_i|X_i = x]$ with $Q_i \equiv q(X_i)$. Notice that q was the same object that needed to be correctly specified for a linear IV estimand to satisfy the rich covariates condition necesary for a weakly causal interpretation (Section 3.8.3).

An imputation approach constructs estimators of each of the conditional means in (36) and then averages across the distribution of X_i. Frölich (2007) derived the asymptotic properties of nonparametric series and local polynomial imputation estimators, however the usual curse of dimensionality suggests these will tend to perform poorly with more than a couple of covariates. Hirano et al. (2000), Yau and Little (2001), and Tan (2006) considered imputation with parametric estimators. Matching on the instrument propensity score Q_i is another possible imputation approach, which was suggested by Frölich (2007, Section 4), but does not seem to have been pursued further in the literature.

Propensity score weighting has been more widely analyzed. In the context of (36), the appropriate weighting expressions are

$$\mathbb{E}[W_i(1)] = \mathbb{E}\left[\frac{Y_i Z_i}{Q_i}\right] \quad \text{and} \quad \mathbb{E}[W_i(0)] = \mathbb{E}\left[\frac{Y_i(1-Z_i)}{1-Q_i}\right], \tag{37}$$

with analogous expressions for $D_i(1)$ and $D_i(0)$. The attraction of propensity score weighting is that only a single function $q(x)$ needs to be modeled and estimated by $\hat{q}(x)$, after which simple sample analogs of the weighting expressions can be formed from $\hat{Q}_i \equiv \hat{q}(X_i)$ and combined to estimate LATE via (36):

$$\frac{\frac{1}{n}\sum_{i=1}^n Y_i Z_i/\hat{Q}_i - \frac{1}{n}\sum_{i=1}^n Y_i(1-Z_i)/(1-\hat{Q}_i)}{\frac{1}{n}\sum_{i=1}^n D_i Z_i/\hat{Q}_i - \frac{1}{n}\sum_{i=1}^n D_i(1-Z_i)/(1-\hat{Q}_i)}. \tag{38}$$

Weighting estimators like this were proposed by Frölich (2007), Tan (2006), Uysal (2011), MaCurdy et al. (2011), and Donald et al. (2014), and have been revisited more recently by Heiler (2022), Sun and Tan (2022), Singh and Sun (2024), and Słoczyński et al. (2024). The latter authors emphasize the importance of normalizing the weights

so that the first term in the numerator of (38) is replaced by

$$\left[\sum_{i=1}^{n} Y_i Z_i / \hat{Q}_i\right] \Big/ \left[\sum_{i=1}^{n} Z_i / \hat{Q}_i\right], \tag{39}$$

and similarly for the other three terms.

The weighting expressions in (37) follow as special cases from a more general argument developed by Abadie (2003). Abadie (2003) showed that for any function ψ of Y_i, D_i, and X_i,

$$\begin{aligned} \mathbb{E}[\psi(Y_i, D_i, X_i)|G_i = \text{CP}] &= \frac{1}{\mathbb{P}[G_i = \text{CP}]} \mathbb{E}\left[\kappa_i \psi(Y_i, D_i, X_i)\right] \\ \text{where} \quad \kappa_i &\equiv 1 - \frac{D_i(1 - Z_i)}{1 - Q_i} - \frac{(1 - D_i)Z_i}{Q_i}, \end{aligned} \tag{40}$$

a result that has been called "Abadie's κ" by Angrist and Pischke (2009).[40] The most common use of Abadie's result is to estimate the distribution of X_i among compliers by taking $\psi(X_i) = X_i$ (e.g. Marx and Turner, 2019; Leung and O'Leary, 2020; Goodman et al., 2020).

Another application of Abadie's result is to estimate parametric specifications of $\mathbb{E}[Y_i(1) - Y_i(0)|G_i = \text{CP}, X_i = x]$, which (40) shows can be achieved by taking ψ to be an appropriate criterion function. For example, if ψ is taken to be a least squares criterion for a linear regression of Y_i onto D_i and X_i, then minimizing the left-hand side of (40) corresponds to running this regression only among compliers, while minimizing the right-hand side corresponds to running a κ_i-weighted regression among the entire population. The former is of interest because $D_i = Z_i$ is exogenous for the subpopulation of compliers (conditional on X_i), but infeasible because compliers are not observed. The latter is feasible because κ_i can be estimated by substituting $\hat{Q}_i$ for Q_i. Using Abadie's result in this way requires correctly specifying both q and the functional form of the linear controls in the weighted regression. The latter condition ends up being the same as requiring rich covariates (Blandhol et al., 2024). The regression-based approach can also potentially be used to examine heterogeneity in complier treatment effects along observables by including interactions between D_i and X_i; for an example, see Angrist et al. (2013).

[40] To obtain (37) from the more general (40), note that $D_i = Z_i$ for compliers and that

$$\mathbb{E}[W_i(1) - W_i(0)] = \text{LATE} \times \mathbb{P}[G_i = \text{CP}] = \mathbb{E}\left[\frac{Y_i D_i}{Q_i} - \frac{Y_i(1 - D_i)}{1 - Q_i}\Big| G_i = \text{CP}\right] \mathbb{P}[G_i = \text{CP}].$$

Applying (40) and simplifying some algebra then produces the difference of the two terms in (37).

4.2.2 Double robustness and machine learning

Doubly robust approaches combine imputation and propensity score weighting to produce an estimator that is consistent under correct specification of either the propensity score or the conditional means, but not necessarily both. See Kang and Schafer (2007) and Słoczyński and Wooldridge (2018) for overviews in the context of selection on observables. For estimating LATEs, Tan (2006) showed that a doubly-robust estimator of the first term in the numerator of (36) is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{\hat{q}(X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i}{\hat{q}(X_i)}-1\right)\hat{\mu}_1(X_i), \tag{41}$$

where $\hat{q}$ is (as before) an estimator of the instrument propensity score q and $\hat{\mu}(x)$ is an estimator of $\mu(x) \equiv \mathbb{E}[Y_i|Z_i = 1, X_i = x]$. Analogous estimators would replace the other terms in (36). Uysal (2011), Ogburn et al. (2015), Sun and Tan (2022), and Słoczyński et al. (2022) analyze various doubly-robust estimators based on (41).

To see what double robustness means, suppose that $\hat{q}$ and $\hat{\mu}$ consistently estimate functions $\tilde{q}$ and $\tilde{\mu}$, so that (41) is consistent for

$$\mathbb{E}\left[\frac{Z_i}{\tilde{q}(X_i)}Y_i\right] - \mathbb{E}\left[\left(\frac{Z_i}{\tilde{q}(X_i)}-1\right)\tilde{\mu}_1(X_i)\right]. \tag{42}$$

If $\tilde{q} = q$ is correctly specified, then the first term of (42) is equal to $\mathbb{E}[W_i(1)]$, as shown in (37), while iterating expectations shows that the second term is zero, regardless of whether $\tilde{\mu}_1 = \mu_1$. On the other hand, if $\tilde{\mu}_1 = \mu_1$, then the second term of (42) satisfies

$$\mathbb{E}\left[\left(\frac{Z_i}{\tilde{q}(X_i)}-1\right)\mu_1(X_i)\right] = \mathbb{E}\left[\frac{Z_i}{\tilde{q}(X_i)}Y_i\right] - \mathbb{E}[\mu_1(X_i)], \tag{43}$$

and so again (42) reduces to $\mathbb{E}[\mu_1(X_i)] = \mathbb{E}[W_i(1)]$, this time regardless of whether $\tilde{q} = q$. This is the double robustness property: the estimator (41) converges to (42), which is equal to $\mathbb{E}[W_i(1)]$ if either $\hat{q}$ or $\hat{\mu}$ is consistent for q or μ, but not necessarily both.

The double robustness property gives the researcher two chances at correct specification. Whether this translates into better finite sample performance when both specifications are wrong is debated, see e.g. Kang and Schafer (2007) and the commenting articles, or Słoczyński and Wooldridge (2018) for a review and unifying analysis. For practitioners, doubly robust estimators may be less attractive than propensity score weighting because they require making more modeling choices.

Machine learning (ML) methods can lessen this concern by allowing for data-driven

model selection. Belloni et al. (2017) and Chernozhukov et al. (2018) propose a method based on (42) and the other three analogous pieces, combined into a single moment condition. They estimate this moment condition using what they term double/debiased machine learning (DDML), which uses cross-fitting to fit flexible ML estimators to the functions $q(x)$, $\mathbb{E}[Y_i|Z_i = z, X_i = x]$, and $\mathbb{E}[D_i|Z_i = z, X_i = x]$ for $z = 0, 1$. With sufficiently flexible ML estimators of these five functions this procedure can be viewed as providing a nonparametric estimator of the unconditional LATE/ACR in (36). The doubly-robust formulation turns out to be important here as it makes the resulting moment condition Neyman orthogonal, an essential property for ensuring that nonparametric ML methods can be used despite their slower-than-parametric rates of convergence (see, e.g. Newey, 1994; Chernozhukov et al., 2018). Ahrens et al. (2024b) show how multiple ML estimators can be averaged together in DDML using "stacking" to reduce dependency on the approximation properties of any specific estimator.

4.2.3 Empirical illustration

We illustrate some of these methods with the well-known extract of the National Longitudinal Survey of Young Men used by Card (1993) in his analysis of the returns to schooling. The sample size of 3010. The outcome Y_i is log wage in 1976. The treatment D_i is years of education. The instrument Z_i is a binary indicator for living near a four-year college in 1966. Our aim is to estimate the unconditional ACR while accounting for the need to control for covariates X_i.

Table 4 compares five estimators across five different sets of covariates. The first two rows report the OLS and linear IV estimators that linearly control for sets of geographic and demographic covariates. Column (4) corresponds to the set of covariates used in Card's Table 5, column (3). Columns (2)–(3) only control for geographic covariates, while column (5) augments Card's specification with interactions between geographic and demographic covariates. The linear IV estimator is larger than the OLS estimator throughout all specifications, a finding that is common in the empirical literature, but conflicts with the classic constant effects reasoning about "ability bias" (Card, 2001).

Reverse engineering an interpretation for the linear IV estimator that allows for UHTE is challenging. The rich covariates condition is required for the corresponding estimand to have an interpretation as a non-negatively weighted average of causal effects (Section 3.8). A Ramsey (1969) RESET test rejects the null of rich covariates in all specifications, with p-values smaller than 10^{-4}. This provides strong statistical evidence that the linear IV estimator cannot be interpreted as weakly causal in this application.

Table 4: Methods of controlling for covariates in Card's (1993) data

	(1)	(2)	(3)	(4)	(5)
OLS	0.052	0.040	0.039	0.075	0.073
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)
Linear IV	0.188	0.091	0.092	0.132	0.133
	(0.026)	(0.056)	(0.056)	(0.054)	(0.055)
PLIV (DDML)	—	—	0.097	—	0.124
			(0.050)		(0.051)
ACR (weighting)	—	0.051	0.041	0.073	0.066
		(0.053)	(0.053)	(0.047)	(0.050)
ACR (DDML)	—	—	0.032	—	0.063
			(0.033)		(0.046)
Geographic controls		✓	✓	✓	✓
Geographic interactions			✓		✓
Demographic controls				✓	✓
Demographic interactions					✓

Notes: *Point estimates and heteroskedasticity-robust standard errors in parentheses. Geographic controls are indicators for region of residence in 1966, residence in an SMSA in 1966 and 1976, and residence in the South in 1966 and 1976. Demographic controls are an indicator for Black, experience, and experience squared. The DDML estimates use an ensemble of ten differently-tuned random forest, gradient boosting, and neural network algorithms, with weights chosen by non-negative least squares through the short-stacking procedure of Ahrens et al. (2024b). DDML estimates are given uninteracted lists of covariates but reported under the with-interactions columns because the methods potentially incorporate interactions on their own. Five folds are used for cross-fitting. The reported point estimate is the median across one hundred replications (different sample splits), with standard errors for the median computed according to Chernozhukov et al. (2018). The propensity score ACR estimates used normalized weights computed with a logit model.*

A modern reaction is to use data-driven machine learning techniques to select the functional form of covariates. The third row of Table 4 reports the DDML estimator for the partially linear IV (PLIV) specification discussed by Chernozhukov et al. (2018), which can be implemented using the `ddml` package for Stata or `R` (Ahrens et al., 2023, 2024a). Three functions are fit in this approach: $\mathbb{E}[Y_i|X_i = x]$, $\mathbb{E}[D_i|X_i = x]$, and $\mathbb{E}[Z_i|X_i = x]$. If the learners used to fit these functions are sufficiently expressive, then the PLIV DDML estimator will converge to the statistically-weighted average of covariate-specific ACRs given in (29), with $\beta_{\text{IV}}(x)$ being interpreted as the covariate-specific ACR via conditional versions of (22) or (23). While weakly causal, this object has a convoluted counterfactual interpretation because the weights depend on the joint distribution of X_i and Z_i.

The linear IV and PLIV estimates are close both when using only geographic con-

trols and in Card's specification that includes demographic controls. Comparing the linear IV and PLIV estimates underscores an important point about reverse engineering. Even if the linear IV and PLIV estimates were identical, this would not justify interpreting the linear IV estimate as weakly causal. There are, of course, an infinite number of ways to write the same single number as a weighted average, whether the weights are non-negative or not. The obvious consequence is that two estimates can be similar even if one estimates a weakly causal estimand and the other does not.

The fourth row of Table 4 reports propensity score weighting estimators (38) with the weights normalized as in (39). The only choice needed for this estimator is a model of $q(x) \equiv \mathbb{E}[Z_i|X_i = x]$. We use a logit model with the same covariate specifications as in the corresponding linear estimators. Słoczyński et al. (2024) show how to construct analytical standard errors for the estimator and provide a Stata package for implementation.[41] The fifth row of Table 4 reports DDML estimates of the unconditional ACR, again implemented using the `ddml` package (Ahrens et al., 2024a). The DDML estimates are computationally intensive to implement and depend on many choices and tuning parameters, which we made only modest attempts to explore.

The weighting and DDML estimates of the unconditional ACR are similar to one another, but substantially smaller than the linear and partially linear estimates, even while the standard errors for all estimates are comparable. The implication is that the difference between an unconditional ACR and a statistically-weighted average of ACRs is considerable in Card's application. Both sets of unconditional ACR estimates are comparable to the OLS estimates, and in some cases even smaller. This provides one answer to Card's (2001) puzzle of why linear IV estimates often exceed their OLS counterparts: the linear IV estimator is estimating an odd statistically-weighted object. Estimates of a more interpretable parameter like the unconditional ACR are not in fact larger than their OLS counterparts.

4.3 Marginal treatment effects

This section contains a development and selected review of marginal treatment effect (MTE) methods for binary treatments. Our focus is on an empirically tractable formulation of the MTE idea that leads to an implementation via linear regression. Surveys on MTE with different emphases are provided by Heckman and Vytlacil (2007b), Cornelissen et al. (2016), and Mogstad and Torgovitsky (2018).

[41] The package is called `kappalate`. We used our own `R` code together with bootstrapped standard errors. The analytical standard errors reported by `kappalate` are 10–20% smaller.

4.3.1 Definitions

MTE methods for binary treatments start with the same underlying assumptions used for identification of LATEs: full exogeneity and monotonicity. Instead of representing these selection assumptions with potential treatments, most authors prefer to use the latent variable notation (14), which we reproduce here, now augmented explicitly with covariates X_i:

$$D_i = \mathbb{1}[V_i \leq \nu(Z_i, X_i)]. \tag{44}$$

The idea is to view V_i as a random variable that captures the unobserved tendency to take treatment and then model the relationship between V_i and $(Y_i(0), Y_i(1))$.

The "marginal" descriptor comes from viewing an individual with $X_i = x$ and $V_i = \nu(z, x)$ as being on the margin between choosing $D_i = 0$ and $D_i = 1$ when faced with an instrument value $Z_i = z$. The average treatment effect for these marginal individuals is $\mathbb{E}[Y_i(1) - Y_i(0)|V_i = \nu(z, x), X_i = x]$. Björklund and Moffitt (1987) appear to have been the first to make use of this interpretation in a Gronau-Heckman normal selection model, but it did not attract much attention until being reintroduced in a nonparametric form by Heckman and Vytlacil (1999, 2005).

Working with (44) is cumbersome because both the function ν and the distribution of V_i are unknown. Yet full exogeneity implies that some features of these unknowns are identified by the propensity score:

$$\underbrace{p(z, x) \equiv \mathbb{P}[D_i = 1|Z_i = z, X_i = x]}_{\text{the (treatment) propensity score}} \overbrace{= \mathbb{P}[V_i \leq \nu(z, x)|X_i = x]}^{\text{by (44) and full exogeneity}} \equiv F_{V|X}(\nu(z, x)|x),$$

where $F_{V|X}$ is the distribution of V_i conditional on X_i. The model can be simplified while incorporating this identified relationship by reparameterizing (or "normalizing") the distribution of V_i. The simplest way to do this is to assume that V_i is continuously distributed and then apply $F_{V|X}$ to both sides of (44), defining a new random variable $U_i \equiv F_{V|X}(V_i|X_i)$:[42]

$$D_i = \mathbb{1}\big[\underbrace{F_{V|X}(V_i|X_i)}_{\equiv U_i} \leq \underbrace{F_{V|X}(\nu(Z_i, X_i)|X_i)}_{=p(Z_i, X_i)}\big] \equiv \mathbb{1}[U_i \leq p(Z_i, X_i)]. \tag{45}$$

The distribution of U_i conditional on $X_i = x$ is always uniform on $[0, 1]$ for any x, a

[42] Assuming that V_i is continuously distributed does not change the Vytlacil equivalence theorem; recall Figure 1.

textbook result known as the probability integral transform (e.g. Hansen, 2022a, pg. 35). Note that this implies that U_i is independent of X_i. However, it's important to remember that U_i is defined as a rank *conditional* on X_i; comparing U_i across different values of X_i can be misleading.

The normalized selection equation (45) is easier to work with because the distribution of U_i is known and the propensity score is identified. The normalization gives U_i an interpretation of the quantile of resistance to treatment. An individual with $U_i = .05$ is more prone to take treatment than 95% of the population or, equivalently, less resistant to taking treatment than only 5% of the population. The Vytlacil equivalence theorem reminds us that these statements should be interpreted as relative to hypothetical variation in the instrument Z_i; the selection model is a model of how treatment choice varies with Z_i, not X_i, and U_i is defined relative to Z_i. If Z_i is a cost shifter, then those with lower U_i require less cost reduction to take treatment than those with higher U_i. A different Z_i would mean a different model of selection and so a different U_i.[43] Even if two different binary instruments have the same propensity scores, their compliers need not reflect the same individuals, and so their U_i's also need not be comparable.

The MTE is defined as the ATE among subpopulations with the same propensity to take treatment:

$$\text{MTE}(u, x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | U_i = u, X_i = x].$$

The MTE is a useful definition because it uses the selection model to partition the population based on all unobservable and observable determinants of their treatment choice except for the instrument, which is the source of exogeneous variation.[44] UHTE is captured through variation in the u component of the MTE, while observed treatment effect heterogeneity is captured through variation in the x component. For modeling purposes it can be advantageous to work with the conditional mean of each treatment arm separately. Mogstad et al. (2018) call these conditional means the marginal

[43] This subtlety is perhaps one benefit of the potential treatment notation, which makes it harder to forget the interpretation of the latent variables being modeled. The latent variable notation makes it tempting to include several instruments without acknowledging the strong implications for choice behavior discussed in Section 3.5 (e.g. Carneiro et al., 2011), or to attempt to port U_i across different environments (e.g. Kowalski, 2023c).

[44] While usually considered in the context of instruments, the definition of the MTE only depends on (44), which allows for $\nu(Z_i, X_i) = \nu(X_i)$. Briggs et al. (2024) use this observation to consider an MTE analysis based on subjective expectations data rather than instruments.

Table 5: Marginal treatment response weights for common target parameters

		MTR weights	
Target parameter	**Expression**	$\boldsymbol{\omega(1\|u,z,x)}$	$\boldsymbol{\omega(0\|u,z,x)}$
Average treated outcome	$\mathbb{E}[Y_i(1)]$	1	0
Average untreated outcome	$\mathbb{E}[Y_i(0)]$	0	1
Average treatment effect (ATE)	$\mathbb{E}[Y_i(1) - Y_i(0)]$	1	-1
Conditional ATE	$\mathbb{E}[Y_i(1) - Y_i(0) \| X_i \in \mathcal{X}]$	$\frac{\mathbb{1}[x \in \mathcal{X}]}{\mathbb{P}[X_i \in \mathcal{X}]}$	$-\omega(1\|u,z,x)$
Average treatment on the treated (ATT)	$\mathbb{E}[Y_i(1) - Y_i(0) \| D_i = 1]$	$\frac{\mathbb{1}[u \leq p(z,x)]}{\mathbb{P}[D_i = 1]}$	$-\omega(1\|u,z,x)$
Average treatment on the untreated (ATU)	$\mathbb{E}[Y_i(1) - Y_i(0) \| D_i = 0]$	$\frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i = 0]}$	$-\omega(1\|u,z,x)$
Generalization of the LATE to $U_i \in [\underline{u}, \overline{u}]$	$\mathbb{E}[Y_i(1) - Y_i(0) \| U_i \in [\underline{u}, \overline{u}]]$	$\frac{\mathbb{1}[\underline{u} < u \leq \overline{u}]}{\overline{u} - \underline{u}}$	$-\omega(1\|u,z,x)$
Average selection on treatment effects	$\mathbb{E}[Y_i(1) - Y_i(0) \| D_i = 1] - \mathbb{E}[Y_i(1) - Y_i(0) \| D_i = 0]$	$\frac{\mathbb{1}[u \leq p(z,x)]}{\mathbb{P}[D_i = 1]} - \frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i = 0]}$	$-\omega(1\|u,z,x)$
Average selection bias	$\mathbb{E}[Y_i(0) \| D_i = 1] - \mathbb{E}[Y_i(0) \| D_i = 0]$	$\frac{\mathbb{1}[u \leq p(z,x)]}{\mathbb{P}[D_i = 1]} - \frac{\mathbb{1}[u > p(z,x)]}{\mathbb{P}[D_i = 0]}$	0
Policy relevant treatment effect (PRTE)	$\frac{\mathbb{E}[Y_i^\circ] - \mathbb{E}[Y_i]}{\mathbb{E}[D_i^\circ] - \mathbb{E}[D_i]}$	$\frac{\mathbb{P}[p^\circ(X_i, Z_i^\circ) \geq u] - \mathbb{P}[p(X_i, Z_i) \geq u]}{\mathbb{E}[p^\circ(X_i, Z_i^\circ)] - \mathbb{E}[p(X_i, Z_i)]}$	$-\omega(1\|u,x)$

Notes: *The weights show how to produce the specified target parameter through the formula*

$$\textit{target parameter} = \mathbb{E}\left[\int_0^1 \text{MTR}(1|u, X_i)\omega(1|u, Z_i, X_i)\,\mathrm{d}u - \int_0^1 \text{MTR}(0|u, X_i)\omega(0|u, Z_i, X_i)\,\mathrm{d}u\right].$$

treatment response (MTR):

$$\text{MTR}(d|u,x) \equiv \mathbb{E}[Y_i(d)|U_i = u, X_i = x]. \tag{46}$$

Any target parameter that reflects a mean or a mean contrast of potential outcomes can be written as a weighted average of the MTR function.[45] For example, the average treatment on the treated (ATT) can be written as

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] = \mathbb{E}\left[\int_0^1 (\text{MTR}(1|u,X_i) - \text{MTR}(0|u,X_i)) \underbrace{\frac{\mathbb{1}[u \leq p(Z_i,X_i)]}{\mathbb{P}[D_i = 1]}}_{\omega(1|u,Z_i,X_i)} \, \mathrm{d}u\right], \tag{47}$$

where the weights $\omega(d|u,z,x)$ are as indicated. For the ATT, the weights are symmetric in the sense that $\omega(0|u,z,x) = -\omega(1|u,z,x)$, but asymmetric weights can arise for target parameters that reflect only one treatment arm, or both arms but weighted differently. Table 5 reports weights for some of the more commonly considered target parameters. Appendix E.1 briefly discusses how to derive weighting expressions like these. Key to these weighting expressions is that the weights themselves are identified. The MTR function is the sole unknown.

4.3.2 Motivation

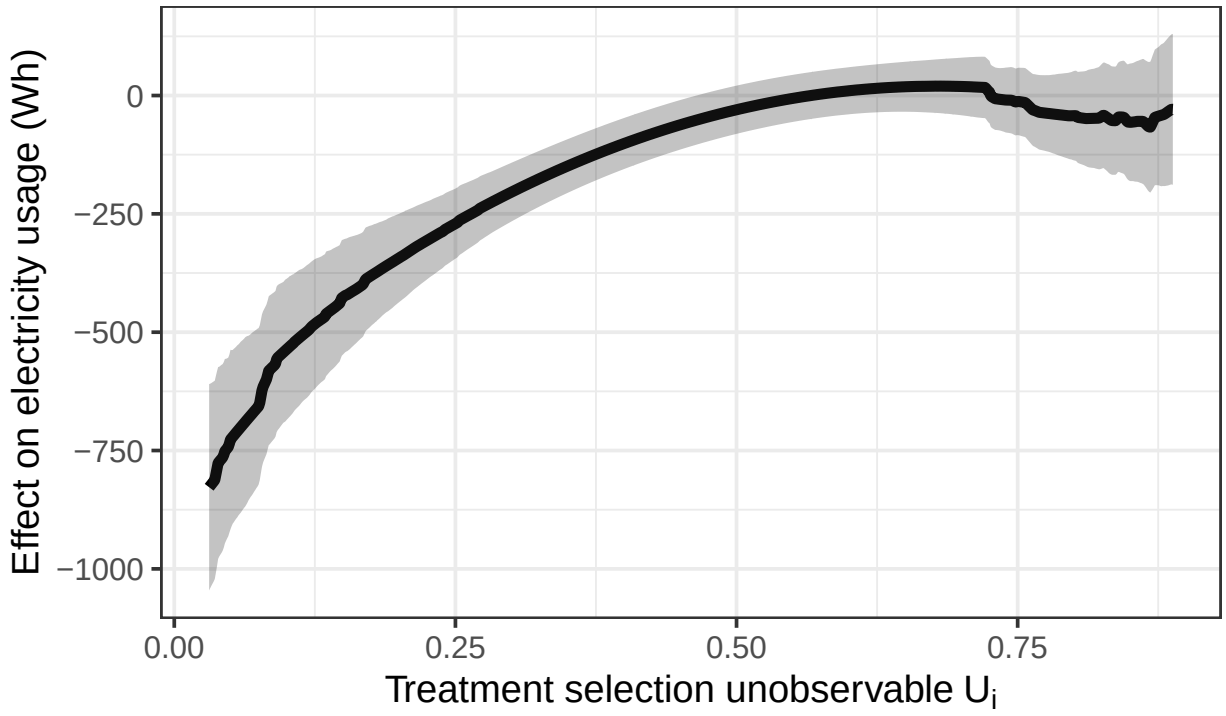
So far, these are just definitions. No additional assumptions beyond monotonicity and full exogeneity have been imposed. The purpose of the definitions is to provide a framework under which additional assumptions can be imposed and their identifying content exploited. The additional assumptions are used to construct estimates of a specific target parameter of interest, such as one of the ones listed in Table 5.

The Ito et al. (2023) study of dynamic electricity pricing provides a concrete example of why a researcher may want to do this. The treatment D_i in their setting indicates whether a household adopts dynamic pricing, meaning that instead of paying a fixed rate throughout the day, they pay considerably more during afternoon peak hours and somewhat less during off-peak hours. The instrument Z_i is a binary indicator of whether a household was randomly assigned a \$60 incentive to adopt dynamic pricing. The outcome Y_i is electricity usage.

This example falls into the baseline LATE setting of Section 3.2: a binary treatment

[45] The MTE idea can be extended beyond means as well. Carneiro and Lee (2009) and Martinez-Iriarte and Sun (2024) consider quantiles, while Acerenza et al. (2024) consider duration outcomes with censoring.

Figure 5: Marginal treatment effect estimates from Ito et al. (2023)



Notes: *Authors' reproduction of Figure 10, Panel A of Ito et al. (2023). We thank Koichiro Ito for providing the necessary data. The point estimate is the estimated MTE evaluated at the sample average of the covariates. The shaded region indicates 95% bootstrapped confidence intervals.*

and a binary instrument that is unconditionally randomly assigned and undoubtedly satisfies the monotonicity condition. The authors estimate the LATE, which provides an evaluation of the effect of an incentive policy matching their experimental policy of a \$60 incentive. But how does it compare to other potential policies?

Answering this question requires understanding which households would be drawn into dynamic pricing under different incentive policies and how dynamic pricing would change their usage. As Ito et al. (2023) discuss, willingness to participate and impact are likely linked: households that can more easily adjust their electricity usage may be both more willing to adopt dynamic pricing and more affected by it. This creates UHTE because "ease of adjustment" is unobserved. The MTE function captures the UHTE by how it changes with u and captures observed heterogeneity by how it changes with x.

The authors' MTE estimates are reproduced in Figure 5 which shows the point estimate of $\text{MTR}(u, x)$ with x evaluated at the sample mean, along with 95% confidence

intervals. The estimated MTE indicates dramatic UHTE with dynamic pricing having much larger impacts on electricity usage for lower values of u (more willing households) than larger values of u (less willing households). This implies that different incentive policies would also have substantially different impacts. Households with the highest impacts are drawn in by relatively small incentives. Larger incentives draw more household into dynamic pricing, but with less impact on usage. Ito et al. (2023) develop a welfare framework that incorporates these considerations, while also accounting for potential costs of adoption (Eisenhauer et al., 2015). They use the framework to estimate optimal incentive policies.

The Ito et al. (2023) MTE estimates in Figure 5 rely on additional assumptions about the MTR/MTE function beyond full exogeneity and monotonicity, assumptions which we discuss in detail ahead. Their estimates of the impacts of alternative incentive policies are necessarily less credible than their LATE estimate, because the former rely on strictly stronger assumptions. Yet the motivation of their analysis was precisely to estimate the impact of alternative policies and to characterize a potentially optimal one. Estimating the LATE alone does not speak to this motivation, but an MTE analysis can. The fact that their MTE analysis relies on stronger assumptions is part of the bargain.

4.3.3 A linear regression formulation

In this section, we describe a general but simple linear regression formulation of the MTE idea. Suppose that the MTR functions are linear in parameters, meaning that

$$\text{MTR}(d|u,x) = \sum_{k=1}^{d_\theta} \theta_k b_k(d|u,x), \tag{48}$$

where b_k are known "basis" functions specified by the researcher and θ_k are unknown parameters, collected into a d_θ dimensional vector θ. If the MTR satisfies (48), then the conditional means of the observed outcomes also turn out to be linear in θ. The relationship is

$$\begin{aligned} \mathbb{E}[Y_i|D_i,P_i,X_i] &= \sum_{k=1}^{d_\theta} \theta_k B_{ik} \\ \text{where} \quad B_{ik} &\equiv \left(\frac{1-D_i}{1-P_i}\right)\int_{P_i}^{1} b_k(0|u,X_i)\,\mathrm{d}u + \frac{D_i}{P_i}\int_{0}^{P_i} b_k(1|u,X_i)\,\mathrm{d}u, \end{aligned} \tag{49}$$

and where $P_i \equiv p(X_i, Z_i)$ is the propensity score evaluated at X_i and Z_i. The derivation of (49) is given in Appendix E.1, but it follows the same type of logic as the weight derivations in Table 5. The regressors B_{ik} are known functions of D_i, P_i, and X_i. The integrals in B_{ik} can be computed analytically for common examples of b_k like polynomials, while numerical integration can be used in other cases.

Equation (49) brings us into standard linear model territory.[46] Collect the regressors into a vector $B_i \equiv [B_{i1}, \ldots, B_{id_\theta}]'$. If the Gram matrix $\mathbb{E}[B_i B_i']$ is invertible, then θ is identified:

$$\theta = \mathbb{E}[B_i B_i']^{-1} \mathbb{E}[B_i Y_i]. \tag{50}$$

Because the B_i are functions of the propensity score, P_i, the invertibility of the Gram matrix is a statement about instrument relevance. Whether it holds will depend on the variation in the propensity score and the flexibility of the MTR specification. If θ is identified, then so too is any target parameter that can be written as an identified function of the MTR, such as the quantities in Table 5.

For example, suppose that there are no covariates and that the MTR is specified as linear in u with different parameters for each treatment arm:

$$\text{MTR}(d|u) = \underbrace{\theta_1(1-d) + \theta_2(1-d)u}_{\text{linear in } u \text{ (untreated)}} + \underbrace{\theta_3 d + \theta_4 du}_{\text{linear in } u \text{ (treated)}}, \tag{51}$$

so that, $b_1(d|u) = (1-d), b_2(d|u) = (1-d)u, b_3(d|u) = d$, and $b_4(d|u) = du$. Then (49) becomes

$$\mathbb{E}[Y_i|D_i, P_i] = \theta_1(1-D_i) + \theta_2(1-D_i)\frac{(1+P_i)}{2} + \theta_3 D_i + \theta_4 D_i \frac{P_i}{2}, \tag{52}$$

which specifies the observed conditional mean as a different linear function of the propensity score for each treatment arm. An alternative way to write (52) is as two separate regressions stratified by treatment arm:

$$\begin{aligned} \mathbb{E}[Y_i|D_i = 0, P_i] &= \theta_1 + \theta_2 \frac{(1+P_i)}{2}, \\ \text{and} \quad \mathbb{E}[Y_i|D_i = 1, P_i] &= \theta_3 + \theta_4 \frac{P_i}{2}. \end{aligned} \tag{53}$$

[46] The approach can be viewed as an example of a two-stage "control function" argument. An early example of it can be found in Heckman and Robb (1985, Section 3.4), although not stated in terms of the MTE. Wooldridge (2015) provides a history and exposition of the general idea of a control function. See also Vella (1998) for a discussion on the various control function approaches for one-sided sample selection models, many of which can be seen as progenitors of the approaches discussed ahead.

From (53), we see clearly what is required for identification of θ: the propensity score must have at least two points of support in each treatment arm. This is satisfied if Z_i is binary and $p(0) \neq p(1)$ with $p(0), p(1) \in (0,1)$.[47]

The general linear-in-parameters formulation (48) can flexibly accomodate more complex MTR specifications. Covariates can be included for each treatment arm, as can interactions between x and u. Nonlinear functions of u such as higher-order polynomials or splines can also be incorporated, just as in a standard linear model framework. The complexity in the u component is limited by the identification requirement that the Gram matrix be invertible, which is in turn determined by the amount of variation in the propensity score net of covariates. We discuss this further in the next section.

If θ is identified via (50) then it can then be consistently estimated as the coefficients in a linear regression of Y_i on B_i. Some components of B_i will generally depend on the propensity score, $P_i \equiv p(Z_i, X_i)$, as in (52), so to make this regression feasible B_i needs to be replaced by an estimate $\hat{B}_i$ based on an estimated propensity score $\hat{P}_i \equiv \hat{p}(Z_i, X_i)$. The same might also be true of the weights in the target parameter. We discuss this further in Section 4.3.6.

4.3.4 Identification

The linear MTR specification (48) was first studied by Brinch et al. (2012, 2017).[48] As those authors observed, the support of the instrument determines how much linearity can be relaxed without losing point identification. If the instrument has three points of support, and if these three points yield three distinct propensity score points per treatment arm, then an MTR that is quadratic in u is point identified. A cubic is identified with four points of support, and so on.

The traditional Gronau-Heckman normal selection model can also be interpreted as specifying (48). For simplicity, suppose there are no covariates. The normal selection model assumes that $(V_i(d), V_i)$ are bivariate normal with mean zero for $d = 0$ and $d = 1$, where $V_i(d) \equiv Y_i(d) - \mathbb{E}[Y_i(d)]$ and V_i has variance one, where V_i is the pre-normalization selection unobservable. Bivariate normals have linear conditional means,

[47] If $p(z) = 0$ or $p(z) = 1$, then there is one-sided non-compliance. For example, suppose that Z_i is treatment assignment and treatment cannot be obtained without being assigned to it, so that $p(0) = 0$. Then $\mathbb{P}[Z_i = 0 | D_i = 1] = 0$, so P_i only has one point of support ($P_i = p(1)$) in the treated arm, and consequently θ_3 and θ_4 are not separately identified. In this case only the compliers are treated, so there is no scope for interpolation or extrapolation among the treated.

[48] See Kowalski (2016, 2023c) for an alternative exposition of the same idea with an application to the impacts of health insurance. Closely-related linear control function assumptions have been used without the context of the MTE by Garen (1984) and Card (1999, 2001); see Wooldridge (2015).

so

$$\text{MTR}(d|u) = \mathbb{E}[Y_i(d)] + \mathbb{E}[V_i(d)|\underbrace{\overbrace{F_V(V_i)}^{U_i=u \Leftrightarrow V_i=F_V^{-1}(u)=\Phi^{-1}(u)} = u}_{U_i}] = \mathbb{E}[Y_i(d)] + \mathbb{C}[V_i(d), V_i]\Phi^{-1}(u), \quad (54)$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function. This can be written in the linear in parameters form (48) as

$$\begin{aligned}\text{MTR}(d|u) = \overbrace{\mathbb{E}[Y_i(0)]}^{\theta_1}\overbrace{(1-d)}^{b_1(d|u)} + \overbrace{\mathbb{C}[V_i(0), V_i]}^{\theta_2}\overbrace{(1-d)\Phi^{-1}(u)}^{b_2(d|u)} + \overbrace{\mathbb{E}[Y_i(1)]d}^{\theta_3 b_3(d|u)} \\ + \underbrace{\mathbb{C}[V_i(1) - V_i(0), V_i]}_{\theta_4}\underbrace{d\Phi^{-1}(u)}_{b_4(d|u)}. \end{aligned} \quad (55)$$

Appendix E.2 shows that computing the integrals in (49) with (55) produces

$$\mathbb{E}[Y_i|D_i, P_i] = \theta_1(1 - D_i) + \theta_2(1 - D_i)\lambda(-\Phi^{-1}(P_i)) + \theta_3 D_i - \theta_4 D_i \lambda(\Phi^{-1}(P_i)), \quad (56)$$

where $\lambda(v) \equiv \phi(v)/\Phi(v)$ is the inverse Mills' ratio. Equation (56) is recognizable as the Heckman selection correction applied to both treatment arms (e.g. Hansen, 2022b, pg. 883, equation 27.7).[49] The coefficients θ_3 and θ_1 are average treated and untreated outcomes for the entire population.

If the instrument is binary and there are no covariates, then both the linear and normal specifications lead to saturated regressions: D_i and P_i have four points of support and θ has four components. The population fitted values from these regressions will therefore exactly reproduce the conditional means $\mathbb{E}[Y_i|D_i = d, P_i = p]$ for $(d, p) \in \{0, 1\} \times \{p(0), p(1)\}$. As Brinch et al. (2017) observe, this means that for either specification (or any other saturated specification) the LATE implied by the MTR coefficients using the weighting in Table 5 must exactly match the usual LATE, even if the MTR specification is incorrect. See Appendix E.3 for a formal justification and Kline and Walters (2019) for an elaboration.[50] The equivalence is particular to saturated specifications and generally breaks in unsaturated specifications, which are more typical in practice, for example when including covariates.

While saturated MTR specifications produce the same LATE, they do not generally produce the same value for other parameters. The reason is that other parameters

[49] Most discussions of the Gronau-Heckman normal selection model would set the index function $\nu(z) = \nu_1 + \nu_2 z$ to be linear. This would then lead $\Phi^{-1}(P_i) = \nu_1 + \nu_2 Z_i$, which is the more familiar argument in the inverse Mills' ratio.

[50] The result for the normal selection model is also implicit in Baker and Lindeman (1994, Appendix I).

are not nonparametrically identified; they are identified given the MTR specification because the parametric assumptions allow for extrapolation (or interpolation). For example, the ATE weights all values of u equally, so the value of the ATE implied by a given MTR specification depends on how that specification extrapolates and interpolates from the observed propensity score support to other values of u. The extrapolation produced by the linear MTR specification is linear in the quantiles of latent resistance, u, whereas the extrapolation produced by the Gronau-Heckman normal selection model is highly nonlinear and diverges at its extremes because $\Phi^{-1}(u)$ diverges as u tends to zero or one. The relative transparency in how the linear MTR specification extrapolates is a good reason to prefer it over the traditional normal selection model.

Another advantage of the linear MTR specification is the transparency with which it can be made more flexible. The practice of including a quadratic or cubic term is familiar to practitioners from standard linear models. The linear-in-parameters specification also allows for the use of more local specifications, such as splines, which are even more transparent in how they extrapolate. A continuous instrument allows for a nonparametric specification, which for the linear-in-parameters formulation can be interpreted as linear sieve (Chen, 2007). This allows for nonparametric interpolation throughout the support of the propensity score, but does not solve the issue of extrapolation beyond the support.

An important distinction about propensity score variation arises when there are covariates. Suppose that $X_i \in \{0, 1\}$ is binary, for simplicity. A nonseparable linear MTR specification interacts the covariate with the linear terms:

$$\text{MTR}(d|u,x) = \underbrace{(1-d)\big(\overbrace{\theta_1 + \theta_2 u + \theta_3 x + \theta_4 ux}^{\text{linear in } u\text{, interacted with } x}\big)}_{\text{untreated arm}} + \underbrace{d\big(\overbrace{\theta_5 + \theta_6 u + \theta_7 x + \theta_8 ux}^{\text{linear in } u\text{, interacted with } x}\big)}_{\text{treated arm}}. \quad (57)$$

Compared to (51), this specification now allows the slope of the linear-in-u MTR to be different for $x = 0$ and $x = 1$. It implies that the conditional mean of Y_i given $D_i = d$ and $X_i = x$ is linear in P_i for each of the four combinations of d and x. Identification requires $p(1, x) \neq p(0, x)$ for each x, so that the instrument is relevant conditional on x.

A separable linear MTR specification sets θ_4 and θ_8 to be zero in (57). This requires the slope of the MTR to be the same for $x = 0$ and $x = 1$ while still allowing the level to be different. In this case, identification only requires $p(1, x) \neq p(0, x)$ for *either* $x = 0$ or $x = 1$. To see this, consider the implied conditional mean of the observed

outcome for the treated arm:

$$\mathbb{E}[Y_i|D_i = 1, P_i, X_i] = \theta_5 + \theta_6 \frac{P_i}{2} + \theta_7 X_i. \tag{58}$$

The coefficients are identified as long as P_i and X_i are not perfectly correlated (given $D_i = 1$). If we write the propensity score out in saturated form, so that

$$P_i = \pi_1 + \pi_2 Z_i + \pi_3 X_i + \pi_4 Z_i X_i \tag{59}$$

we can see that P_i and X_i will not be perfectly correlated if Z_i and X_i aren't and if either $\pi_2 \neq 0$ or $\pi_4 \neq 0$, which is the same as $p(0, x) \neq p(0, x)$ for either $x = 0$ or $x = 1$.

This discussion is related to the well-known critique about the Gronau-Heckman model being identified even without a relevant instrument (e.g. Goldberger, 1983; Puhani, 2000). The critique is really about the model of the propensity score: if both $\pi_2 = 0$ and $\pi_4 = 0$ in (59), so that Z_i is irrelevant, then P_i is perfectly correlated with X_i, so that θ_6 and θ_7 are not separately identified in (58). On the other hand, suppose that we start with an unsaturated probit model for the propensity score, so that

$$p(x, z) = \Phi(\pi_1 + \pi_2 Z_i + \pi_3 X_i), \tag{60}$$

where Φ is the standard normal cumulative distribution function. Now even if $\pi_2 = 0$, $P_i = \Phi(\pi_1 + \pi_3 X_i)$ is not perfectly correlated with X_i, because Φ is a nonlinear function. This shows that the critique about not needing an instrument is not related to selection modeling per se, but rather to the fact that statistical models commonly used for the propensity score of a binary treatment are nonlinear. Viewed from this perspective, this classic critique of selection modeling seems much less damning.

Separable specifications require less instrument variation. The flip side of this statement is that separable specifications can be more flexible in u than the variation in the instrument alone would suggest. Brinch et al. (2017) show that a separable MTR that is quadratic in u is generally identified with a binary instrument and binary covariate:

$$\text{MTR}(d|u, x) = \underbrace{(1 - d)\overbrace{\left(\theta_1 + \theta_2 u + \theta_3 x + \theta_4 u^2\right)}^{\text{quadratic in } u\text{, separable in } x}}_{\text{untreated arm}} + \underbrace{d\overbrace{\left(\theta_5 + \theta_6 u + \theta_7 x + \theta_8 u^2\right)}^{\text{quadratic in } u\text{, separable in } x}}_{\text{treated arm}}. \tag{61}$$

The intuition can be seen in the linear separable conditional mean (58) and the linear (saturated) propensity score specification (59). There are two excluded variables in

(59)—Z_i and Z_iX_i—but only one "endogenous" variable $P_i/2$ in (58).[51] So there's room to include an additional endogenous variable in (58) by adding the u^2 term in (61). Nonlinear propensity score specifications like (60) are typically used in practice, and these effectively create interactions between all values of the covariates and the instrument because $p(1,x) - p(0,x) = \Phi(\pi_1 + \pi_2 + \pi_3 x) - \Phi(\pi_1 + \pi_3 x)$ differs for all values of x.

The Ito et al. (2023) estimates in Figure 5 are based on a separable specification with a binary instrument. The authors provide evidence that takeup differs heavily by baseline household characteristics, in particular a measure of expected savings from switching to dynamic pricing given historical usage. The incentive has an impact throughout the distribution of household characteristics, leading to wide variation in the propensity score. This is what allows the authors to identify and estimate a flexible MTE curve across a wide range of u. The cost is the separability assumption, which requires the pattern of UHTE to not depend on observables.

Is the cost worth it? Should we be extrapolating at all? Certain target parameters, such as the ATE, depend on the MTE at extreme quantiles of u, and so require some extrapolation whenever these extreme quantiles are not represented in the propensity score, which is common. So whether extrapolation is necessary is a matter of the research question and how much variation there is in the data. The Ito et al. (2023) study provides a good example: the authors observed one incentive policy, but the purpose of their analysis was to compare different incentive policies and estimate an optimal one. There is no way to do this without interpolating and extrapolating.

The central role of extrapolation in IV methods with UHTE means that different specifications should be compared for sensitivity. This is already common practice in applications of MTE. Partial identification analysis provides a formal way to incorporate specification ambiguity by allowing for models that are too rich to generate a single value of the target parameter. Mogstad et al. (2018) develop a partial identification approach for MTE analysis. For applications of this approach see Mogstad et al. (2017), Rose and Shem-Tov (2021), Gulotty and Yu (2023), and Daljord et al. (2023).[52] A major benefit of considering partial identification is that it allows one to harness the

[51] The scare quotes are because $P_i/2$ is a function of Z_i and X_i, so not really endogenous. However, it arises in (58) because the MTR depends on u, and the dependence between U_i and $(Y_i(0), Y_i(1))$ is the source of endogenity.

[52] Han and Yang (2024) and Marx (2024) consider partial identification approaches that exploit the full independence between the instrument and potential outcomes, rather than the mean-independence considered here and in Mogstad et al. (2018). Kowalski (2023b,a) considers special cases that arise with a binary treatment and binary instrument, derives closed-form bounds that are easy to implement, and applies the bounds to study mammography and the overdiagnosis of breast cancer.

identifying content of nonparametric shape restrictions, such as monotonicity or concavity, which often have a clear economic interpretation. The major challenge with partial identification is computation, estimation, and inference; see Canay and Shaikh (2017) and Molinari (2020) for general discussions, and Shea and Torgovitsky (2023) for a discussion in the context of MTE methods.

4.3.5 Unstratified regressions and local instrumental variables

The implied conditional mean for Y_i considered in (49) is stratified in the sense that it conditions on the treatment indicator, D_i. Heckman and Vytlacil (2007b, Section 4.8) call this the selection or control function approach, while Brinch et al. (2017) call it the "separate" approach. An alternative is an unstratified regression where the conditioning on D_i is dropped and the coarser conditional mean of Y_i given only P_i and X_i is used instead. This produces the relationship

$$\mathbb{E}[Y_i|P_i,X_i] = \mathbb{E}[Y_i(0)|X_i] + \underbrace{\int_0^{P_i} \textsc{mte}(u,X_i)\,\mathrm{d}u}_{\mathbb{E}[D_i(Y_i(1)-Y_i(0))|P_i,X_i];\text{ see Appendix E.1}}. \tag{62}$$

Heckman and Vytlacil (2007b, Section 4.8) describe approaches based on (62) as "IV approaches" in contrast to control function approaches. We adopt the terminology stratified for (49) and unstratified for (62) because both use the variation in P_i produced by Z_i. The two approaches are more similar than they are different.

Heckman and Vytlacil (1999) observed that the derivative of the unstratified regression (62) identifies the MTE:

$$\underbrace{\textsc{liv}(u,x) \equiv \frac{\partial}{\partial u}\,\mathbb{E}[Y_i|P_i=u,X_i=x]}_{\text{local instrumental variable}} = \textsc{mte}(u,x). \tag{63}$$

They describe this derivative as the local instrumental variable (LIV) estimand due to its interpretation as a limiting case of the usual reduced-form-to-first-stage ratio in traditional linear IV estimands. If $u = p(z,x)$ is set to be an observed propensity score value, then

$$\textsc{liv}(p(z,x),x) \approx \frac{\mathbb{E}[Y_i|\overbrace{p(Z_i,X_i)}^{P_i} = p(z',x),X_i=x] - \mathbb{E}[Y_i|p(Z_i,X_i)=p(z,x),X_i=x]}{\underbrace{\mathbb{E}[D_i|Z_i=z',X_i=x]}_{p(z',x)} - \underbrace{\mathbb{E}[D_i|Z_i=z,X_i=x]}_{p(z,x)}},$$

where the approximation is for $p(z', x) \approx p(z, x)$. Because it is a derivative, $\text{LIV}(u, x)$ is only well-defined if there is continuous variation in P_i around u, conditional on $X_i = x$, which requires continuous variation in Z_i. Assuming such variation is available, (63) shows that the MTE is also identified at that point, a relationship that can be used to view the MTE at specific points of evaluation as limiting versions of the LATE. Carneiro et al. (2011) develop a semiparametric local polynomial estimator of the LIV using Robinson's (1988) approach for partially linear models; see Cornelissen et al. (2016) and Andresen (2018) for more details on implementation.

Continuous instrument variation is a luxury that is not available in many IV applications. Continuous covariate variation can be used as a substitute if the MTE is assumed to be separable so that $\text{MTE}(u, x) = m_U(u) + m_X(x)$ for two functions m_U and m_X. Estimation based on the LIV requires one or the other, so may not be applicable or attractive in many situations.

Alternatively, one can start with the unstratified regression (62) and use a linear-in-parameters specification:

$$\mathbb{E}[Y_i(0)|X_i = x] = x'\vartheta_0 \quad \text{and} \quad \text{MTE}(u, x) = \sum_{k=1}^{d_\vartheta} \vartheta_k b_k(u, x), \tag{64}$$

where b_k are known basis functions specified by the researcher and ϑ_k are unknown parameters. Notice that in contrast to (48), which parameterized the two treatment arms in the MTR separately, now we are parameterizing their difference—the MTE—as well as the baseline covariate relationship in the untreated state. Substituting these forms into (62) produces

$$\mathbb{E}[Y_i|P_i, X_i] = X_i'\vartheta_0 + \sum_{k=1}^{d_\vartheta} \vartheta_k B_k \quad \text{where} \quad B_k \equiv \int_0^{P_i} b_k(u, X_i)\,\mathrm{d}u. \tag{65}$$

Identification is again a matter of whether the Gram matrix for this linear regression is invertible, which requires having sufficient variation in the propensity score P_i, controlling for X_i.

Compared to the stratified regression (49), the unstratified regression (65) exploits less of the observed variation in the data, because it does not condition on D_i. An implication is that more instrument variation is needed for identification when considering comparable specifications. For example, suppose that there are no covariates and assume that the MTE is linear in u, so that

$$\mathbb{E}[Y_i(0)] = \vartheta_0 \quad \text{and} \quad \text{MTE}(u) = \vartheta_1 + \vartheta_2 u.$$

A linear MTE is implied by the linear MTR in (51). In principle, it is a weaker parameterization, although as a practical matter it is probably not substantively different.[53] However, when substituted into (65), the linear MTE produces a regression that is quadratic in P_i:

$$\mathbb{E}[Y_i|P_i, X_i] = \vartheta_0 + \vartheta_1 P_i + \vartheta_2 \frac{P_i^2}{2}. \tag{66}$$

This is in contrast to (52), which was linear in the propensity score, but stratified by treatment arm. Whereas a binary instrument was sufficient for invertibility with the stratified regression, three points of instrument support are needed for the comparable unstratified approach.

4.3.6 Estimation and inference

Estimating θ in the stratified regression requires first estimating the treatment propensity score to obtain estimates $\hat{P}_i \equiv \hat{p}(Z_i, X_i)$ of P_i. Typically one would use a logit or probit model for this purpose so that the $\hat{P}_i$ lie between 0 and 1, but a linear model could also be used. Replacing P_i with $\hat{P}_i$ in the definition of B_{ik} gives an estimate $\hat{B}_{ik}$ of B_{ik}, collected into a vector $\hat{B}_i$. Then θ can be estimated with ordinary least squares:

$$\hat{\theta} = \left(\sum_{i=1}^{n} \hat{B}_i \hat{B}_i' \right)^{-1} \left(\sum_{i=1}^{n} \hat{B}_i Y_i \right). \tag{67}$$

Estimating ϑ in the unstratified regression (62) proceeds the same way, except that B_{ik} and $\hat{B}_{ik}$ are defined differently and baseline covariates X_i are included additively, so that the regression is of Y_i on X_i and $\hat{B}_i$.

The parameters θ of the MTR (or ϑ or the MTE) are usually not of ultimate interest; instead we are interested in the target parameter that can be constructed from the MTR (or MTE). Suppose in particular that the target parameter takes the form

$$\tau = \sum_{d \in \{0,1\}} \mathbb{E}\left[\int_0^1 \text{MTR}(d|u, X_i) \omega^\star(d|u, Z_i, X_i) \mathrm{d}u \right], \tag{68}$$

where the weights $\omega^\star$ are assumed to be identified, but may need to be estimated.

[53] For example suppose that we specify the MTR as $\text{MTR}(d|u,x) = \theta_1(1-d) + \theta_2(1-d)u + \theta_3 d + \theta_4 du + \theta_5 u^2$, where the quadratic term does not depend on d. The implied MTE is then $(\theta_3 - \theta_1) + (\theta_4 - \theta_2)u$, which is linear in u.

Substituting (48) into this expression gives

$$\tau = \sum_{k=1}^{K} \theta_k b_k^\star \quad \text{where} \quad b_k^\star \equiv \mathbb{E}\left[\sum_{d\in\{0,1\}} \int_0^1 b_k(d|u, X_i)\omega^\star(d|u, Z_i, X_i)\,\mathrm{d}u \right]. \tag{69}$$

Each of the $b_k^\star$ are identified but need to be estimated if $\omega^\star$ needs to be estimated, or if b_k or $\omega^\star$ depend on X_i or Z_i. A natural estimator of $\hat{b}_k^\star$ is

$$\hat{b}_k^\star \equiv \frac{1}{n}\sum_{i=1}^{n} \sum_{d\in\{0,1\}} \int_0^1 b_k(d|u, X_i)\hat{\omega}^\star(d|u, Z_i, X_i)\,\mathrm{d}u, \tag{70}$$

where $\hat{\omega}^\star$ is an estimator of $\omega^\star$. For example, if the target parameter is the ATT, then

$$\hat{\omega}^\star(d|u, Z_i, X_i) = \frac{\mathbb{1}[u \leq \hat{P}_i]}{n^{-1}\sum_{j=1}^{n} \hat{P}_j}. \tag{71}$$

The final estimate of the target parameter τ is then

$$\hat{\tau} \equiv \sum_{k=1}^{d_\theta} \hat{\theta}_k \hat{b}_k^\star. \tag{72}$$

Estimating a target parameter with an unstratified regression uses exactly the same procedure, just that τ must only depend on the MTE, and not the two arms of the MTR separately.

Computing $\hat{\theta}$ and $\hat{\tau}$ requires calculating the integrals in the definitions of B_i and $b^\star$. The `mtefe` package for Stata (Andresen, 2018) and the `ivmte` package for `R` (Shea and Torgovitsky, 2023) both contain functionality that automates this task.[54] Given the integration, computation is simply a matter of one logistic or other binary response regression to estimate the propensity score p and one linear regression to estimate the MTR parameters θ. The `ivmte` package for `R` also contains functionality for implementing the partial identification approach developed by Mogstad et al. (2018), but estimation and inference is more complicated.

The formal asymptotic theory for $\hat{\theta}$ and $\hat{\tau}$ does not appear to have been worked out yet.[55] There is no reason to expect that these estimators would not be asymptotically normal, but their asymptotic variances will be complicated by the fact that $\hat{B}_{ik}$ and $\hat{b}_k^\star$ are estimated in a first step. Simulation evidence by Andresen (2018) bears out the

[54] The `mtefe` package improves on the earlier `margte` package by Brave and Walstrum (2014).

[55] Carneiro and Lee (2009) and Sasaki and Ura (2023) have derived formal results for semiparametric approaches that are not linear-in-parameters.

normality and suggests that naive variance calculations that ignore first step estimation error may still be roughly accurate. Bootstrapping the entire procedure will account for this first step estimation error and is easy to do. This already appears to be standard practice among empirical practitioners. All aspects of the estimation procedure are smooth, so there is no reason to expect that the bootstrap would not be consistent (Fang and Santos, 2019), at least assuming that the instruments are sufficiently strong.[56]

These statements apply equally to the stratified and unstratified regressions. Which should be used? One consideration is the amount of instrument variation available, as less is required for the stratified regressions. Another consideration is the target parameter: a stratified approach estimates the MTE, and so will not provide enough information to compute target parameters that depend on the MTR components themselves. This comes up in the Ito et al. (2023) study, where the target parameter considered depends asymmetrically on electricity usage under dynamic and static pricing, which create different marginal surpluses. Assuming that enough variation is available to consider both and that the target parameter depends only on the MTE, the only remaining consideration is presumably statistical precision. Andresen (2018) provides some simulation evidence that suggests the stratified approach tends to lead to more precise estimates. On balance then, stratified MTR approaches seem preferable, although more research on the statistical differences would be useful.

4.3.7 Applications and uses of marginal treatment effects

Applications of MTE methods are widespread and have been proliferating rapidly in the past fifteen years. Table 6 provides a list of some empirical applications of MTE methods. All of these applications use the MTE to investigate patterns of UHTE. These patterns add depth and nuance to the empirical analysis and can sometimes speak to questions about mechanisms.

Another complementary use of MTE methods is for estimating the impacts of explicit policy counterfactuals. The Ito et al. (2023) study is one example. Another example is given by Cornelissen et al. (2018), who study the impact of publicly provided childcare on childrens' outcomes, and then use their MTE estimates to simulate the aggregate impacts of expanding publicly provided childcare. A third example is Mogstad et al. (2017), who use MTE estimates to compare the cost effectiveness of different subsidies for encouraging the use of mosquito nets.

[56] As with linear IV estimators, weak instruments could make the asymptotic approximation a poor description of the finite sample behavior, and would also render the bootstrap inconsistent (e.g. Andrews et al., 2019).

Table 6: Empirical applications of marginal treatment effects

Labor and human capital	Moffitt (2008), Carneiro et al. (2011), Kaufmann (2014), Carneiro et al. (2016), Joensen and Nielsen (2016), Nybom (2017), Dal Bó et al. (2021), De Groote and Declercq (2021), Gathmann et al. (2021), Heinesen and Stenholt Lange (2022), Westphal et al. (2022), Dutz et al. (2022), Humlum et al. (2023),
Development	Mogstad et al. (2017), Bandiera et al. (2020), Berry et al. (2020), Manda et al. (2020), Li et al. (2021a), Mellon Bedi et al. (2021), Sarr et al. (2021)
Health	Basu et al. (2007), Johar and Maruyama (2014), Basu et al. (2014), Alessie et al. (2020), Depalo (2020), Gong et al. (2020), Zeng et al. (2020), Kowalski (2023c,a), Wilding et al. (2023), Gupta et al. (2024)
Family and childhood development	Doyle Jr. (2007), Brinch et al. (2017), Cornelissen et al. (2018), Felfe and Lalive (2018), Priebe (2020), Hojman and Lopez Boo (2022), Liu et al. (2022)
Crime	Doyle Jr. (2008), Arnold et al. (2018), Bhuller et al. (2020), Arnold et al. (2022), Baron and Gross (2022), Arbour (2022), Arteaga (2023), Agan et al. (2023), Possebom (2023), Gonçalves and Mello (2023)
Public programs	Maestas et al. (2013), French and Song (2014), Moffitt (2019), Moffitt and Zahn (2019), Aizawa et al. (2023)
Energy	Wang et al. (2020), Li et al. (2021b), Ito et al. (2023)
Other	Galasso et al. (2013) (innovation), Kasahara et al. (2016) (international trade), Daljord et al. (2023) (marketing), Coury et al. (2022) (history), Heldring et al. (2022) (history)

Notes: *A list of papers that have applied MTE methods to empirical problems. We limit the list to papers that fit the binary treatment setting in Section 4.3.*

In each case, the authors measure the policy counterfactual using what Heckman and Vytlacil (2001a, 2005) called policy-relevant treatment effects (PRTEs). A PRTE is defined by a hypothetical modification of the instrument and/or propensity score from p and Z_i to p° and Z_i°. This results in a different hypothetical treatment selection,

$$D_i^\circ \equiv \mathbb{1}[U_i \leq p^\circ(Z_i^\circ, X_i)],$$

and so also different realized outcomes,

$$Y_i^\circ \equiv (1 - D_i^\circ)Y_i(0) + D_i^\circ Y_i(1).$$

Let $p^{\circ\circ}$ and $Z_i^{\circ\circ}$ denote some other policy that leads to treatment choices $D_i^{\circ\circ}$ and

realized outcomes $Y_i^{\circ\circ}$. Then the PRTE for these two policies is defined as

$$\text{PRTE} \equiv \frac{\mathbb{E}[Y_i^{\circ}] - \mathbb{E}[Y_i^{\circ\circ}]}{\mathbb{E}[D_i^{\circ}] - \mathbb{E}[D_i^{\circ\circ}]},$$

which gives the average change in outcomes per net change in treatment participation. An alternative definition omits the denominator and just measures the average change in outcomes (Heckman and Vytlacil, 2001a; Carneiro et al., 2010).[57] The contrasting policy is often taken to be the baseline status quo, $p^{\circ\circ} = p$, $Z_i^{\circ\circ} = Z_i$, so that $D_i^{\circ\circ} = D_i$ and $Y_i^{\circ\circ} = Y_i$ are the observed treatment and outcomes.

Constructing p° and Z_i° may require some extrapolation or speculation. Cornelissen et al. (2018) consider a policy that takes $p^{\circ}(Z_i, X_i) = \min\{1.5p(Z_i, X_i), 1\}$ and so increases the likelihood of attendance for every child by one and a half times. As the authors point out, it is not clear what type of concrete intervention would achieve this new level of attendance. Another counterfactual policy intervention the authors consider is a direct increase in their instrument, the number of available childcare seats per capita, from Z_i to $Z_i^{\circ} = Z_i + .4$. Here the intervention is more clear, but the impact that this has on attendance depends on $p(Z_i^{\circ}, X_i)$, which involves some extrapolation beyond the observed support of Z_i. Two types of extrapolation (or interpolation) are required for evaluating a PRTE of this sort: the effect of changing the instrument on treatment takeup, and the effect of treatment on outcomes for those induced to change their treatment status under the new policy.

The definition of the PRTE is premised on the fundamentals of the environment remaining stable under the new policy, an assumption that Heckman and Vytlacil (2005) describe as policy invariance. The need for policy invariance is not specific to MTE methods; it's a necessity for any sort of counterfactual policy analysis. In the context of the MTE, policy invariance means that the distribution of $(Y_i(0), Y_i(1), U_i)$ remains the same under different policies. There are certainly good reasons to be skeptical of such an assumption. For example, if the childcare expansion entertained by Cornelissen et al. (2018) is achieved by adding poorer quality childcare facilities, then the treatment effect $Y_i(1) - Y_i(0)$ could change, leading to a failure of policy invariance. Again, this is not a drawback of MTE methods or even IV strategies more generally, but rather an inherent limitation of evaluating a policy with a model that doesn't model all possible effects of the policy.

[57] Carneiro et al. (2010) also define a limiting version of the PRTE for small policy changes, which they call the marginal PRTE or MPRTE. The advantage of the MPRTE is that it doesn't require any extrapolation, and so in principle can be estimated nonparametrically.

4.4 Binary treatments when monotonicity is violated

The monotonicity condition plays a central role when estimating both LATEs and MTEs. Yet as we saw in Section 3.4, it can be unattractive in some settings, such as in judge designs. Multiple instruments are also difficult to square with the traditional monotonicity condition (Section 3.5). What can be done in these cases?

The simplest solution is to redefine the instrument in a way that makes the monotonicity condition more plausible. For example, instead of using all judges individually, Dahl et al. (2014) consider estimates based on a binary instrument that defines whether the judge is one of the most strict or one of the least strict, with moderate judges being omitted from the analysis. Monotonicity violations that might occur among a variety of similar judges are perhaps less likely to occur when comparing extreme judges, making the monotonicity condition with the binary instrument more plausible. This point was recently recycled by Sigstad (2024b,a). Binarizing the instrument also makes it easier to assess the impact that violations of monotonicity would have through a sensitivity analysis like the one in Section 3.4.

The same idea can also be applied to multiple instruments. The problem in that case was the difficulty in comparing treatment choice behavior under pairs of instrument values that were not ordered in a natural way, such as $(Z_{i1}, Z_{i2}) = (0, 1)$ and $(Z_{i1}, Z_{i2}) = (1, 0)$. One solution is to remove these instrument values and only consider $(Z_{i1}, Z_{i2}) = (0, 0)$ and $(1, 1)$, which can be naturally ordered if both Z_{i1} and Z_{i2} are incentives to take treatment. Versions of this idea have been used by Frölich (2007), Goff (2024), and van 't Hoff et al. (2024). An alternative is to use only one component of the instrument at a time, conditioning on the rest of the components as covariates. Monotonicity in each instrument separately allows for the estimation of separate LATEs and separate MTE curves, one for each instrument; see Mogstad et al. (2021) for an empirical illustration. Mogstad et al. (2024) show how MTE curves for different instruments can be aggregated in a partial identification framework.

Recording the instrument or conditioning on subcomponents are simple solutions, but they reduce the amount of effective exogenous variation. A more ambitious approach is to design a new selection model that allows for deviations from monotonicity. The major obstacle is identification. With a binary treatment, monotonicity enables identification of the shares of each choice group. Removing monotonicity requires adding a new assumption or allowing for the possibility of partial identification.

Gautier and Hoderlein (2015) and Gautier (2021) consider random coefficient versions of the threshold-crossing model (44) and show that point identification can be obtained under extreme large support assumptions on the available of instrument vari-

ation. Ura and Zhang (2024) and Han and Kaido (2024) provide partial identification approaches that are applicable to models that do not satisfy monotonicity, such as a random coefficients model, but the approaches come with some of the familiar computational and statistical challenges of partial identification. Dutz et al. (2022) develop a simple non-monotonic model of survey response and apply it under conditions that lead to either point or partial identification.

Arnold et al. (2022, Section 4) point out that MTE-style regressions of outcomes on propensity scores can still be estimated even if monotonicity does not hold. Assuming that these regressions are correctly parameterized, they can still be extrapolated to estimate unconditional potential outcome means. Their approach effectively replaces low-level behavioral assumptions about monotonicity with higher-level statistical assumptions about the relationship between outcomes and the propensity score. Arnold et al. (2022, Section 5) develop a parametric selection model that does not impose monotonicity and show how to estimate the model, but do not establish identification.

Measurement error in the treatment provides another source of monotonicity violations. Even if monotonicity is satisfied for the correctly measured (but latent) binary treatment variable, misclassification will mean that it is violated for the observed, mismeasured treatment variable. Ura (2018), Calvi et al. (2022), and Tommasi and Zhang (2024) consider identification of the LATE in the presence of this type of measurement error, while Possebom (2023), Acerenza et al. (2023), and Acerenza (2024) consider identification of the MTE. Partial identification emerges in all of these analyses, with the exception of Calvi et al. (2022).

4.5 Ordered treatments

The linear MTE approach for binary treatments extends to ordered treatments with one important caveat: the natural generalization of the threshold-crossing model (44) is no longer equivalent to the monotonicity condition.

4.5.1 Threshold-crossing with multiple treatments

Suppose as in Section 3.6 that the treatment variable takes values $d_0, d_1, \ldots, d_J$ arranged in increasing order. Instead of (44), assume that

$$\mathbb{1}[D_i \geq d_j] = \mathbb{1}[V_i \leq \nu(d_j|Z_i, X_i)] \quad \text{for each } j. \tag{73}$$

The function ν now depends on d_j, but the unobservable V_i is the same for all d_j, an important distinction that we will return to ahead. Because D_i is no smaller than d_0

we can set $\nu(d_0|Z_i, X_i) = +\infty$. It's also convenient to add an artificial value $d_{J+1} > d_J$ with $\nu(d_{J+1}|Z_i, X_i) = -\infty$ to reflect that D_i must always be strictly smaller than d_{J+1}.

Normalizing (73) as in the binary case simplifies it to

$$\mathbb{1}[D_i \geq d_j] = \mathbb{1}\big[\underbrace{F_{V|X}(V_i|X_i)}_{\equiv U_i} \leq \underbrace{F_{V|X}(\nu(d_j|Z_i, X_i)|X_i)}_{=p(d_j|Z_i,X_i)}\big] \equiv \mathbb{1}[U_i \leq p(d_j|Z_i, X_i)], \quad (74)$$

where $p(d_j|z, x) \equiv \mathbb{P}[D_i \geq d_j|Z_i = z, X_i = x]$ is a generalization of the propensity score (the "greater than" propensity score or the conditional survival function when viewed as a function of d_j). As in the binary case, $p(d_j|z, x)$ is identified. In the multivalued case, it is decreasing in d_j, with $p(d_0|z, x) = 1$ and $p(d_{J+1}|z, x) = 0$.

Writing (74) in terms of individual levels makes it look a bit more familiar:

$$D_i = d_0 + \sum_{j=1}^{J}(d_j - d_0)\underbrace{\mathbb{1}\left[p(d_{j+1}|Z_i, X_i) < U_i \leq p(d_j|Z_i, X_i)\right]}_{\mathbb{1}[D_i < d_{j+1} \text{ and } D_i \geq d_j] = \mathbb{1}[D_i = d_j]}. \quad (75)$$

This is an ordered response model (e.g. Greene and Hensher, 2009; Wooldridge, 2010, Chapter 16). The binary threshold-crossing model (5) is recovered by setting $J = 1$, $d_0 = 0$, and $d_1 = 1$, so that $p(d_0|z, x) = 1$, $p(d_1|z, x) = \mathbb{P}[D_i = 1|z, x]$ is the usual binary propensity score, and $p(d_2|z, x) = 0$.

In the multivalued case, the ordered response model is no longer equivalent to the monotonicity condition. This was shown by Vytlacil (2006) in a follow-up to Vytlacil (2002). The reason can be seen with three values ($J = 2$), a binary instrument, no covariates, and with monotonicity in the direction $D_i(1) \geq D_i(0)$. Monotonicity allows for the choice groups $G_i = (d_0, d_2)$ and $G_i = (d_1, d_1)$ to both exist. In terms of (75), the first group would consist of those individuals with U_i strictly larger than $p(d_1|0)$, giving $D_i(0) = d_0$, and weakly smaller than $p(d_2|1)$, giving $D_i(1) = d_2$. So for this first group to exist, it must be that $p(d_1|0) < p(d_2|1)$. On the other hand, the second group consists of those values of U_i that lie in the intervals $(p(d_2|z), p(d_1|z)]$ for both $z = 0$ and $z = 1$. But if the first group exists, these two intervals must be disjoint:

$$\underbrace{p(d_2|0) < p(d_1|0)}_{U_i \text{ here if } G_i(0) = d_1} \overbrace{<}^{\text{if } G_i = (d_0, d_2) \text{ exists}} \underbrace{p(d_2|1) < p(d_1|1)}_{U_i \text{ here if } G_i(1) = d_1}. \quad (76)$$

Intuitively, the ordered response model restricts how much treatment can respond to the instrument: in this case it can either respond a lot ($G_i = (d_0, d_2)$) or not at all ($G_i = (d_1, d_1)$), but not both.

This finding makes sense in the context of reverse engineering linear IV with ordered treatments (Section 3.6). The weights in the Angrist and Imbens (1995) ACR were identified, but only because they combined ("double counted") multiple choice groups. A counting exercise reveals that the shares of each of the choice groups cannot be point identified. With $J = 2$ and a binary instrument, there are six choice groups consistent with monotonicity: G_i can be $(d_0, d_0), (d_0, d_1), (d_0, d_2), (d_1, d_1), (d_1, d_2)$, or (d_2, d_2). Yet there are only four independent choice probabilities: $\mathbb{P}[D_i = d|Z_i = z]$ for $d = d_0, d_1$, and $z = 0, 1$, with the probability for $D_i = d_2$ being implied by the sum-to-one constraint. Five independent choice group probabilities cannot be uniquely pinned down by four independent choice probabilities. The ordered threshold model (75) effectively rules out an additional choice group a priori, restoring point identification of the group shares.

Is this additional restriction attractive? Vytlacil (2006) shows that it's not inherent to latent variable notation or even to a threshold-crossing structure. Vytlacil (2006) extends his equivalence result from the binary case to a more flexible class of ordered response models with thresholds that vary according to additional latent variables, which he shows is again equivalent to the monotonicity condition stated with potential choices notation. The number of latent variables in these models makes it clear that they are not point identified without additional distributional structure or extreme assumptions on the available instrument variation (Cunha et al., 2007). We are left with a familiar trilemma: (i) use a selection model that admits point identification but makes potentially restrictive behavioral assumptions; (ii) relax the behavioral assumptions but impose additional parametric structure; or (iii) allow for partial identification. Most empirical applications of forward engineered with ordered treatments have taken the first option and used (75) as the selection model.[58]

4.5.2 A linear regression formulation

Using (75) as the selection model makes it possible to directly adapt the linear regression formulation of MTE from the binary case.[59] The marginal treatment response

[58] An example of an application of the third option is Goldin et al. (2021), who maintain the usual monotonicity condition. Their partial identification argument is further developed in Vohra and Goldin (2024). Kamat et al. (2024) develop and apply a partial identification approach under a different model of ordered choice that they describe as "latent monotonicity." Their approach uses a linear-in-parameters framework similar to the one discussed ahead, but without the benefit of a point identified selection model. Arteaga (2023) uses the same latent monotonicity model with additional assumptions that effectively return the problem back to a binary treatment case, to which she then applies MTE methods.

[59] Heckman et al. (2006) and Heckman and Vytlacil (2007b) provide the earliest discussions phrased in terms of local instrumental variable estimands.

(MTR) is defined as before, except now there are more values of d for its first argument. We again assume that the MTR function has the linear-in-parameters form (48). For notation, let $P_i(d_j) \equiv p(d_j|Z_i, X_i)$ be the d_j-specific greater-than propensity score and collect these scores into $P_i \equiv (P_i(d_1), \ldots, P_i(d_J))$. Then (49) can be generalized to

$$\mathbb{E}[Y_i|D_i, P_i, X_i] = \sum_{k=1}^{d_\theta} \theta_k B_{ik}$$
$$\text{where} \quad B_{ik} \equiv \sum_{j=0}^{J} \mathbb{1}[D_i = d_j] \left(\frac{1}{P_i(d_j) - P_i(d_{j+1})} \int_{P_i(d_{j+1})}^{P_i(d_j)} b_k(d_j|u, X_i)\, du \right). \quad (77)$$

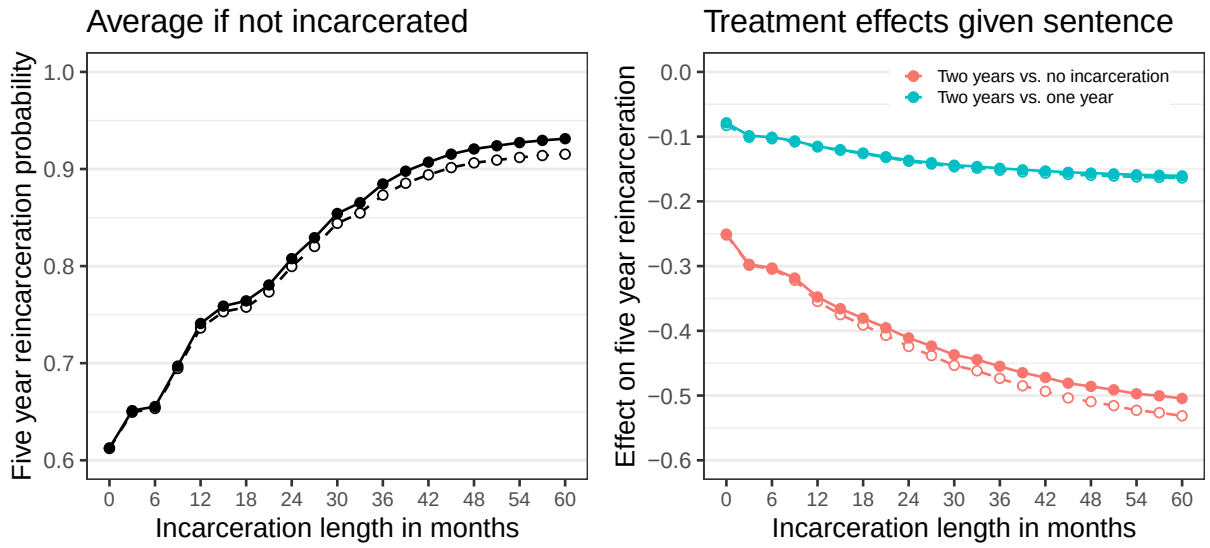
First step estimates of $P_i(d_j)$ can be constructed by estimating an ordered response model and then used to construct estimates $\hat{B}_{ik}$ of B_{ik}. At that point the story becomes the same as in the binary treatment case: a linear regression of Y_i onto $\hat{B}_{ik}$ to estimate the θ_k's, which can then be used to estimate a variety of target parameters.

The primary difference with the binary treatment case is that it might also be attractive to parameterize the d dimension of the MTR function when D_i has a cardinal interpretation. For example, the linear specification (51) can be extended so that each treatment value d has its own linear-in-u function, leading to $d_\theta = 2(J+1)$ parameters. These can be point identified if the binary instrument satisfies $p(d_j|0) \neq p(d_j|1)$ for $j = 1, \ldots, J$. A more parsimonious specification could be to interact the levels of d and u, so that there are only four basis functions $(1, d, u, du)$ with four parameters to estimate. Many other types of parameterizations are possible. Point identification is a matter of whether the resulting Gram matrix formed from the B_{ik} variables is invertible, which requires the $P_i(d_j)$ scores to vary sufficiently with Z_i and/or X_i.

Rose and Shem-Tov (2021) apply a linear-in-parameters ordered treatment MTR analysis in their study of the effect of incarceration on recidivism.[60] The authors' treatment is incaracteration measured in months, which can be expected to have an impact on recidivism that is both nonlinear over length and heterogeneous across individuals. The authors use discontinuities in sentencing guidelines as instruments to produce linear IV estimates, but they correctly recognize that the interpretation of these estimates is opaque with both a multivalued treatment and a multivalued instrument (not to mention covariates). Even if weakly causal, the linear IV estimates smear UHTE and nonlinearities into a single hard-to-interpret number. The authors provide suggestive evidence of both UHTE and nonlinearities by showing that their linear IV estimates change substantially when using different instruments or when including

[60] Other recent applications are Cornelissen et al. (2018) and Rivera (2023).

Figure 6: Marginal treatment effect estimates from Rose and Shem-Tov (2021)



Notes: *Authors' reproduction of Figure 6 of Rose and Shem-Tov (2021). We thank Evan Rose for providing the necessary data. The solid lines are upper bound estimates and the dotted lines are lower bound estimates. The left-hand panel reports estimates of* $\mathbb{E}[Y_i(0)|D_i = d]$*, where* Y_i *is an indicator for any reincarceration within five years and* D_i *is incarceration length in months. The right-hand panel reports estimates of* $\mathbb{E}[Y_i(24) - Y_i(0)|D_i = d]$ *in red and* $\mathbb{E}[Y_i(24) - Y_i(12)|D_i = d]$ *in blue.*

nonlinear treatment terms.

Figure 6 reproduces a central empirical finding from Rose and Shem-Tov (2021, Figure 6). The outcome is an indicator for any reincarceration within five years of sentencing. The authors assume that the MTR function is a fifth degree polynomial in u that is additively separable between u and x, but make no assumptions about how the MTR varies across d. Because the MTR is so flexible in u, the authors proceed as if it (and any target parameters generated from it) are potentially only partially identified and extend the partial identification framework of Mogstad et al. (2018) to the ordered treatment case. While some of their bound estimates are wide, many are quite narrow, including the ones in Figure 6, which are essentially point estimates.

The left panel of Figure 6 plots estimates of the counterfactual probability of recidivism (being reincarcerated) if counterfactually not incarcerated, conditional on sentence length. The strong upward trend indicates strong selection patterns: judges assign longer sentences to individuals more likely to offend if not incarcerated. The right panel shows that the effects of a hypothetical two year sentence are large and negative. The effect of the two year sentence increases dramatically with an individual's actual sentence length, signaling important treatment effect heterogeneity. On top of

the heterogeneity, the right panel shows evidence of nonlinearity, with the effect being driven by the first year of the hypothetical two year sentence.

Rose and Shem-Tov (2021) point out that accounting for the type of heterogeneity and nonlinearity visible in Figure 6 is important when considering sentencing policy. They use their estimates to consider the impacts of a budget-neutral change in sentencing that increases the rate of incarceration while reducing the length of longer sentences. As expected from Figure 6, they find that this type of reallocation can produce large reductions in the rate and duration of reincarceration.

Linear IV estimates are not up to the task of evaluating this type of nuanced policy counterfactual. If viewed as correctly-specified, a linear IV estimate for a specification that is linear in treatment mechanically produces no aggregate effects from reallocating sentence lengths. A reverse engineering interpretation allows the coefficient on the single linear treatment variable to be interpreted as a non-negatively weighted average across different treatment intensities (Section 3.6), but knowing the components in this weighted average is what's needed for evaluating a reallocation of sentence lengths. Adding in a nonlinear treatment term to the linear IV specification puts one in a reverse engineering no-mans land of no known results other than to assume constant effects. But the assumption of constant effects is both implausible a priori and strongly at odds with the right-hand panel of Figure 6. Ignoring treatment effect heterogeneity would overstate the benefit from reducing sentence lengths by not accounting for the higher effect of incarceration on more severe offenders.

4.5.3 Continuous treatments

In some cases it might be reasonable to think of the treatment as being ordered and continuous. The model analogous to (75) for a continuous treatment can be written as

$$D_i = \nu(Z_i, X_i, V_i), \tag{78}$$

where ν is an unknown function that is assumed to be strictly decreasing in V_i, which is still a single scalar unobservable, and still assumed to satisfy full exogeneity together with the potential outcomes. It is again possible to normalize V_i by replacing it with $U_i \equiv F_{V|X}(V_i|X_i)$, a transformation that can be absorbed into the definition of ν (Matzkin, 2003). Upon doing so, we see that ν is identified by the survival function

$$p(d|Z_i, X_i) \equiv \mathbb{P}[D_i \geq d|Z_i, X_i] = \mathbb{P}[U_i \leq \nu^{-1}(Z_i, X_i, d)|Z_i, X_i] = \nu^{-1}(Z_i, X_i, d), \tag{79}$$

where ν^{-1} is the inverse of ν in its V_i argument. The selection model is then

$$D_i = p^{-1}(U_i|Z_i, X_i) = \nu(Z_i, X_i, U_i), \tag{80}$$

which is just a function of a scalar uniform random variable and point identified objects, as before. The MTR $m(d|u, x)$ is defined the same way as before, except now the first argument can take a continuum of values.

There is an econometrics literature that analyzes models of continuous treatments together with a selection equation like (80).[61] Imbens and Newey (2009) show that if no assumptions are placed on the MTR, then extreme instrument variation is necessary for point identification of $\mathbb{E}[Y_i(d)] = \mathbb{E}[\text{MTR}(d|U_i, X_i)]$. Torgovitsky (2015, 2017) shows that the strong assumption of rank invariance across different potential outcomes (e.g. Heckman et al., 1997) enables nonparametric point identification of average and quantile potential outcomes even with only a binary instrument. Viewed in terms of the MTR function, these results are fully nonparametric in d, u, and x. They represent polar cases that are likely to be unattractive for most applications.

Imposing some parametric assumptions seems reasonable. Masten and Torgovitsky (2016) show that if

$$\text{MTR}(d|u, x) = \rho_0(u) + \rho_1(u)d + \rho_2(u)'x, \tag{81}$$

then the functions $\rho_0(u), \rho_1(u)$, and $\rho_2(u)$ can be identified for all u with only a binary instrument, implying identification of the entire MTR function. Masten and Torgovitsky (2014) analyze a kernel-based linear regression estimator, which can be implemented with the Stata command `ivcrc` (Benson et al., 2022); see Gollin and Udry (2021) and Carrillo et al. (2023) for empirical applications. Florens et al. (2008) consider a rth degree polynomial specification that omits covariates but includes an additional unknown function of d:

$$\text{MTR}(d|u) = \rho_0(u) + \rho_1(u)d + \cdots + \rho_r(u)d^r + \bar{\rho}(d). \tag{82}$$

Their identification results require continuous instrument variation and they do not consider estimation.

Chernozhukov et al. (2020) and Newey and Stouli (2021) extend the analysis of

[61] The selection equation is traditionally assumed to be strictly increasing in U_i rather than strictly decreasing, but this is not material. What matters is the invertibility created by strict monotonicity. The literature also typically uses nonseparable models for Y_i rather than potential outcomes; we have translated the notation and findings to the concept of an MTR so as to stick with the potential outcomes notation.

both Florens et al. (2008) and Masten and Torgovitsky (2016) to allow for more general specifications of the MTR, as well as quantile counterparts. These include the linear-in-parameters form of the MTR (48), which should be particularly attractive and flexible for applications. If we let $P_i \equiv p(D_i|Z_i, X_i)$, which is equal to U_i by (80), then

$$\mathbb{E}[Y_i|D_i, P_i, X_i] = \sum_{k=1}^{d_\theta} \theta_k B_{ik} \quad \text{where} \quad B_{ik} \equiv b_k(D_i|P_i, X_i). \tag{83}$$

Implementation proceeds as before: estimate p, now using distribution or quantile regression (e.g. Chernozhukov et al., 2013) to construct $\hat{B}_{ik}$, then regress Y_i onto $\hat{B}_{ik}$ to estimate the θ_k's. Relative to (81) and (82), the linear-in-parameters specification allows for parameterizations in both the u and d dimensions. This can be used to lessen the demands on instrument variation.

4.5.4 Selection models that do not allow for heterogeneity

The selection equations (44), (75), and (78) for the binary, ordered discrete, and continuous cases all have a single unobservable, but one that enters the equation non-additively. This is important. As we noted, for multivalued cases the selection models are not equivalent to the monotonicity condition, which generally requires additional latent variables (Vytlacil, 2006). However, they do still allow for unobserved heterogeneity in how the instrument affects treatment choice.

In contrast, Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008) consider linear and additive selection models like

$$D_i = \nu_0 + \nu_1 Z_i + V_i, \tag{84}$$

where V_i is assumed to be mean-independent of Z_i. The authors show that under this condition and the linear-in-treatment specification (81), the linear IV estimand is equal to $\mathbb{E}[\rho_1(U_i)]$, and so the average partial effect $\mathbb{E}[\text{MTR}(d|U_i) - \text{MTR}(d'|U_i)]$ is identified for any pairs d and d'.

This result comes at a high cost. While (84) may look like the usual statistical first stage regression (7), the regression only ensures that V_i and Z_i are orthogonal, which is weaker than mean-independence. Imposing mean-independence requires thinking of (84) as a selection model that describes choices under counterfactual manipulation of the instrument. It is a particularly restrictive selection model because it implies that the effect of Z_i on D_i is ν_1—constant for all individuals—so that there is no unobserved heterogeneity in the effect of the instrument on the treatment. This seems

like an unacceptably asymmetric assumption given the motivating goal of allowing for UHTE in treatment effects.

4.6 Unordered treatments

The major obstacle in extended these ideas to unordered treatments is identification of the selection equation. Suppose that D_i takes one of J unordered values $d_0, d_1, \ldots, d_J$, and that selection follows the discrete choice model

$$D_i = \underset{d_j \in \{d_0, d_1, \ldots, d_J\}}{\arg\max} \; \nu(d_j | Z_i, X_i) + U_{ij}, \tag{85}$$

where ν is again an unknown function and U_{ij} are unobservables, with $\nu(d_0|z,x) = 0$ and $U_{i0} = 0$ as a normalization. Relative to the binary and ordered cases, there are now multiple unobservables, one for each choice, which we collect as the vector $U_i \equiv (U_{i1}, \ldots, U_{iJ})$. The familiar econometric interpretation of (85) views the arguments of the argmax as indirect utilities for choosing option d_j, with the observed choice D_i being the one with the highest utility. These indirect utilities can differ with the instrument, the observed covariates, and the unobservables.

The definition of the MTR extends immediately with the change that now u is a vector, not a scalar. We can still consider a linear-in-parameters specification like (48). As one example, Kline and Walters (2016) consider a case with three choices ($J = 2$) and assume that

$$\text{MTR}(d_j | \underbrace{u_1, u_2}_{u}, x) = \rho_0(d_j)'x + \rho_1(d_j)u_1 + \rho_2(d_j)u_2, \tag{86}$$

where ρ_j are unknown coefficients that are different for each treatment state d_j. This MTR specification can be written in the linear basis form with nine components by including indicators for each treatment state. Where things become difficult is the following step: what does (86) imply about the conditional mean of the observed outcome? It can still be related to the MTR via

$$\mathbb{E}[Y_i(d)|D_i = d_j, Z_i = z, X_i] = \mathbb{E}\left[\text{MTR}(d_j|U_i, x) \,\middle|\, \underbrace{U_i \in \mathcal{U}^\star(d_j|z, X_i)}_{\text{set of } U_i \text{ for which } d_j \text{ is optimal}}, X_i\right], \tag{87}$$

where $\mathcal{U}^\star(d_j|z,x)$ is the subset of U_i for which d_j is the maximizer of (85). But evaluating this expression further requires knowing something about the distribution of U_i, even if the MTR function is assumed to have a linear-in-parameters form.

This problem is the same issue that arose for binary treatments without mono-

tonicity (Section 4.4). Counterfactual choice probabilities—let alone the distribution of U_i—are not point identified in a traditional discrete choice model like (85) without parametric assumptions or extreme instrument variation (Tebaldi et al., 2023). This is in contrast to the models for ordered choice considered in the previous section, all of which admitted nonparametric point identification of choice probabilities. The uniform normalization that produced U_i came from folding the unknown distribution of the original latent variable V_i into the definition of the MTR, permitting focus on a single unknown object. These nonparametric simplifications are not available for the unordered case, at least not with a model like (85).

One path is to embrace the need for parameterization and leverage insights from the well-developed literature on discrete choice (e.g. Train, 2009). A pioneering early example is Dubin and McFadden (1984), who used a multinomial logit model for the selection equation and also made an assumption like (86) to derive a linear-in-parameters expression for the conditional mean.[62] See Abdulkadiroğlu et al. (2020) for a recent application of their approach. Kline and Walters (2016) replaced the logit with a multinomial probit that allows for correlation between U_{ij} and U_{ik}. They derived the resulting expression for the observed conditional outcome mean when the MTR is given by (86). The result looks like a multivariate generalization of the inverse Mills' ratio expression (56). Kline and Walters (2016) show that if there is only a binary instrument that affects the utility of one option, then additional variation in choice probabilities due to covariates is needed for identification if the MTR is separable between x and u, as in (86). Hull (2020) uses a similar approach to estimate hospital quality.

These types of parametric distributional assumptions may of course be unattractive, or at least not sufficiently easy to flexibly modify. Dahl (2002) replaced the explicit parametric distributional assumptions with higher-level index-sufficiency assumptions. Heckman and Pinto (2018), Lee and Salanié (2023), and Navjeevan et al. (2023) consider selection models for unordered treatments that are more restrictive than (85), which can lead to point identification of average treatment effects for certain choice groups either nonparametrically or with a little bit of added parametric structure (Pinto, 2022).[63] Lee and Salanié (2023) and Kamat (2024) consider partial identification. All of these approaches are relatively tailored and may require a fair amount of case-specific work to be applied.

[62] This is again an example of a control function approach (Heckman and Robb, 1985; Vella, 1998; Wooldridge, 2015), in contrast to fully parameterizing the entire model and basing estimation off of the likelihood function, which requires stronger assumptions. See Geweke et al. (2003) for an example of this type of approach for unordered treatments.

[63] See Navjeevan et al. (2023) and Xie (2024) for estimation methods that incorporate covariates gracefully.

As usual, extreme amounts of instrument variation can solve these difficulties. Heckman et al. (2008) showed that instruments that drive choice probabilities to one effectively reduce the problem back to the binary treatment setting. Lee and Salanié (2018) showed that natural generalizations of the local instrumental variable argument extend to the unordered treatment setting for a variety of choice models *if* these models are point identified, which typically requires extreme instrument variation (or parametric assumptions). Also as usual, instruments with extreme variation don't exist in practice.

Mountjoy (2022) shows how continuous but local (not extreme) instrument variation can be used nonparametrically with unordered treatments. Mountjoy's argument is based on having choice-specific instruments in (85).[64] In his application with $J = 2$ representing the choice of two-year or four-year college, this means that $\nu(d_1|Z_{i1}, Z_{i2}, X_i) = \nu(d_1|Z_{i1}, X_i)$ and $\nu(d_2|Z_{i1}, Z_{i2}, X_i) = \nu(d_2|Z_{i2}, X_i)$. Mountjoy (2022) argues that this can be satisfied with separate two- and four-year distance instruments. He then shows that marginal shifts in these instruments can be used to separately identify marginal treatment effects of d_2 relative to d_0 and of d_4 relative to d_0 for those on the margin of indifference between these choices. Identifying the ν function or the distribution of U_i can by bypassed because of the assumption that each instrument affects only one choice. Humphries et al. (2023b) show that if this assumption is dropped, then one can still apply Mountjoy's argument by estimating ν and then using the estimated $\nu(d_j|Z_i, X_i)$ as themselves choice-specific instruments.

Unordered choice problems can also arise out of dynamic or multistage decision problems. Versions of the above ideas have been applied to these settings as well. For examples, see Heckman et al. (2016, 2018), Walters (2018), and Humphries et al. (2023a).

4.7 No selection model

All of the reverse and forward engineering approaches for allowing UHTE discussed so far have maintained a selection model. This is certainly not innocuous. The behavioral assumptions imposed by a selection model, such as the monotonicity condition, can sometimes be unattractive. The full exogeneity condition that is invariably imposed with a selection model places strong requirements on the instrument, as we noted in Section 2.5. In this section, we consider an approach due to Manski (1990) that allows for UHTE but does not impose a selection model.

[64] Mountjoy shows that the full structure of (85) is not necessary, although it nicely captures the key elements. Heckman and Pinto (2018) and Loeser (2023) provide equivalence results relating discrete choice models of treatment selection to restrictions on their implied potential choices.

4.7.1 Manski-Robins and IV intersection bounds

Suppose that outcome exogeneity is satisfied, so that $\mathbb{E}[Y_i(d)|Z_i = z] = \mathbb{E}[Y_i(d)]$ for all d and z. Assume that the treatment is binary, for simplicity.[65] Then

$$\underbrace{\mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(1)|Z_i = z]}_{\text{outcome exogeneity}} = \underbrace{\mathbb{E}[Y_i|D_i = 1, Z_i = z]p(z)}_{\text{identified}} + \underbrace{\mathbb{E}[Y_i(1)|D_i = 0, Z_i = z]}_{\text{not directly identified}} \underbrace{(1 - p(z))}_{\text{identified}}.$$

The only term on the right-hand side that is not identified by the data is the counterfactual treated mean for those in the untreated state. Assume that it lies in $[y_{\text{lb}}, y_{\text{ub}}]$. This assumption could be based either on the logical support of Y_i, in which case it's not restrictive, or it could be based on substantive restrictions about a reasonable range for the conditional mean of Y_i. Substituting these bounds for the unidentified counterfactual gives upper and lower bounds on $\mathbb{E}[Y_i(1)]$ that depend on z. The tightest bounds are then found by taking the largest lower bound and the smallest upper bound across z:

$$\begin{aligned}\mathbb{E}[Y_i(1)] \in \Big[&\max_z \mathbb{E}[Y_i|D_i = 1, Z_i = z]p(z) + y_{\text{lb}}(1 - p(z)), \\ &\min_z \mathbb{E}[Y_i|D_i = 1, Z_i = z]p(z) + y_{\text{ub}}(1 - p(z))\Big]. \end{aligned} \tag{88}$$

A symmetric set of bounds can be derived for the untreated mean, $\mathbb{E}[Y_i(0)]$. Bounds for the ATE are then formed by taking the difference of the bounds for the potential outcome means.

This argument was first considered without an instrument (Z_i deterministic) by Manski (1989) and Robins (1989), a result often described as the "worstcase" bounds, although that phrase is a bit misleading, so we will describe these as the Manski-Robins bounds.[66] Manski (1990, 1994) observed that an instrument allows for the construction of the IV bounds in (88), which are often described as "intersection bounds" due to their max-min structure. The intersection bounds only depend on the outcome exogeneity assumption. They do not require full exogeneity nor any behavioral assumptions about selection, such as the monotonicity condition.

4.7.2 Empirical illustration

Table 7 reports estimates of Manski-Robins and IV intersection bounds using data from Gelbach (2002), who estimated the impact of public school availability on maternal

[65] The following argument applies equally well for non-binary treatments, but the bounds will tend to be wider.

[66] All bounds are achieved at the "worst case."

Table 7: Manski-Robins and instrumental variable bounds in Gelbach (2002)

		Logical ($y_{\text{lb}} = 0, y_{\text{ub}} = 1$)		Substantive ($y_{\text{lb}} = 0.4, y_{\text{ub}} = 0.8$)	
	p(z)	**LB**	**UB**	**LB**	**UB**
			Manski bounds		
$\mathbb{E}[Y_i(0)]$	.632	.275	.908	.528	.781
$\mathbb{E}[Y_i(1)]$	.632	.425	.793	.572	.719
$\mathbb{E}[Y_i(1) - Y_i(0)]$		-.482	.518	-.209	.191
$\mathbb{E}[Y_i(0)\|Z_i = 75\text{:Q1}]$	.313	.496	.809	.621	.746
$\mathbb{E}[Y_i(0)\|Z_i = 74\text{:Q4}]$	.553	.344	.897	.565	.787
$\mathbb{E}[Y_i(0)\|Z_i = 74\text{:Q3}]$	.793	.159	.952	.476	.793
$\mathbb{E}[Y_i(0)\|Z_i = 74\text{:Q2}]$	.834	.127	.961	.461	.795
$\mathbb{E}[Y_i(1)\|Z_i = 75\text{:Q1}]$	.313	.192	.879	.467	.742
$\mathbb{E}[Y_i(1)\|Z_i = 74\text{:Q4}]$	.553	.355	.802	.534	.712
$\mathbb{E}[Y_i(1)\|Z_i = 74\text{:Q3}]$	.793	.545	.752	.628	.711
$\mathbb{E}[Y_i(1)\|Z_i = 74\text{:Q2}]$	.834	.582	.748	.648	.715
			Instrumental variable bounds		
$\mathbb{E}[Y_i(0)]$		.496	.809	.621	.746
$\mathbb{E}[Y_i(1)]$		.582	.748	.648	.711
$\mathbb{E}[Y_i(1) - Y_i(0)]$		-.227	.252	-.098	.090

Notes: *Sample analog estimates of the components of* (88)*, the symmetric expressions for* $\mathbb{E}[Y_i(0)]$*, and bounds on the ATE formed from the difference. The data is the sample of 10,932 single mothers whose youngest child was five years old in 1980. The outcome variable is an indicator for employment in 1979. The values of* Z_i *indicate the birth quarter of this child. heteroskedasticity-robust standard errors (not shown) for* $\mathbb{E}[Y_i|D_i = d, Z_i = z]$ *are smaller than* .02 *for all values of* d *and* z *and for* $p(z)$ *are smaller than* .01 *for all values of* z*.*

labor supply. The sample is restricted to mothers whose youngest child was five years old in 1980. The treatment D_i is an indicator for whether the mother's five-year-old was enrolled in public school. The outcome Y_i is an indicator for the whether the mother was employed in the previous year. Gelbach instruments for D_i with indicators Z_i for the quarter when the five-year-old was born, an instrument that is relevant because of age-at-entry rules for public kindergartens.

The top portion of Table 7 reports two sets of estimated Manski-Robins bounds on the untreated means. In the first set of bounds, we take these to be the logically possible values for a binary variable of $y_{\text{lb}} = 0$ and $y_{\text{ub}} = 1$. In the second set of bounds, we make the substantive (potentially incorrect) assumption that counterfactual employment probabilities lie between $y_{\text{lb}} = .4$ and $y_{\text{lb}} = .8$, compared to estimated conditional employment probabilities $\mathbb{E}[Y_i|D_i = d, Z_i = z]$ that range between .49 and

.71 over different values of d and z.

The first three rows of Table 7 do not condition on the instrument. The unconditional treatment propensity is .632, so bounds on the treated mean are narrower than on the untreated mean. The implied bounds on the ATE are quite wide, even when placing substantive prior bounds on counterfactual employment probabilities. These bounds do not make use of the assumption that the instrument satisfies outcome exogeneity.

The subsequent rows report Manski-Robins bounds that condition on the instrument. These are just conditional-on-Z_i versions of the first three rows. The propensity score varies with the conditioning value of the instrument, leading to variation in the width of the bounds. The bounds for the untreated conditional mean are narrowest for the youngest children (75:Q1), who are least likely to be in public kindergarten. For the treated conditional mean, they are narrowest for the oldest children (74:Q2). Outcome exogeneity allows bounds for the youngest and oldest children to be combined through the intersection bounds in (88). The result is shown in the final three rows together with the implied bounds on the ATE.

While narrower than the unconditional Manski-Robins bounds, both the logical and substantive IV intersection bounds on the ATE are still quite wide. As a point of reference, an uncontrolled OLS estimate is $-.076$ (standard error .009), which increases to $-.013$ (SE: .008) when adding state fixed effects and demographic controls (Gelbach, 2002, Table 7, columns (1)–(2)). An uncontrolled linear IV estimate is .036 (SE: .022), which increases to .040 (SE: .020) in Gelbach's preferred linear IV specification that includes controls (Gelbach, 2002, Table 7, column (3)).[67] The logical IV bounds on the ATE by contrast range from $-.227$ to .252. The substantive bounds of $-.098$ to .090 are considerably narrower, but are still consistent with negative or positive effects larger than any of Gelbach's estimates. This is despite the relatively wide variation in the propensity score between younger and older children from .313 to .834. That the IV bounds are inconclusive about the ATE even in a setting with this type of propensity score variation is probably why they are not often used in practice.[68]

[67] The bounds in Table 7 do not control for covariates. Controlling for covariates in a parsimonious way is a challenge for this type of bounding analysis.

[68] Applications of IV bounds can be found in Pepper (2000), Siddique (2013), and Shurtz et al. (2022). A more commonly-applied bounding approach uses an assumption that Manski and Pepper (2000) call monotone instrumental variables (MIV). The MIV assumption is that $\mathbb{E}[Y_i(d)|Z_i = z]$ is increasing or decreasing in z, instead of constant (both increasing and decreasing) as under outcome exogeneity. Bounds based on MIV can be compelling tools for causal inference, but the name MIV is perhaps a misnomer, as the whole point of the assumption is that the variables used as Z_i no longer need to be excluded and exogenous; monotone covariates would be an equally appropriate description. For applications of MIV, see Kreider and Pepper (2007), Blundell et al. (2007), Kreider et al. (2012), and De Haan and Leuven (2020).

Inconclusive does not mean useless. The bounds shown in Table 7 are sharp, meaning the best possible given the assumption of outcome exogeneity. So, far from being useless, they indicate the central role played by making additional assumptions. A linear IV estimate based on a binarized version of the instrument that groups the two earliest and two latest quarters is .034 (SE: .024). If we assume away UHTE, then this estimate is a consistent estimate of the ATE. If we allow for UHTE, then the most we can conclude about the ATE is that it is contained within the bounds given in Table 7. In this way, the IV bounds quantify the empirical importance of assuming constant treatment effects.

4.7.3 The role of a selection model

By the same reasoning, the IV bounds also quantify the empirical importance of using a selection model. Under full exogeneity and monotonicity, the binarized linear IV estimate of .034 is a consistent estimate of the LATE.[69] Comparing the LATE estimate to the ATE bounds measures how different the treatment effect for compliers could be from the treatment effect for the overall population. In this case, a positive LATE is consistent with an ATE that is positive and considerably larger, or negative and equally large. Similarly, using the selection model together with a linear MTE extrapolation produces an estimated ATE of .025 (SE .022). Selecting this number from the ATE bounds depends on the validity of the selection model and the extrapolation of the MTE.

The absence of a selection model helps clarify why one can be so useful. Modeling how counterfactual objects relate to observable ones is the fundamental challenge in all causal inference. For IV with a binary treatment, the counterfactual object is the conditional mean $\mathbb{E}[Y_i(1)|D_i = 0, Z_i = z]$. Bounding this object to $[y_{\text{lb}}, y_{\text{ub}}]$ is one simple model. For developing a more involved model, one effectively has two components to work with: the potential treatment arm, $Y_i(1)$ vs. $Y_i(0)$, or the conditioning event, $D_i = 0, Z_i = z$ vs. $D_i = d, Z_i = z$ for different values of d and z. One option with the former is to assume that $Y_i(1) \geq Y_i(0)$, an assumption Manski (1997) described as monotone treatment response. Apart from that, there is little scope for additional

[69] Perhaps surprisingly, the ATE bounds in Table 7 are still sharp even after imposing monotonicity and full exogeneity (Balke and Pearl, 1993, 1997; Heckman and Vytlacil, 2001b; Bai et al., 2024). To be more precise, these selection model assumptions have testable implications, as explored by the literature cited in Section 2.7, but if the testable implications are satisfied, then the sharp bounds on the ATE are the same as they are without the selection model. This result applies more generally to the entire marginal distributions of potential outcomes (Kitagawa, 2009, 2021), but not to the joint distribution (Kamat, 2021). It breaks down when additional assumptions are added; see Flores et al. (2018) for a survey and Shaikh and Vytlacil (2011), Bhattacharya et al. (2012), Huber et al. (2015), and Machado et al. (2019) for some specific results.

assumptions that still allow for UHTE, at least with a binary treatment.[70] That leaves modeling the conditioning event of the treatment and instrument value, which means modeling the treatment selection process.

Selections models become a necessity when the goal is to evaluate a policy change that affects treatment choice. The Cornelissen et al. (2018) evaluation of publicly provided childcare and the Ito et al. (2023) study of dynamic pricing provide two clear examples (see Section 4.3.7). There is little hope for conclusive inference about these types of policy counterfactuals without imposing assumptions on how the instrument affects treatment selection.

4.8 Summary of forward engineering

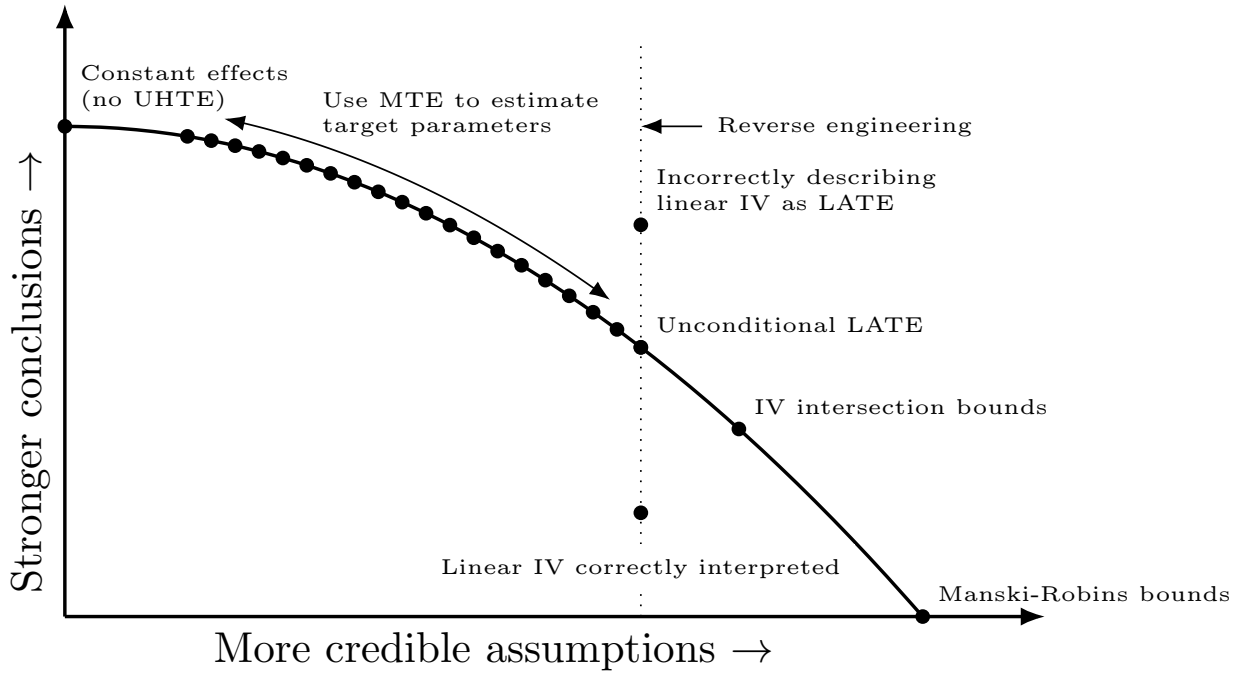
In this section we've discussed some forward engineering approaches for incorporating UHTE into IV models. These run the gamut from the always-possible option of assuming there is no UHTE, to estimating LATEs directly, to extrapolating MTEs for binary or multivalued treatments, to bounding analyses that use only the most essential properties of an IV. What unites all of these approaches as forward engineering is that take the target parameter as the focus and design an estimator suitable for estimating it. What divides them are the target parameters they focus on and the assumptions they use to estimate those target parameters.

Figure 7 provides a familiar graphical tool for comparing these different approaches. It depicts a stylized production possibilities frontier for empirical research using IV with a binary treatment, with the attendant trade-off between assumptions and conclusions that Manski (2003) has described as the Law of Decreasing Credibility. At one corner is the assumption of no UHTE, the strongest assumption that we have considered in this chapter, under which the assumed-to-be-constant treatment effect is point identified under the classical linear IV assumptions. In the other corner are Manski-Robins and IV intersection bounds that do not impose full exogeneity or a selection model, but provide only a bound on the ATE. Selection models allow one to explore the area between these two corners, either at a single point with an unconditional LATE, or at multiple points by extrapolating an MTE curve under additional parametric or shape restrictions.

Our two-dimensional figure can't capture the many dimensions on which empirical

[70] One option that has been explored is rank invariance (Chernozhukov and Hansen, 2005), which can provide point identification without a selection model, albeit with a more complicated relevance condition on how the instrument affects the treatment. Vuong and Xu (2017) show how to combine rank invariance with the usual monotonicity condition to point identify the entire marginal distributions of potential outcomes under a standard instrument relevance condition.

Figure 7: The empirical production possibility frontier for IV methods



Notes: *Two primary trade-offs involved in producing empirical research with a binary treatment.*

approaches differ. What is a stronger conclusion, a bound on an ATE or a point estimate of a LATE? How does one trade-off statistical precision with a broader and more ambitious target parameter? What is the appropriate value for a researcher's time, and how should this value be priced against the quality considerations of the empirical research they produce?

Yet we think that the idea of an empirical production function still nicely describes some useful points that should seem uncontroversial to most economists. Locating inside the frontier is suboptimal: instead of estimating a difficult-to-interpret statistically-weighted average of LATEs, it's possible under essentially the same assumptions to estimate an unconditional LATE using the estimators in Section 4.2. Locating outside of the frontier is not possible: using a linear IV estimator as if it estimates an ATE, or even an unconditional LATE, might be typographically possible to put on paper, but it is logically incoherent under the baseline LATE assumptions. Locating on the frontier is the goal, but reasonable people can disagree about which point on the frontier they prefer.

5 Recommendations for Practice

In this section we distill the discussion of this chapter into concrete recommendations for practitioners. We organize our recommendations around three steps:

Three steps for incorporating UHTE into an IV analysis

1. Assess the likely role of UHTE.
2. Reverse engineer, cautiously.
3. Forward engineer estimates of interpretable target parameters.

Throughout, we take it for granted that the instrument satisfies exclusion and outcome exogeneity. Supporting these assumptions is an important part of a compelling IV analysis, with or without UHTE.

Step 1: Assess the likely role of UHTE

All of the complications, subtleties, and caveats discussed in this chapter vanish under the assumption that there is no UHTE. Our view is that UHTE is probably a generic feature of economic environments. But it still makes sense to assess the source of UHTE and its likely magnitude before embarking on an econometric quest that would be much simpler if it could be convincingly assumed away.

The first task is to think about the nature of the treatment and outcome. Why could treatment effects vary? Is there any reason to think that treatment effects *would not* vary? We expect that the answer to this first question will be negative in almost all cases: heterogeneous treatment effects cannot be ruled out on a priori grounds alone. For example, if the outcome Y_i is binary or discrete, treatment effects are almost necessarily heterogeneous.[71]

In order to create difficulties, however, the UHTE needs to be systematic, in the sense of being correlated with treatment choice. How plausible is it to assume that the UHTE is asystematic? Answering this question requires considering the source of endogeneity for which the IV is intended as a remedy. Does the treatment represent the choice that some economic agent is making? If so, is it possible to write down a plausible model in which the treatment choice is endogenous, but the agent makes their choice without regard to its possible effects on the outcome variable? A compelling

[71] If the outcome is binary and the treatment is also binary, then $Y_i(1) - Y_i(0)$ can take three values: -1, 0, or 1. The assumption that treatment effects are constant implies that the average treatment effect across the population or within any subgroup must also be -1, 0, or 1.

positive answer may make it reasonable to assume that there is no UHTE.

To defend such an assumption, it seems like a reasonable exercise to consider *observed* heterogeneity in treatment effects (OHTE). Given an assumption of no UHTE, patterns of OHTE can be estimated with linear IV by interacting the treatment variable with pre-determined covariates, such as sociodemographic characteristics. It's possible for there to be UHTE but no OHTE, or for there to be OHTE but no UHTE. Nevertheless, it seems reasonable to support any assumption of no UHTE with compelling evidence that there is also no OHTE. Compelling evidence in practice would mean precisely estimated zeros for the interactions between treatment and background characteristics.

Assessing the likely role of UHTE

☐ Is there any a prior reason to believe that treatment effects are constant?

☐ Is there a reasonable behavioral model under which there is UHTE, but economic agents choose treatment without taking it into account?

☐ Is there compelling evidence on the lack of OHTE?

→ If the answer is "no" to all of these questions, then proceed to step two. Otherwise, ask yourself the following question:

☐ Are you willing to maintain an explicit assumption that there is no UHTE?

→ If the answer is "yes," then use linear IV estimators, and be sure to include the explicit assumption of no UHTE when describing your empirical results. If the answer is "no," then proceed to step two.

Step 2: Reverse engineer with caution

Our discussion in Section 3 landed on the conclusion that reverse engineered interpretations of linear IV estimates are often not applicable. However, there are important cases in which these interpretations do apply. Reverse engineering arguments can be successful in these cases, but the necessary assumptions need to be assessed carefully and the interpretations reported accurately.

The first task is determining which setting in Table 3 is applicable and how this interacts with the assumptions about the selection process. In particular, the number of values that the treatment and instrument take is crucial for assessing whether the standard monotonicity condition is reasonable. The most favorable cases are when

both the treatment and instrument are binary or ordered. Unordered treatments require careful thinking about the appropriate selection model. Unordered instruments require careful thinking about whether the standard monotonicity condition is likely to be violated. Unfortunately, there are currently no appealing general-purpose alternatives to the monotonicity condition. Average monotonicity has been suggested recently (Frandsen et al., 2023), but as our discussion in Section 3.4 showed, it is generally no easier to justify than the usual monotonicity condition, except on the narrow technical grounds that it is mathematically weaker. Full exogeneity is required for reverse engineering in all settings.

The second task is determining whether the linear IV specification actually delivers a weakly causal interpretation under full exogeneity and an appropriate monotonicity condition. The primary concern here is satisfying the rich covariates condition. If the instrument is independent of covariates, then rich covariates is automatically satisfied. Otherwise, the covariate specification needs to be scrutinized. The Ramsey (1969) RESET test for a linear regression of Z_i on X_i is the primary tool for doing so; a rejection of the RESET test is a rejection of rich covariates and so also a rejection of the null hypothesis that the linear IV specification produces a weakly causal estimand. If evidence is found that the estimand is not weakly causal, then there's little point in proceeding to its interpretation. One can try adjusting the specification by hand or using machine learning tools to help select the specification, as illustrated in Section 4.2.3.

Assuming that the estimand is weakly causal, the third task is giving it a more concrete interpretation. What factors determine the weights? Which subgroups receive the most weight? The least weight? What counterfactual would the estimand describe? How would one describe the counterfactual in a sentence, or explain it in words to a colleague? If these task are hard for the researcher, it suggests that the interpretation of the estimand is also going to be difficult for the consumer of the research as well. In some cases, such as with ordered treatments, there may be multiple competing interpretations to choose from or discuss jointly. Binarizing the instrument can ease the interpretation challenges created by having multiple complier groups.

The fourth task is the simplest but most important: clearly and transparently communicate the interpretation of the estimand and the assumptions on which the interpretation rests. Be both correct and honest. Incorrectly describing an estimand as "the LATE" is no better than incorrectly describing the exclusion or exogeneity assumptions. Both errors amount to logically incoherent descriptions of causal inference.

The four tasks of compelling reverse engineering can be remembered with the acronym JOSH:

The JOSH method for reverse engineering

☐ **J**udge the setting.

☐ **O**btain a weakly causal interpretation.

☐ **S**crutinize the interpretation.

☐ ***H****onestly* communicate to the audience.

Step 3: Forward engineer estimates of interpretable target parameters

Valiant efforts at implementing the JOSH method of reverse engineering will still end in failure if the setting is simply too complicated to obtain a weakly causal interpretation. Our survey of reverse engineering suggests that this will often be so, as even moderate departures from the baseline LATE setting can make it difficult to obtain a weakly causal interpretation. Even when it is possible to reverse engineer a weakly causal interpretation, an honest description of this interpretation may be convoluted, unclear, or have only a loose connection to the motivating research question. These are all reasons for pursuing forward engineering as a complement to reverse engineering.

Forward engineering requires choosing some target parameters. Which target parameters are useful necessarily depends on the context and the researcher's motivation. What can be said about any given target parameter depends on what assumptions are made. This reflects the necessary trade-off captured in the empirical production frontier (Figure 7). We recommend that researchers explore this frontier by reporting estimates of interesting target parameters under several different sets of assumptions. The linear-in-parameters specification that we used throughout much of Section 4 makes this relatively easy to do, at least if one sticks to point identified settings. Considering partial identification provides further flexibility for exploring the empirical production frontier but raises the difficulty of implementation.

Some target parameters will necessarily be easier to draw conclusions about than others. The unconditional LATE is often point identified nonparametrically while the ATE seldom is. We don't see this as a good reason to omit target parameters, either because they are interesting but too difficult, or easy but not particularly interesting. It is hard to disagree with Imbens's (2010, pp. 414–15) advice to report both estimates of nonparametrically identified quantities, like LATEs, together with estimates of target parameters with higher "external validity." As part of this, Imbens (2010) emphasizes being clear about the degree to which estimates of these different quantities depend on

different assumptions, another point that is hard to disagree with.

Our recommendation for forward engineering is to embrace these two points: estimate LATEs and estimate other target parameters that are relevant to the empirical question. Through it all, be clear and upfront about the role of the maintained assumptions, an important part of which is reporting estimates under different sets of assumptions. This recommendation shouldn't be controversial. It comes with a cost of more difficult implementation. How large of a cost this is depends on the setting.

Forward engineering is now relatively low cost for binary or multivalued treatments. Estimating unconditional LATEs or the unconditional ACR with propensity score weighting is simply a matter of estimating a logistic regression. Estimating MTE curves involves estimating a binary or ordered logistic regression together with some properly-specified linear regressions. Software is available in both `R` and Stata to streamline either task, although the MTE software is currently limited to binary treatments. Even applying machine learning methods like DDML is relatively low cost, albeit potentially demanding computationally.

Deviating from binary or ordered treatments to unordered treatments raises the cost of forward engineering considerably, and at current puts one into less-charted methodological territory. There are a number of successful empirical examples of forward engineering that one can try to follow as a guide (see Section 4.6), however implementation will often require bespoke analysis and coding. This is not a reflection of the difficulty of forward engineering, but rather the difficulty of unordered treatments, a case for which there are also few meaningful reverse engineering results. The same comments apply to the binary or ordered treatment case without the usual monotonicity condition.

6 Conclusion

The literature on including UHTE for IV methods now spans several decades and has been recognized in two Nobel prize awards to three scholars. Reflections on this work have often focused on the question of whether LATEs are interesting quantities. See for example the exchange between Deaton (2010), Imbens (2010), and Heckman and Urzua (2010).

Our review suggests that this question is a bit of a red herring. As we showed in Section 3, the LATE result is so specialized that empirical researchers, who by and large use linear IV, often *aren't actually estimating LATEs*. That's not a problem with the LATE, it's a problem with the practice of what we've described as reverse engineering: starting with an estimator and working backwards to an interpretation.

The concept of a LATE is amenable to reverse engineering in a couple of stylized baseline cases, but by and large the types of complications usually found in empirical work invalidate simple interpretations of linear IV estimates as LATEs, or even as "weakly causal" estimands. We expect that the same type of fragility will also be found in other applications of reverse engineering once researchers start to consider how different forms of misspecification interact with each other.[72]

The obvious alternative is also the oldest one: work forward, instead of backward. As we showed in Section 4, there are now many well-developed and relatively low-cost tools that can be used to forward engineer estimators that estimate specific target parameters with clear interpretations, including LATEs. These estimators do not rely on fundamentally different assumptions, they simply make the assumptions harder to hide and easier to adjust than when reverse engineering. Many of the challenges about modeling selection that arise in forward engineering also arise in reverse engineering. But solving them is easier without self-imposing the straightjacket of the linear IV estimator.

[72] As one example, Blandhol et al. (2022) show that Angrist's (1998) interpretation of the OLS estimand under selection on observables collapses unless the propensity score has implicitly been correctly specified.

Appendices

A Potential outcomes or latent variables? It's just notation ...

Start with a latent variable model of form (3) with $Y_i = f(D_i, \epsilon_i)$. Let the potential outcomes be defined as $Y_i(d) \equiv f(d, \varepsilon_i)$. If treatment states take values in a finite set $\mathcal{D}$, then

$$Y_i = f(D_i, \epsilon_i) = \sum_{d \in \mathcal{D}} \mathbb{1}[D_i = d] f(d, \epsilon_i) \equiv \sum_{d \in \mathcal{D}} \mathbb{1}[D_i = d] Y_i(d) = Y_i(D_i).$$

So, starting with a latent variable model we have constructed potential outcomes that generate the same observed outcome, Y_i. The assumption that $\mathcal{D}$ is finite is just to preserve the familiarity of the summation; if $\mathcal{D}$ is infinite, then consider only the first and final equalities.

Conversely, suppose that there are potential outcomes $Y_i(d)$ for each treatment state in a set $\mathcal{D}$, which we again begin by assuming is finite and enumerated as $\mathcal{D} = \{d_0, d_1, \ldots, d_J\}$. Let $\epsilon_{ij} \equiv Y_i(d_j)$ and $\epsilon_i \equiv (Y_i(d_0), Y_i(d_1), \ldots, Y_i(d_J))$. Then take

$$f(D_i, \epsilon_i) \equiv \sum_{j=0}^{J} \mathbb{1}[D_i = d_j] \epsilon_{ij}.$$

This implies that

$$Y_i = \sum_{j=0}^{J} \mathbb{1}[D_i = d_j] Y_i(d_j) \equiv \sum_{j=0}^{J} \mathbb{1}[D_i = d_j] \epsilon_{ij} \equiv f(D_i, \epsilon_i).$$

So, starting with potential outcomes, we have constructed a latent variable model that generates the same observed outcome, Y_i. The assumption that $\mathcal{D}$ is finite is again unimportant; if $\mathcal{D}$ were infinite then ϵ_i would represent the random function $d \mapsto Y_i(d)$, and f would have domain that includes these random functions.

B Definition of a weakly causal estimand

Blandhol et al. (2022) introduced the definition of a weakly causal estimand as a way of extending the logic of a non-negatively weighted average to estimands that might not be expressible as a weighted average at all. To see how this could arise, consider an OLS estimator of Y_i on D_i and covariates X_i, and let β denote the estimand corresponding to the coefficient on D_i. Suppose that $D_i \in \{0, 1\}$ is binary and that $(Y_i(0), Y_i(1))$ is independent of D_i conditional on X_i. Then a bit of Frisch-Waugh-Lovell algebra shows that

$$\begin{aligned} \beta &= \mathbb{E}\left[\omega_0(X_i)\, \mathbb{E}[Y_i(0)|X_i]\right] + \mathbb{E}\left[\omega_1(X_i)\, \mathbb{E}[Y_i(1)|X_i]\right] \\ \text{where} \quad \omega_d(x) &= \begin{cases} \mathbb{E}[(D_i - X_i'\delta)^2]^{-1}(x'\delta)(1 - \mathbb{E}[D_i|X_i = x]), & \text{if } d = 0 \\ \mathbb{E}[(D_i - X_i'\delta)^2]^{-1}\, \mathbb{E}[D_i|X_i = x](1 - x'\delta), & \text{if } d = 1 \end{cases} \\ \text{with} \quad \delta &\equiv \mathbb{E}[X_i X_i']^{-1}\, \mathbb{E}[X_i D_i]. \end{aligned} \tag{89}$$

In general, there is no way to rewrite (89) as a weighted average of treatment effects, like (9), because $\omega_0(x) \neq -\omega_1(x)$ except when $\mathbb{E}[D_i|X_i] = X_i'\delta$. The notion that is captured by analyzing the sign of the weights for an estimand with a form like (9) becomes more complicated for an estimand that does not have that form, like β in (89).

Here we will develop that notion in a way that is a bit more abstract than in Blandhol et al. (2022) and also allows for unordered treatments. Let F denote a distribution for $(\{Y_i(d)\}_{d\in\mathcal{D}}, \{D_i(z)\}_{z\in\mathcal{Z}})|X_i, Z_i$, where $\mathcal{D}$ and $\mathcal{Z}$ are the set of values that the treatment and instrument take, respectively. Each F generates a distribution $\chi(F)$ of observables (Y_i, D_i, X_i, Z_i) through the definition of potential outcomes and the distribution of (X_i, Z_i), which is not modeled, and is viewed as part of the definition of χ. An estimand is a function that takes a distribution of observables such as $\chi(F)$ and maps it into a number or vector of numbers $\tau(\chi(F))$.[73]

Let $\mathcal{F}$ denote the subset of F that satisfy a set of assumptions. For example, in an IV context with full exogeneity, $\mathcal{F}$ only includes F for which the joint distribution of $Y_i(d)$ and $D_i(z)$ across all d and z is independent of Z_i, conditional on X_i. Usually, $\mathcal{F}$ will also include further assumptions, such as the monotonicity condition, which rules out F that allow for certain types of choice groups, such as defiers.

Let $\mathcal{F}^\diamond \subseteq \mathcal{F}$ denote a subset of F in $\mathcal{F}$ that share a property that we want to be reflected in the estimand, $\tau(\chi(F))$. Let $\mathcal{T}^\diamond$ be a set of values for $\tau(\chi(F))$ that reflect this property. We say that the estimand τ is faithful to the pair $(\mathcal{F}^\diamond, \mathcal{T}^\diamond)$ if

$$F \in \mathcal{F}^\diamond \quad \Rightarrow \quad \tau(\chi(F)) \in \mathcal{T}^\diamond. \tag{90}$$

For example, suppose D_i is ordered, taking values $d_0, d_1, \ldots, d_J$. Then $\mathcal{F}_+^\diamond$ could be the subset of F in $\mathcal{F}$ for which $\mathbb{E}_F[Y_i(d_j) - Y_i(d_{j-1})|G_i = g, X_i = x]$ is non-negative for all $j \geq 1$, all g, and all x, where $\mathbb{E}_F$ denotes expectation taken under F, and G_i is the usual group notation derived from $\{D_i(z)\}_{z\in\mathcal{Z}}$. The set $\mathcal{T}_+^\diamond$ could be the set of non-negative numbers. Then τ is faithful to $(\mathcal{F}_+^\diamond, \mathcal{T}_+^\diamond)$ if $\tau(\chi(F)) \geq 0$ whenever $F \in \mathcal{F}$ is such that all treatment effects are non-negative.

For the ordered treatment case, Blandhol et al. (2022) say that τ is weakly causal if it is faithful to both $(\mathcal{F}_+^\diamond, \mathcal{T}_+^\diamond)$ and $(\mathcal{F}_-^\diamond, \mathcal{T}_-^\diamond)$, where $\mathcal{F}_-^\diamond$ is the subset of $\mathcal{F}$ for which all treatment effects are non-*positive* and $\mathcal{T}_-^\diamond$ is the set of non-positive numbers. Given (90), this means that (i) if all treatment effects are non-negative, then the estimand is also non-negative, and (ii) if all treatment effects are non-positive, then the estimand is also non-positive. This is the same notion that is usually captured by the sign of the weights being non-negative in a decomposition like (9), but formalized in a way that can also be applied to estimands like (89) that do not have a weighted average decomposition.[74]

We can use this framework to extend the definition of a weakly causal estimand to the

[73] In what follows, τ could also just be some quantity that is determined by $\chi(F)$; it need not necessarily be an estimand in the sense of being the limit of some estimator.

[74] To show that an estimand is not weakly causal it suffices to find an $F \in \mathcal{F}_+^\diamond$ for which $\tau(\chi(F)) < 0$. For example, to show that (89) is not weakly causal, one can construct an F such that $\mathbb{E}_F[Y_i(1) - Y_i(0)|X_i = x] \geq 0$ for all x, but with $\mathbb{E}_F[Y_i(0)|X_i = x]$ chosen to make $\beta < 0$. This is always possible when $\omega_0(x) \neq -\omega_1(x)$,

unordered case by choosing different pairs $(\mathcal{F}^\diamond, \mathcal{T}^\diamond)$. Let $\mathcal{F}^\diamond_{0\to\ell,+}$ denote the subset of $\mathcal{F}$ for which the contrast $\mathbb{E}_F[Y_i(d_\ell) - Y_i(d_0)|G_i = g, X_i = x]$ between a particular treatment state d_ℓ and a base state d_0 is non-negative for all g and x, ignoring the other treatment states. Keep $\mathcal{T}_+$ defined as the set of non-negative numbers. Let $\mathcal{F}^\diamond_{0\to\ell,-}$ and $\mathcal{T}_-$ be defined symmetrically for the non-positive case. Then an estimand τ is weakly causal for the treatment contrast between d_ℓ and d_0 if it is faithful to both $(\mathcal{F}^\diamond_{0\to\ell,+}, \mathcal{T}_+)$ and $(\mathcal{F}^\diamond_{0\to\ell,-}, \mathcal{T}_-)$.

Notice that because $\mathcal{F}^\diamond \subseteq \mathcal{F}$, whether an estimand is faithful to $(\mathcal{F}^\diamond, \mathcal{T}^\diamond)$ depends on the maintained assumptions on $\mathcal{F}$. For example, an estimand may not be weakly causal if $\mathcal{F}$ includes F that do not satisfy a monotonicity condition, but might become weakly causal when $\mathcal{F}$ is restricted to only include F that do satisfy a monotonicity condition. The nature of (90) means that making $\mathcal{F}$ a smaller set makes it easier for an estimand to be faithful to any given $(\mathcal{F}^\diamond, \mathcal{T}^\diamond)$ pair. The literature on reverse engineering can be seen as an effort to choose τ and $\mathcal{F}$ in a way that ensures τ is weakly causal.

C Deriving the average causal response and an alternative decomposition

We first derive (22), which to our knowledge has only appeared in the literature for the case when $d_j = j$ are the integers. Start by decomposing the outcome as

$$\begin{aligned} Y_i &= \sum_{j=0}^{J} \mathbb{1}[D_i = d_j] Y_i(d_j) \\ &= \sum_{j=0}^{J} \left(\mathbb{1}[D_i \geq d_j] - \mathbb{1}[D_i \geq d_{j+1}]\right) Y_i(d_j) = Y_i(d_0) + \sum_{j=1}^{J} \mathbb{1}[D_i \geq d_j] \left(Y_i(d_j) - Y_i(d_{j-1})\right), \end{aligned} \tag{92}$$

where the final equality follows from a change of variables and d_{J+1} in the summand when $j = J$ can be interpreted as any value larger than d_J. Taking the difference in conditional expectation with respect to z and applying full exogeneity then gives

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] = \sum_{j=1}^{J} \mathbb{E}\left[\left(\mathbb{1}[D_i(1) \geq d_j] - \mathbb{1}[D_i(0) \geq d_j]\right) \left(Y_i(d_j) - Y_i(d_{j-1})\right) \right]. \tag{93}$$

Under the monotonicity condition, the event that $D_i(1) < d_j$ and $D_i(0) \geq d_j$ has probability zero, so the difference in indicators only takes the value zero or one with positive probability.

---

because one can write β as

$$\beta = \mathbb{E}[(\omega_0(X_i) + \omega_1(X_i))\, \mathbb{E}_F[Y_i(0)|X_i]] + \mathbb{E}[\omega_1(X_i)(\mathbb{E}_F[Y_i(1) - Y_i(0)|X_i])]. \tag{91}$$

Even if $\omega_1(x) \geq 0$ for all x, the fact that $\omega_0(x) \neq -\omega_1(x)$ for all x means that $\beta = \tau(\chi(F))$ depends on some $\mathbb{E}_F[Y_i(0)|X_i = x]$, and therefore can change signs as F varies across $\mathcal{F}^\diamond_+$.

Conditioning on the event that it is one leaves

$$\begin{aligned}&\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]\\&\quad=\sum_{j=1}^{J}\mathbb{P}\left[D_i(1)\geq d_j>D_i(0)\right]\mathbb{E}\left[Y_i(d_j)-Y_i(d_{j-1})|D_i(1)\geq d_j>D_i(0)\right].\end{aligned}$$

The same algebraic argument can be applied to the denominator of the Wald estimand with Y_i replaced by $D_i=\sum_{j=0}^{J}\mathbb{1}[D_i=d_j]d_j$, yielding

$$\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]=\sum_{j=1}^{J}\mathbb{P}\left[D_i(1)\geq d_j>D_i(0)\right](d_j-d_{j-1}).$$

Taking the ratio gives the right-hand side of (22), which is equal to the simple IV estimand when the instrument is binary (recall (13)).

To derive the alternative group-based decomposition (23), let $\mathcal{G}(j)\equiv\{(g(0),g(1))\in\mathcal{G}:g(1)\geq d_j>g(0)\}$ denote the subset of groups represented by the conditioning event in the unit causal response (22), recalling that $\mathcal{G}$ is the set of groups who could have non-zero probability under the monotonicity condition. Then (22) can be written as

$$\frac{\mathbb{C}[Y_i,Z_i]}{\mathbb{C}[D_i,Z_i]}=\sum_{j=1}^{J}\frac{\mathbb{P}[G_i\in\mathcal{G}(j)]}{\sum_{\ell=1}^{J}\mathbb{P}[G_i\in\mathcal{G}(\ell)](d_\ell-d_{\ell-1})}\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})|G_i\in\mathcal{G}(j)].$$

Focusing on the numerator, write

$$\begin{aligned}&\sum_{j=1}^{J}\mathbb{P}[G_i\in\mathcal{G}(j)]\,\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})|G_i\in\mathcal{G}(j)]\\&\qquad=\sum_{j=1}^{J}\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})\mathbb{1}[G_i\in\mathcal{G}(j)]]\\&\qquad=\sum_{j=1}^{J}\sum_{g\in\mathcal{G}}\mathbb{1}[g(1)\geq d_j>g(0)]\,\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})\mathbb{1}[G_i=g]]\\&\qquad=\sum_{g\in\mathcal{G}}\sum_{j=g(0)+1}^{g(1)}\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})|G_i=g]\,\mathbb{P}[G_i=g]\\&\qquad=\sum_{g\in\mathcal{G}}\mathbb{P}[G_i=g]\,\mathbb{E}[Y_i(g(1))-Y_i(g(0))|G_i=g].\end{aligned}$$

The same argument applied to the denominator yields

$$\sum_{\ell=1}^{J}\mathbb{P}[G_i\in\mathcal{G}(\ell)](d_\ell-d_{\ell-1})=\sum_{g'\in\mathcal{G}}\mathbb{P}[G_i=g'](g'(1)-g'(0)).$$

Expression (23) follows after multiplying and dividing each term in the numerator by $g(1)-g(0)$.

D Estimating the average causal response with covariates

Suppose that D_i is multivalued and ordered, taking values $d_0, d_1, \ldots, d_J$, as in Section 3.6. Define $W_i(z) \equiv Y_i(D_i(z))$ as before. The second half of (36) did not depend on how many values D_i takes, so we still have:

$$\frac{\mathbb{E}[W_i(1) - W_i(0)]}{\mathbb{E}[D_i(1) - D_i(0)]} = \frac{\mathbb{E}\left[\mathbb{E}[Y_i|Z_i = 1, X_i] - \mathbb{E}[Y_i|Z_i = 0, X_i]\right]}{\mathbb{E}\left[\mathbb{E}[D_i|Z_i = 1, X_i] - \mathbb{E}[D_i|Z_i = 0, X_i]\right]},$$

But the interpretation of the left-hand side of this equality changes now that D_i takes multiple ordered values. Write $W_i(z) \equiv Y_i(D_i(z))$ like (92):

$$W_i(z) = Y_i(d_0) + \sum_{j=1}^{J} \mathbb{1}[D_i(z) \geq d_j] \left(Y_i(d_j) - Y_i(d_{j-1})\right).$$

Then $\mathbb{E}[W_i(1) - W_i(0)]$ matches the expression (92) for the Z_i-differenced conditional mean of Y_i derived when Z_i was unconditionally exogenous:

$$\mathbb{E}[W_i(1) - W_i(0)] = \sum_{j=1}^{J} \mathbb{E}\left[\left(\mathbb{1}[D_i(1) \geq d_j] - \mathbb{1}[D_i(0) \geq d_j]\right) \left(Y_i(d_j) - Y_i(d_{j-1})\right) \right].$$

The rest of the derivation in Appendix C shows that the right-hand side is the numerator of the ACR. The argument for the denominator follows the same logic with $W_i(z) \equiv Y_i(D_i(z))$ replaced by simply $D_i(z) = \sum_{j=0}^{J} \mathbb{1}[D_i(z) = d_j] d_j$.

E Derivations for marginal treatment effects

This appendix contains some derivations relevant binary marginal treatment effects for binary treatments.

E.1 Derivations of weighting expressions

Consider the ATT expression given in (47) and Table 5. We derive this by iterating expectations:

$$\begin{aligned}
\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] &= \mathbb{E}\left[\left(Y_i(1) - Y_i(0)\right) \frac{\mathbb{1}[U_i \leq p(Z_i, X_i)]}{\mathbb{P}[D_i = 1]} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(1) - Y_i(0)|U_i, X_i, Z_i\right] \frac{\mathbb{1}[U_i \leq p(Z_i, X_i)]}{\mathbb{P}[D_i = 1]} \right] \\
&= \mathbb{E}\left[\left(\text{MTR}(1|U_i, X_i) - \text{MTR}(0|U_i, X_i)\right) \frac{\mathbb{1}[U_i \leq p(Z_i, X_i)]}{\mathbb{P}[D_i = 1]} \right] \\
&= \mathbb{E}\left[\int_0^1 \left(\text{MTR}(1|u, X_i) - \text{MTR}(0|u, X_i)\right) \frac{\mathbb{1}[u \leq p(Z_i, X_i)]}{\mathbb{P}[D_i = 1]} \, \mathrm{d}u \right].
\end{aligned}$$

The third equality used full exogeneity and the definition of the MTR. The fourth equality iterated expectation on U_i given X_i and Z_i and used the normalization that the conditional distribution of U_i is uniformly distributed.

The linear-in-parameters representation (49) for the conditional mean of the observed outcome is derived like this:

$$\begin{aligned}\mathbb{E}[Y_i|D_i=1,P_i=\bar{p},X_i] &= \mathbb{E}[Y_i(1)|U_i\leq\bar{p},P_i=\bar{p},X_i]\\ &= \mathbb{E}[Y_i(1)|U_i\leq\bar{p},X_i]\\ &= \mathbb{E}\left[\mathbb{E}\left[Y_i(1)|U_i,X_i\right]\big|U_i\leq\bar{p},X_i\right]\\ &= \mathbb{E}\left[\textsc{mtr}(1|U_i,X_i)\big|U_i\leq\bar{p},X_i\right]=\frac{1}{\bar{p}}\int_0^{\bar{p}}\textsc{mtr}(1|u,X_i)\,\mathrm{d}u,\end{aligned}$$

where the second equality uses full exogeneity that the fact that $P_i\equiv p(Z_i,X_i)$ is a function of Z_i and X_i. A symmetric derivation for the untreated arm gives

$$\mathbb{E}[Y_i|D_i=0,P_i=\bar{p},X_i]=\frac{1}{(1-\bar{p})}\int_{1-\bar{p}}^{1}\textsc{mtr}(0|u,X_i)\,\mathrm{d}u.$$

Putting them together and applying the linear-in-parameters assumption (48) produces (49):

$$\begin{aligned}\mathbb{E}[Y_i|D_i,P_i,X_i] &= \frac{(1-D_i)}{(1-P_i)}\int_{1-P_i}^{1}\textsc{mtr}(0|u,X_i)\,\mathrm{d}u+\frac{D_i}{P_i}\int_0^{P_i}\textsc{mtr}(1|u,X_i)\,\mathrm{d}u\\ &= \frac{(1-D_i)}{(1-P_i)}\int_{1-P_i}^{1}\sum_{k=1}^{d_\theta}\theta_k b_k(0|u,X_i)\,\mathrm{d}u+\frac{D_i}{P_i}\int_0^{P_i}\sum_{k=1}^{d_\theta}\theta_k b_k(1|u,X_i)\,\mathrm{d}u\\ &= \sum_{k=1}^{d_\theta}\theta_k B_{ik},\end{aligned}$$

where B_{ik} is as defined in (49).

Lastly, the expression (62) used to derive the LIV follows from

$$\begin{aligned}\mathbb{E}[D_i(Y_i(1)-Y_i(0))|P_i,X_i] &= \mathbb{E}\left[\mathbb{1}[U_i\leq P_i](Y_i(1)-Y_i(0))|P_i,X_i\right]\\ &= \mathbb{E}\left[\mathbb{1}[U_i\leq P_i]\textsc{mte}(U_i,X_i)|P_i,X_i\right]=\int_0^{P_i}\textsc{mte}(U_i,X_i)\,\mathrm{d}u,\end{aligned}$$

again making use of the normalization that U_i is uniform given Z_i and X_i, and so also given P_i and X_i.

E.2 The normal selection model

This section provides detail on the derivation of (56). With the pre-normalization selection equation the propensity score satisfies

$$p(z)\equiv\mathbb{P}[D_i=1|Z_i=z]=\mathbb{P}[V_i\leq\nu(z)]=\Phi\left(\nu(z)\right),$$

noting that V_i has mean zero and variance one. Consider $b_4(d|u) = d\Phi^{-1}(u)$. Then

$$B_{i4} = D_i \frac{1}{P_i} \int_0^{P_i} \Phi^{-1}(u)\,\mathrm{d}u \quad \text{where} \quad P_i = \Phi\left(\nu(Z_i)\right).$$

A change of variables to $v \equiv \Phi^{-1}(u)$ leads to a mean of a standard normal truncated at $\Phi^{-1}(P_i)$:

$$B_{i4} = D_i \int_{-\infty}^{\Phi^{-1}(P_i)} v\, \frac{\phi(v)}{\Phi(\Phi^{-1}(P_i))}\,\mathrm{d}v.$$

This can be expressed in terms of the inverse Mills' ratio (e.g Hansen, 2022a, pg. 116), so that

$$B_{i4} = -D_i \frac{\phi\left(\Phi^{-1}(P_i)\right)}{\Phi\left(\Phi^{-1}(P_i)\right)} \equiv -D_i \lambda(\Phi^{-1}(P_i)).$$

The corresponding term for the untreated arm is derived similarly.

E.3 Saturated MTR specifications reproduce the LATE

We give a simple and general justification of the finding in Brinch et al. (2017) and Kline and Walters (2019) that in saturated settings even a misspecified MTR reproduces the usual LATE. For simplicity, suppose that Z_i is binary and there are no covariates. Consider $\mathbb{E}[Y_i|Z_i = z]$ as a target parameter. An MTR function parameterized by θ implies the following values for this target parameter:

$$\mathbb{E}_\theta[Y_i|Z_i = z] = \mathbb{E}_\theta[Y_i(0)] + \mathbb{E}_\theta[\mathbb{1}[U_i \leq p(z)](Y_i(1) - Y_i(0))],$$

where $\mathbb{E}_\theta$ indicates expectation taken under the assumption that the MTR function is parameterized by θ. When the MTR follows the linear-in-parameters specification (48),

$$\mathbb{E}_\theta[Y_i|Z_i = z] = \sum_{k=1}^{d_\theta} \theta_k \int_0^1 b_k(0|u)\,\mathrm{d}u + \theta_k \int_0^{p(z)} \left(b_k(1|u) - b_k(0|u)\right)\,\mathrm{d}u.$$

The LATE formed from θ is then

$$\frac{\mathbb{E}_\theta[Y_i|Z_i = 1] - \mathbb{E}_\theta[Y_i|Z_i = 0]}{p(1) - p(0)} = \sum_{k=1}^{d_\theta} \theta_k \int_0^1 \left(b_k(1|u) - b_k(0|u)\right) \frac{\mathbb{1}[p(0) < u \leq p(1)]}{p(1) - p(0)}\,\mathrm{d}u,$$

which matches the expression given in Table 5, with $\underline{u} = p(0)$, $\overline{u} = p(1)$, and $\textsc{mtr}(d|u) = \sum_{k=1}^{d_\theta} \theta_k b_k(d|u)$.

Given this observation, it suffices to show that if the implied regression specification (49) is saturated, then $\mathbb{E}_m[Y_i|Z_i = 1] = \mathbb{E}[Y_i|Z_i = 1]$ even if the MTR function m is misspecified. Saturation implies that for all values of d and z,

$$\mathbb{E}[Y_i|D_i = d, Z_i = z] = \mathbb{E}[Y_i|D_i = d, P_i = p(z)] = \sum_{k=1}^{d_\theta} \theta_k B_k(d, p(z)), \tag{94}$$

where

$$B_k(d, p(z)) \equiv \left(\frac{1-d}{1-p(z)}\right) \int_{p(z)}^{1} b_k(0|u)\,\mathrm{d}u + \frac{d}{p(z)} \int_{0}^{p(z)} b_k(1|u)\,\mathrm{d}u.$$

So, in particular,

$$\begin{aligned}
\mathbb{E}[Y_i|Z_i = z] &= \mathbb{E}[Y_i|D_i = 0, Z_i = z](1 - p(z)) + \mathbb{E}[Y_i|D_i = 1, Z_i = z]p(z) \\
&= \sum_{k=1}^{d_\theta} \theta_k \left(B_k(0, p(0))(1 - p(0)) + B_k(1, p(1))p(1)\right) \\
&= \sum_{k=1}^{d_\theta} \theta_k \left(\int_{p(z)}^{1} b_k(0|u)\,\mathrm{d}u + \int_{0}^{p(z)} b_k(1|u)\,\mathrm{d}u\right) \\
&= \sum_{k=1}^{d_\theta} \theta_k \int_{0}^{1} b_k(0|u)\,du + \sum_{k=1}^{d_\theta} \theta_k \int_{0}^{p(z)} \left(b_k(1|u) - b_k(0|u)\right)\,du = \mathbb{E}_\theta[Y_i|Z_i = z].
\end{aligned}$$

Note that the same argument works for the sample as well: a saturated specification will continue to satisfy the sample analog of (94) with $\mathbb{E}[Y_i|D_i = d, Z_i = z]$ and $p(z)$ replaced by their conditional sample means, and with θ_k replaced by its estimator. The argument holds outside of a regression context and doesn't depend on the linear-in-parameters specification. The driving observation is just that in a saturated model, even an incorrectly-specified MTR function that is fit to the data will produce the same value of $\mathbb{E}[Y_i|D_i = d, Z_i = z]$ for all values of d and z, and so will also produce the same estimate of the LATE.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263. 53

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do Parents Value School Effectiveness?" *American Economic Review*, 110, 1502–1539. 87

Acerenza, S. (2024): "Partial Identification of Marginal Treatment Effects with Discrete Instruments and Misreported Treatment*," *Oxford Bulletin of Economics and Statistics*, 86, 74–100. 78

Acerenza, S., K. Ban, and D. Kédagni (2023): "Marginal Treatment Effects with a Misclassified Treatment," . 78

Acerenza, S., V. Possebom, and P. H. C. Sant'Anna (2024): "Was Javert Right to Be Suspicious? Unpacking Treatment Effect Heterogeneity of Alternative Sentences on Time-to-Recidivism in Brazil," . 61

Agan, A., J. L. Doleac, and A. Harvey (2023): "Misdemeanor Prosecution," *The Quarterly Journal of Economics*, 138, 1453–1505.

Ahrens, A., C. B. Hansen, M. E. Schaffer, and T. Wiemann (2023): "Ddml: Double/Debiased Machine Learning in Stata," . 56

——— (2024a): *Ddml: Double/Debiased Machine Learning*. 56, 57

——— (2024b): "Model Averaging and Double Machine Learning," . 55, 56

Aizawa, N., C. Mommaerts, and S. L. Rennane (2023): "Firm Accommodation After Disability: Labor Market Impacts and Implications for Social Insurance," .

Alessie, R. J. M., V. Angelini, J. O. Mierau, and L. Viluma (2020): "Moral Hazard and Selection for Voluntary Deductibles," *Health Economics*, 29, 1251–1269.

Alvarez, L. A. and R. Toneto (2024): "The Interpretation of 2SLS with a Continuous Instrument: A Weighted LATE Representation," *Economics Letters*, 237, 111658. 21

Andresen, M. E. (2018): "Exploring Marginal Treatment Effects: Flexible Estimation Using Stata," *The Stata Journal: Promoting communications on statistics and Stata*, 18, 118–158. 71, 73, 74

Andresen, M. E. and M. Huber (2021): "Instrument-Based Estimation with Binarised Treatments: Issues and Tests for the Exclusion Restriction," *The Econometrics Journal*, 24, 536–558. 34

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753. 74

Angrist, J. D. (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288. 4, 100

——— (2022): "Empirical Strategies in Economics: Illuminating the Path From Cause to Effect," *Econometrica*, 90, 2509–2539. 45

Angrist, J. D. and W. N. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*, 88, 450–477. 25, 26, 27

Angrist, J. D. and I. Fernández-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," in *Advances in Economics and Econometrics*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, 401–434. 49, 50

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527. 32

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 20, 32, 33, 34, 41, 80

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67. 44

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. 10, 16, 25

Angrist, J. D. and A. B. Krueger (1999): "Chapter 23 Empirical Strategies in Labor Economics," Elsevier, vol. Volume 3, Part A, 1277–1366. 45

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013): "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27. 53

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. 33, 41, 44, 45, 46, 53

Arbour, W. (2022): "Can Recidivism Be Prevented from Behind Bars? Evidence from a Behavioral Program," *Working Paper*.

Arnold, D., W. Dobbie, and P. Hull (2022): "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, 112, 2992–3038. 78

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

Arteaga, C. (2023): "Parental Incarceration and Children's Educational Attainment," *Review of Economics and Statistics*, 105, 1394–1410. 37, 80

Bai, Y., S. Huang, S. Moon, A. M. Shaikh, and E. J. Vytlacil (2024): "On the Identifying Power of Monotonicity for Average Treatment Effects," . 92

Bailey, M. J. and A. Goodman-Bacon (2015): "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans," *American Economic Review*, 105, 1067–1104. 45

Baker, S. G. and K. S. Lindeman (1994): "The Paired Availability Design: A Proposal for Evaluating Epidural Analgesia during Labor," *Statistics in Medicine*, 13, 2269–2278. 20, 66

Balke, A. and J. Pearl (1993): "Nonparametric Bounds on Causal Effects from Partial Compliance Data," *Technical Report R-199, University of California, Los Angeles, Computer Science Department*. 92

——— (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176. 17, 92

Balla-Elliott, D. (2023): "Identifying Causal Effects in Information Provision Experiments," . 48

Bandiera, O., N. Buehren, R. Burgess, M. Goldstein, S. Gulesci, I. Rasul, and M. Sulaiman (2020): "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa," *American Economic Journal: Applied Economics*, 12, 210–259.

Baron, E. J. and M. Gross (2022): "Is There a Foster Care-To-Prison Pipeline? Evidence from Quasi-Randomly Assigned Investigators," .

Basu, A., J. J. Heckman, S. Navarro-Lozano, and S. Urzua (2007): "Use of Instrumental Variables in the Presence of Heterogeneity and Self-Selection: An Application to Treatments of Breast Cancer Patients," *Health Economics*, 16, 1133–1157.

Basu, A., A. B. Jena, D. P. Goldman, T. J. Philipson, and R. Dubois (2014): "Heterogeneity in Action: The Role of Passive Personalization in Comparative Effectiveness Research," *Health Economics*, 23, 359–373.

Becker, G. (1964): *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition*, National Bureau of Economic Research, Inc. 7

Becker, G. S. (1967): *Human Capital and the Personal Distribution of Income: An Analytical Approach*, 1, Institute of Public Administration. 7

Behaghel, L., B. Crepon, and M. Gurgand (2013): "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial," . 35, 36, 37

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681. 44

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017): "Program Evaluation and Causal Inference With High-Dimensional Data," *Econometrica*, 85, 233–298. 55

Benson, D., M. A. Masten, and A. Torgovitsky (2022): "Ivcrc: An Instrumental-Variables Estimator for the Correlated Random-Coefficients Model," *The Stata Journal: Promoting communications on statistics and Stata*, 22, 469–495. 84

Berry, J., G. Fischer, and R. Guiteras (2020): "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana," *Journal of Political Economy*, 128, 1436–1473.

Bhattacharya, J., A. M. Shaikh, and E. Vytlacil (2012): "Treatment Effect Bounds: An Application to Swan-Ganz Catheterization," *Journal of Econometrics*, 168, 223–243. 17, 92

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2020): "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 128, 1269–1324. 27, 43

Bhuller, M. and H. Sigstad (2024): "2SLS with Multiple Treatments," *Journal of Econometrics*, 242, 105785. 37, 38

Björklund, A. and R. Moffitt (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *The Review of Economics and Statistics*, 69, 42–49. 58

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When Is TSLS Actually LATE?" Tech. Rep. w29709, National Bureau of Economic Research, Cambridge, MA. 18, 40, 41, 43, 44, 45, 100, 101, 102

——— (2024): "When Is TSLS Actually LATE?" Tech. Rep. w29709, National Bureau of Economic Research, Cambridge, MA. 53

Blundell, R. and M. C. Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics," *The Journal of Human Resources*, 44, 565–640. 45

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75, 323–363. 91

Brave, S. and T. Walstrum (2014): "Estimating Marginal Treatment Effects Using Parametric and Semiparametric Methods," *The Stata Journal: Promoting communications on statistics and Stata*, 14, 191–217. 73

Briggs, J., G. Sachs, A. Caplin, S. Leth-Petersen, and C. Tonetti (2024): "Identification of Marginal Treatment Effects Using Subjective Expectations," . 59

Brinch, C. N., M. Mogstad, and M. Wiswall (2012): "Beyond LATE with a Discrete Instrument," *Working paper*. 65

——— (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039. 65, 66, 68, 70, 107

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326. 41

Calvi, R., A. Lewbel, and D. Tommasi (2022): "LATE With Missing or Mismeasured Treatment," *Journal of Business & Economic Statistics*, 40, 1701–1717. 78

Canay, I. A. and A. M. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 271–306. 70

Card, D. (1993): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," Tech. rep., National Bureau of Economic Research. 55

——— (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. N. Christofides, K. E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222. 13

——— (1999): "Chapter 30 The Causal Effect of Education on Earnings," Elsevier, vol. Volume 3, Part A, 1801–1863. 6, 65

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127–1160. 7, 55, 57, 65

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2453–2483. 5

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," *Econometrica*, 78, 377–394. 76

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101, 2754–81. 59, 71

Carneiro, P. and S. Lee (2009): "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality," *Journal of Econometrics*, 149, 191–208. 61, 73

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

Carr, T. and T. Kitagawa (2023): "Testing Instrument Validity with Covariates," . 17

Carrillo, P., D. Donaldson, D. Pomeranz, and M. Singhal (2023): "Misallocation in Firm Production: A Nonparametric Analysis Using Procurement Lotteries," Tech. Rep. w31311, National Bureau of Economic Research. 84

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2016): "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *The Journal of Politics*, 78, 1229–1248. 5

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 5549–5632. 67

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. 55, 56

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268. 85

Chernozhukov, V., I. Fernández-Val, W. Newey, S. Stouli, and F. Vella (2020): "Semiparametric Estimation of Structural Functions in Nonseparable Triangular Models," *Quantitative Economics*, 11, 503–533. 84

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. 93

Chyn, E., B. Frandsen, and E. C. Leslie (2024): "Examiner and Judge Designs in Economics: A Practitioner's Guide," . 29

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions," *Labour Economics*, 41, 47–60. 57, 71

——— (2018): "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance," *Journal of Political Economy*, 126, 2356–2409. 74, 76, 81, 93

Coury, M., T. Kitagawa, A. Shertzer, and M. Turner (2022): "The Value of Piped Water and Sewers: Evidence from 19th Century Chicago," .

Cunha, F., J. J. Heckman, and S. Navarro (2007): "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds," *International Economic Review*, 48, 1273–1309. 80

Cunningham, S. (2021): *Causal Inference: The Mixtape*, Yale University Press. 28, 46

Dahl, G. B. (2002): "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70, 2367–2420. 87

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014): "Family Welfare Cultures," *The Quarterly Journal of Economics*, 129, 1711–1752. 27, 77

Dal Bó, E., F. Finan, N. Y. Li, and L. Schechter (2021): "Information Technology and Government Decentralization: Experimental Evidence From Paraguay," *Econometrica*, 89, 677–701.

Daljord, Ø., C. F. Mela, J. M. T. Roos, J. Sprigg, and S. Yao (2023): "The Design and Targeting of Compliance Promotions," *Marketing Science*, 42, 866–891. 69

de Chaisemartin, C. (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8, 367–396. 30

de Chaisemartin, C. and X. D'Haultfoeuille (2018): "Fuzzy Differences-in-Differences," *The Review of Economic Studies*, 85, 999–1028. 45

de Chaisemartin, C. and Z. Lei (2023): "More Robust Estimators for Instrumental-Variable Panel Designs,

With An Application to the Effect of Imports from China on US Employment," . 45
De Groote, O. and K. Declercq (2021): "Tracking and Specialization of High Schools: Heterogeneous Effects of School Choice," *Journal of Applied Econometrics*, 36, 898–916.
De Haan, M. and E. Leuven (2020): "Head Start and the Distribution of Long-Term Education and Labor Market Outcomes," *Journal of Labor Economics*, 38, 727–765. 91
Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48, 424–455. 16, 99
Depalo, D. (2020): "Explaining the Causal Effect of Adherence to Medication on Cholesterol through the Marginal Patient," *Health Economics*, n/a.
Dinkelman, T. (2011): "The Effects of Rural Electrification on Employment: New Evidence from South Africa," *American Economic Review*, 101, 3078–3108. 13
Dobbie, W., J. Goldin, and C. S. Yang (2018): "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–240. 43
Dobbie, W. and J. Song (2015): "Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection," *American Economic Review*, 105, 1272–1311. 43
Donald, S. G., Y.-C. Hsu, and R. P. Lieli (2014): "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT," *Journal of Business & Economic Statistics*, 32, 395–415. 52
Doyle Jr., J. J. (2007): "Child Protection and Child Outcomes: Measuring the Effects of Foster Care," *The American Economic Review*, 97, 1583–1610. 27
——— (2008): "Child Protection and Adult Crime: Using Investigator Assignment to Estimate Causal Effects of Foster Care," *Journal of Political Economy*, 116, 746–770.
Dubin, J. A. and D. L. McFadden (1984): "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, 52, 345. 87
Dutz, D., M. Greenstone, A. Hortaçsu, S. Lacouture, M. Mogstad, D. Roumis, A. M. Shaikh, A. Torgovitsky, and W. Van Dijk (2023a): "Selection Bias in Voluntary Random Testing: Evidence from a COVID-19 Antibody Study," *AEA Papers and Proceedings*, 113, 562–566. 23
Dutz, D., M. Greenstone, A. Hortaçsu, S. Lacouture, M. Mogstad, A. M. Shaikh, A. Torgovitsky, and W. van Dijk (2023b): "Representation and Hesitancy in Population Health Research: Evidence from a COVID-19 Antibody Study," . 23
Dutz, D., I. Huitfeldt, S. Lacouture, M. Mogstad, A. Torgovitsky, and W. van Dijk (2022): "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias," Tech. Rep. w29549, National Bureau of Economic Research, Cambridge, MA. 23, 78
Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015): "The Generalized Roy Model and the Cost-Benefit Analysis of Social Programs," *Journal of Political Economy*, 123, 413–443. 63
Fang, Z. and A. Santos (2019): "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86, 377–412. 74
Felfe, C. and R. Lalive (2018): "Does Early Child Care Affect Children's Development?" *Journal of Public Economics*, 159, 33–53.
Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 84, 85
Flores, C. A., X. Chen, et al. (2018): *Average Treatment Effect Bounds with an Instrumental Variable: Theory and Practice*, Springer. 92
Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging Judge Fixed Effects," *American Economic Review*, 113, 253–277. 17, 27, 28, 97
Frangakis, C. E. and D. B. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. 12
French, E. and J. Song (2014): "The Effect of Disability Insurance Receipt on Labor Supply," *American Economic Journal: Economic Policy*, 6, 291–337.
Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75. 33, 52, 77
Galasso, A., M. Schankerman, and C. J. Serrano (2013): "Trading and Enforcing Patent Rights," *The RAND Journal of Economics*, 44, 275–312.
Garen, J. (1984): "The Returns to Schooling: A Selectivity Bias Approach with a Continuous Choice Variable," *Econometrica*, 52, 1199. 65
Gathmann, C., C. Vonnahme, A. Busse, and J. Kim (2021): *Marginal Returns to Citizenship and Educational Performance*, DE: RWI.
Gautier, E. (2021): "Relaxing Monotonicity in Endogenous Selection Models and Application to Surveys," in *Advances in Contemporary Statistics and Econometrics*, ed. by A. Daouia and A. Ruiz-Gazen, Cham: Springer International Publishing, 59–78. 77
Gautier, E. and S. Hoderlein (2015): "A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation," *arXiv:1109.0362 [math, stat]*. 77
Gelbach, J. B. (2002): "Public Schooling for Young Children and Maternal Labor Supply," *The American Economic Review*, 92, 307–322. 13, 89, 90, 91

Geweke, J., G. Gowrisankaran, and R. J. Town (2003): "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71, 1215–1238. 87

Goff, L. (2024): "A Vector Monotonicity Assumption for Multiple Instruments," *Journal of Econometrics*, 241, 105735. 77

Goldberger, A. S. (1983): "ABNORMAL SELECTION BIAS," in *Studies in Econometrics, Time Series, and Multivariate Statistics*, ed. by SAMUEL. Karlin, TAKESHI. Amemiya, and L. A. Goodman, Academic Press, 67–84. 68

Goldin, J., I. Z. Lurie, and J. McCubbin (2021): "Health Insurance and Mortality: Experimental Evidence from Taxpayer Outreach," *The Quarterly Journal of Economics*, 136, 1–49. 80

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2024): "Contamination Bias in Linear Regressions," . 5

Gollin, D. and C. Udry (2021): "Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture," *Journal of Political Economy*, 129, 1–80. 84

Gonçalves, F. M. and S. Mello (2023): "Police Discretion and Public Safety," .

Gong, J., Y. Lu, and H. Xie (2020): "The Average and Distributional Effects of Teenage Adversity on Long-Term Health," *Journal of Health Economics*, 71, 102288.

Goodman, J., O. Gurantz, and J. Smith (2020): "Take Two! SAT Retaking and College Enrollment Gaps," *American Economic Journal: Economic Policy*, 12, 115–158. 53

Goodman-Bacon, A. (2021): "Difference-in-Differences with Variation in Treatment Timing," *Journal of Econometrics*, 225, 254–277. 5

Greene, W. H. and D. A. Hensher (2009): *Modeling Ordered Choices*, Cambridge University Press. 32, 79

Griliches, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22. 7

Gronau, R. (1974): "Wage Comparisons–A Selectivity Bias," *Journal of Political Economy*, 82, 1119–1143. 5, 10, 49

Gulotty, R. and A. Z. Yu (2023): "Must Watch Propaganda: The Marginal Treatment Effect of Foreign Media among Always-Takers," *Political Science Research and Methods*, 1–18. 69

Gupta, A., S. T. Howell, C. Yannelis, and A. Gupta (2024): "Owner Incentives and Performance in Healthcare: Private Equity Investment in Nursing Homes," *The Review of Financial Studies*, 37, 1029–1077.

Hahn, J., G. Kuersteiner, A. Santos, and W. Willigrod (2024): "Overidentification in Shift-Share Designs," . 45

Hahn, J., P. Todd, and W. Van der Klauuw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209. 41

Han, S. and H. Kaido (2024): "Set-Valued Control Functions," . 78

Han, S. and S. Yang (2024): "A Computational Approach to Identification of Treatment Effects for Policy Evaluation," *Journal of Econometrics*, 240, 105680. 69

Hansen, B. (2022a): *Probability and Statistics for Economists*, Princeton University Press. 59, 107

Hansen, B. E. (2022b): *Econometrics*, Princeton University Press. 45, 66

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422. 44

Heckman, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–694. 5, 10, 49

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

Heckman, J., N. Hohmann, J. Smith, and M. Khoo (2000): "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *The Quarterly Journal of Economics*, 115, 651–694. 38

Heckman, J. and E. Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return Is Correlated with Schooling," *The Journal of Human Resources*, 33, 974–987. 85

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*. 5, 10, 49

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. 5, 10

——— (2010): "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48, 356–98. 13

Heckman, J. J. and B. E. Honoré (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149. 10

Heckman, J. J., J. E. Humphries, and G. Veramendi (2016): "Dynamic Treatment Effects," *Journal of Econometrics*, 191, 276–292. 88

——— (2018): "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking," *Journal of Political Economy*, 126, S197–S246. 88

Heckman, J. J., R. J. Lalonde, and J. A. Smith (1999): "Chapter 31 The Economics and Econometrics of Active Labor Market Programs," Elsevier, vol. Volume 3, Part A, 1865–2097. 16

Heckman, J. J. and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86, 1–35. 12, 87, 88

Heckman, J. J. and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer, Cambridge University Press. 64, 87

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. 84

Heckman, J. J. and S. Urzua (2010): "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156, 27–37. 16, 99

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389–432. 31, 33, 80

——— (2008): "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case," *Annales d'Economie et de Statistique*, 91/92, 151–174. 88

Heckman, J. J. and E. Vytlacil (2001a): "Policy-Relevant Treatment Effects," *The American Economic Review*, 91, 107–111. 15, 75, 76

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 5, 15, 17, 31, 58, 75, 76

Heckman, J. J. and E. J. Vytlacil (1999): "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences of the United States of America*, 96, 4730–4734. 5, 58, 70

——— (2001b): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner and F. Pfeiffer, Heidelberg and Berlin: Physica. 92

——— (2007a): "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4779–4874. 8

——— (2007b): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. Volume 6, Part 2, 4875–5143. 57, 70, 80

Heiler, P. (2022): "Efficient Covariate Balancing for the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 40, 1569–1582. 52

Heinesen, E., C. Hvid, L. Kirkebøen, E. Leuven, and M. Mogstad (2022): "Instrumental Variables with Unordered Treatments: Theory and Evidence from Returns to Fields of Study," . 35, 37

Heinesen, E. and E. Stenholt Lange (2022): "Vocational versus General Upper Secondary Education and Earnings," *Journal of Human Resources*, 0221–11497R2.

Heldring, L., J. A. Robinson, and S. Vollmer (2022): "The Economic Effects of the English Parliamentary Enclosures," .

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000): "Assessing the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88. 52

Hojman, A. and F. Lopez Boo (2022): "Public Childcare Benefits Children and Mothers: Evidence from a Nationwide Experiment in a Developing Country," *Journal of Public Economics*, 212, 104686.

Hong, H. and D. Nekipelov (2010): "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1, 279–304. 51

Huber, M., L. Laffers, and G. Mellace (2015): "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations Under Endogeneity and Noncompliance," *Journal of Applied Econometrics*, 32, 56–79. 92

Huber, M. and G. Mellace (2014): "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints," *Review of Economics and Statistics*, 97, 398–411. 17

Hull, P. (2020): "Estimating Hospital Quality with Quasi-Experimental Data," SSRN Scholarly Paper ID 3118358, Social Science Research Network, Rochester, NY. 87

Humlum, A., J. Munch, and M. Rasmussen (2023): "What Works for the Unemployed? Evidence from Quasi-Random Caseworker Assignments," .

Humphries, J. E., J. S. Joensen, and G. F. Veramendi (2023a): "Complementarities in High School and College Investments," *SSRN Electronic Journal*. 88

Humphries, J. E., A. Ouss, K. Stavreva, M. T. Stevenson, and W. van Dijk (2023b): "Conviction, Incarceration, and Recidivism: Understanding the Revolving Door," . 37, 38, 88

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423. 16, 98, 99

——— (2015): "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50, 373–419. 51

——— (2022): "Causality in Econometrics: Choice vs Chance," *Econometrica*, 90, 2541–2566. 45

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 4, 5, 10, 11, 12, 20, 21, 23, 27, 30, 46

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 84

Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *The Review of Economic Studies*, 64, 555–574. 17

Ito, K., T. Ida, and M. Tanaka (2023): "Selection on Welfare Gains: Experimental Evidence from Electricity Plan Choice," *American Economic Review*, 113, 2937–2973. 14, 15, 16, 23, 61, 62, 63, 69, 74, 93

Joensen, J. S. and H. S. Nielsen (2016): "Mathematics and Gender: Heterogeneity in Causes and Consequences," *The Economic Journal*, 126, 1129–1163.

Johar, M. and S. Maruyama (2014): "DOES CORESIDENCE IMPROVE AN ELDERLY PARENT'S HEALTH?: DOES CORESIDENCE IMPROVE AN ELDERLY PATIENT'S HEALTH?" *Journal of Applied Econometrics*, 29, 965–983.

Kamat, V. (2021): "On the Identifying Content of Instrument Monotonicity," . 92

——— (2024): "Identifying the Effects of a Program Offer with an Application to Head Start," *Journal of Econometrics*, 240, 105679. 87

Kamat, V., S. Norris, and M. Pecenco (2024): "Conviction, Incarceration, and Policy Effects in the Criminal Justice System," . 37, 38, 80

Kane, T. J. and C. E. Rouse (1995): "Labor-Market Returns to Two- and Four-Year College," *The American Economic Review*, 85, 600–614. 13

Kang, J. D. Y. and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," 22, 523–539. 54

Kasahara, H., Y. Liang, and J. Rodrigue (2016): "Does Importing Intermediates Increase the Demand for Skilled Workers? Plant-level Evidence from Indonesia," *Journal of International Economics*, 102, 242–261.

Kaufmann, K. M. (2014): "Understanding the Income Gradient in College Attendance in Mexico: The Role of Heterogeneity in Expected Returns: Income Gradient in College Attendance," *Quantitative Economics*, 5, 583–630.

Kédagni, D. and I. Mourifié (2020): "Generalized Instrumental Inequalities: Testing the Instrumental Variable Independence Assumption," *Biometrika*, 107, 661–675. 17

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings, and Self-Selection," *The Quarterly Journal of Economics*, 131, 1057–1111. 35, 36, 37, 38

Kitagawa, T. (2009): "Identification Region of the Potential Outcome Distributions under Instrument Independence," *Cemmap working paper*. 92

——— (2015): "A Test for Instrument Validity," *Econometrica*, 83, 2043–2063. 17

——— (2021): "The Identification Region of the Potential Outcome Distributions under Instrument Independence," *Journal of Econometrics*, 225, 231–253. 92

Kline, P. and A. Santos (2013): "Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the U.S. Wage Structure: Sensitivity to Missing Data Assumptions," *Quantitative Economics*, 4, 231–267. 27

Kline, P. and C. R. Walters (2016): "Evaluating Public Programs with Close Substitutes: The Case of Head Start*," *The Quarterly Journal of Economics*, 131, 1795–1848. 19, 38, 86, 87

——— (2019): "On Heckits, LATE, and Numerical Equivalence," *Econometrica*, 87, 677–696. 66, 107

Kling, J. R. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876. 27

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity," . 44

Kowalski, A. E. (2023a): "Behaviour within a Clinical Trial and Implications for Mammography Guidelines," *The Review of Economic Studies*, 90, 432–462. 69

——— (2023b): "How to Examine External Validity within an Experiment," *Journal of Economics & Management Strategy*, 32, 491–509. 69

——— (2023c): "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform," *Review of Economics and Statistics*, 105, 646–664. 59, 65

Kowalski, Amanda. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," *NBER Working paper 22363*. 65

Kreider, B. and J. V. Pepper (2007): "Disability and Employment," *Journal of the American Statistical Association*, 102, 432–441. 91

Kreider, B., J. V. Pepper, C. Gundersen, and D. Jolliffe (2012): "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, 107, 958–975. 91

Lee, D. S. (2008): "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 142, 675–697. 5

Lee, K., E. Miguel, and C. Wolfram (2019): "Experimental Evidence on the Economics of Rural Electrification," *Journal of Political Economy*, 128, 1523–1565. 23

Lee, S. (2018): "A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs," *Journal of Business & Economic Statistics*, 36, 400–410. 46

Lee, S. and B. Salanié (2018): "Identifying Effects of Multivalued Treatments," *Econometrica*, 86, 1939–1963. 88

——— (2023): "Treatment Effects with Targeting Instruments," . 87

Leung, P. and C. O'Leary (2020): "Unemployment Insurance and Means-Tested Program Interactions: Evidence from Administrative Data," *American Economic Journal: Economic Policy*, 12, 159–192. 53

Lewis, H. G. (1974): "Comments on Selectivity Biases in Wage Comparisons," *Journal of Political Economy*, 82, 1145–1155. 10

Li, H., Y. Liu, X. Zhao, L. Zhang, and K. Yuan (2021a): "Estimating Effects of Cooperative Membership on Farmers' Safe Production Behaviors: Evidence from the Rice Sector in China," *Environmental Science and Pollution Research*, 28, 25400–25418.

Li, M., T. Jin, S. Liu, and S. Zhou (2021b): "The Cost of Clean Energy Transition in Rural China: Evidence Based on Marginal Treatment Effects," *Energy Economics*, 97, 105167.

Liu, S., T. Jin, M. Li, and S. Zhou (2022): "Fertility Policy Adjustments and Female Labor Supply: Estimation of Marginal Treatment Effect Using Chinese Census Data," *Journal of Human Resources*.

Loeser, J. (2023): "Modeling Selection into Unordered Treatments: An Equivalence Result," *Working Paper*. 88

Machado, C., A. M. Shaikh, and E. J. Vytlacil (2019): "Instrumental Variables and the Sign of the Average Treatment Effect," *Journal of Econometrics*. 92

MaCurdy, T., X. Chen, and H. Hong (2011): "Flexible Estimation of Treatment Effect Parameters," *American Economic Review*, 101, 544–551. 52

Maestas, N., K. J. Mullen, and A. Strand (2013): "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *The American Economic Review*, 103, 1797–1829. 13

Manda, J., A. D. Alene, A. H. Tufa, T. Abdoulaye, A. Y. Kamara, O. Olufajo, O. Boukar, and V. M. Manyong (2020): "Adoption and Ex-post Impacts of Improved Cowpea Varieties on Productivity and Net Returns in Nigeria," *Journal of Agricultural Economics*, 71, 165–183.

Manski, C. (1994): "The Selection Problem," in *Advances in Econometrics, Sixth World Congress*, vol. 1, 143–70. 89

Manski, C. F. (1989): "Anatomy of the Selection Problem," *The Journal of Human Resources*, 24, 343–360. 89

——— (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80, 319–323. 88, 89

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334. 92

——— (2003): *Partial Identification of Probability Distributions*, Springer. 93

——— (2007): "Partial Indentification of Counterfactual Choice Probabilities," *International Economic Review*, 48, 1393–1410. 12

Manski, C. F. and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010. 91

Marshall, J. (2016): "Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates," *Political Analysis*, 24, 157–171. 34

Martinez-Iriarte, J. and Y. Sun (2024): "Identification and Estimation of Unconditional Policy Effects of an Endogenous Binary Treatment: An Unconditional MTE Approach," . 61

Marx, B. M. and L. J. Turner (2019): "Student Loan Nudges: Experimental Evidence on Borrowing and Educational Attainment," *American Economic Journal: Economic Policy*, 11, 108–141. 53

Marx, P. (2024): "Sharp Bounds in the Latent Index Selection Model," *Journal of Econometrics*, 238, 105561. 69

Masten, M. A. and A. Poirier (2020): "Inference on Breakdown Frontiers," *Quantitative Economics*, 11, 41–111. 27

——— (2021): "Salvaging Falsified Instrumental Variable Models," *Econometrica*, 89, 1449–1469. 27

Masten, M. A. and A. Torgovitsky (2014): "Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model," *cemmap working paper 02/14*. 84

——— (2016): "Identification of Instrumental Variable Correlated Random Coefficients Models," *Review of Economics and Statistics*, 98, 1001–1005. 84, 85

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. 83

Mellon Bedi, S., C. Azzarri, B. Hundie Kotu, L. Kornher, and J. von Braun (2021): "Scaling-up Agricultural Technologies: Who Should Be Targeted?" *European Review of Agricultural Economics*, jbab054.

Miyaji, S. (2024a): "Instrumented Difference-in-Differences with Heterogeneous Treatment Effects," . 45

——— (2024b): "Two-Way Fixed Effects Instrumental Variable Regressions in Staggered DID-IV Designs," . 45

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

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

Moffitt, R. A. and M. V. Zahn (2019): "The Marginal Labor Supply Disincentives of Welfare: Evidence from Administrative Barriers to Participation," .

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters," *NBER Working Paper*. 69, 74

——— (2018): "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters," *Econometrica*, 86, 1589–1619. 59, 69, 73, 82

Mogstad, M. and A. Torgovitsky (2018): "Identification and Extrapolation of Causal Effects with Instrumental Variables," *Annual Review of Economics*, 10. 57

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–3698. 13, 31, 45, 77

——— (2024): "Policy Evaluation with Multiple Instrumental Variables," *Journal of Econometrics*, 105718. 77

Molinari, F. (2020): "Microeconometrics with Partial Identification," *arXiv:2004.11751 [econ]*. 70

Mountjoy, J. (2022): "Community Colleges and Upward Mobility," *American Economic Review*, 112, 2580–2630. 38, 88

Mourifié, I. and Y. Wan (2016): "Testing Local Average Treatment Effect Assumptions," *The Review of Economics and Statistics*, 99, 305–313. 17

Mueller-Smith, M. (2015): "THE CRIMINAL AND LABOR MARKET IMPACTS OF INCARCERATION," *Working Paper*, 59. 43

Navjeevan, M., R. Pinto, and A. Santos (2023): "Identification and Estimation in a Class of Potential Outcomes Models," . 87

Newey, W. and S. Stouli (2021): "Control Variables, Discrete Instruments, and Identification of Structural Functions," *Journal of Econometrics*, 222, 73–88. 84

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. 55

Noack, C. (2021): "Sensitivity of LATE Estimates to Violations of the Monotonicity Assumption," . 27

Nobel Committee (2021): "Scientific Background on the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 202," 48. 45

Norris, S., M. Pecenco, and J. Weaver (2021): "The Effects of Parental and Sibling Incarceration: Evidence from Ohio," *American Economic Review*, 111, 2926–2963. 43

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

Ogburn, E. L., A. Rotnitzky, and J. M. Robins (2015): "Doubly Robust Estimation of the Local Average Treatment Effect Curve," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 373–396. 54

Pearl, J. (2011): "Principal Stratification – a Goal or a Tool?" *The International Journal of Biostatistics*, 7, 1–13. 16

Pepper, J. V. (2000): "The Intergenerational Transmission of Welfare Receipt: A Nonparametric Bounds Analysis," *The Review of Economics and Statistics*, 82, 472–488. 91

Permutt, T. and J. R. Hebel (1989): "Simultaneous-Equation Estimation in a Clinical Trial of the Effect of Smoking on Birth Weight," *Biometrics*, 45, 619–622. 20

Phillips, G. D. A. and C. Hale (1977): "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review*, 18, 219–228. 44

Pinto, R. (2022): "Beyond Intention-to-Treat: Using the Incentives of Moving to Opportunity to Identify Neighborhood Effects," *Working Paper*. 87

Poirier, A. and T. Słoczyński (2024): "Quantifying the Internal Validity of Weighted Estimands," . 18

Possebom, V. (2023): "Crime and Mismeasured Punishment: Marginal Treatment Effect with Misclassification," *The Review of Economics and Statistics*, 1–42. 78

Priebe, J. (2020): "Quasi-Experimental Evidence for the Causal Link between Fertility and Subjective Well-Being," *Journal of Population Economics*, 33, 839–882.

Puhani, P. (2000): "The Heckman Correction for Sample Selection and Its Critique," *Journal of Economic Surveys*, 14, 53–68. 68

Ramsey, J. B. (1969): "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, 31, 350–371. 41, 55, 97

Rivera, R. (2023): "Release, Detail, or Surveil?" . 81

Robins, J. M. (1989): "The Analysis of Randomized and Non-Randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," *Health Service Research Methodology : A Focus on AIDS*, 113–159. 89

Robins, J. M. and S. Greenland (1992): "Identifiability and Exchangeability for Direct and Indirect Effects," *Epidemiology*, 3, 143. 12

——— (1996): "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association*, 91, 456–458. 16

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. 71

Rose, E. K. and Y. Shem-Tov (2021): "How Does Incarceration Affect Reoffending? Estimating the Dose-Response Function," *Journal of Political Economy*, 000–000. 69, 81, 82, 83

——— (2023): "On Recoding Ordered Treatments as Binary Indicators," . 34

Roy, A. D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146. 10

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, 66, 688–701. 8

Sargan, J. D. (1958): "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26, 393–415. 17

Sarr, M., M. Bezabih Ayele, M. E. Kimani, and R. Ruhinduka (2021): "Who Benefits from Climate-Friendly Agriculture? The Marginal Returns to a Rainfed System of Rice Intensification in Tanzania," *World Development*, 138, 105160.

Sasaki, Y. and T. Ura (2023): "Estimation and Inference for Policy Relevant Treatment Effects," *Journal of Econometrics*, 234, 394–450. 73

Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955. 92

Shea, J. and A. Torgovitsky (2023): "Ivmte: An R Package for Extrapolating Instrumental Variable Estimates Away From Compliers*," *Observational Studies*, 9, 1–42. 70, 73

Shurtz, I., A. Eizenberg, A. Alkalay, and A. Lahad (2022): "Physician Workload and Treatment Choice: The Case of Primary Care," *The RAND Journal of Economics*, 53, 763–791. 91

Siddique, Z. (2013): "Partially Identified Treatment Effects Under Imperfect Compliance: The Case of Domestic Violence," *Journal of the American Statistical Association*, 108, 504–513. 91

Sigstad, H. (2024a): "Marginal Treatment Effects and Monotonicity," . 77

——— (2024b): "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs," *SSRN Electronic Journal*. 29, 77

Singh, R. and L. Sun (2024): "Double Robustness for Complier Parameters and a Semi-Parametric Test for Complier Characteristics," *The Econometrics Journal*, 27, 1–20. 52

Słoczyński, T. (2020): "When Should We (Not) Interpret Linear IV Estimands as LATE?" . 42, 43, 48

——— (2022): "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economics and Statistics*, 104, 501–509. 48

Słoczyński, T., S. D. Uysal, and J. M. Wooldridge (2022): "Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment," . 54

——— (2024): "Abadie's Kappa and Weighting Estimators of the Local Average Treatment Effect," *Journal of Business & Economic Statistics*, 1–14. 52, 57

Słoczyński, T. and J. M. Wooldridge (2018): "A GENERAL DOUBLE ROBUSTNESS RESULT FOR ESTIMATING AVERAGE TREATMENT EFFECTS," *Econometric Theory*, 34, 112–133. 54

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. 8

Stevenson, M. T. (2018): "Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes," *The Journal of Law, Economics, and Organization*, 34, 511–542. 28

Sun, B. and Z. Tan (2022): "High-Dimensional Model-Assisted Inference for Local Average Treatment Effects With Instrumental Variables," *Journal of Business & Economic Statistics*, 40, 1732–1744. 52, 54

Sun, L. and S. Abraham (2021): "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 225, 175–199. 5, 45

Sun, Z. (2023): "Instrument Validity for Heterogeneous Causal Effects," *Journal of Econometrics*, 237, 105523. 17

Swanson, S. A. and M. A. Hernán (2014): "Think Globally, Act Globally: An Epidemiologist's Perspective on Instrumental Variable Estimation," *Statistical Science*, 29, 371–374. 16

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101, 1607–1618. 52, 54

Tebaldi, P., A. Torgovitsky, and H. Yang (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91, 107–146. 87

Theil, H. (1971): *Principles of Econometrics*, John Wiley & Sons. 7

Thornton, R. L. (2008): "The Demand for, and Impact of, Learning HIV Status," *American Economic Review*, 98, 1829–63. 31

Tommasi, D. and L. Zhang (2024): "Bounding Program Benefits When Participation Is Misreported," *Journal of Econometrics*, 238, 105556. 78

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83, 1185–1197. 84

——— (2017): "Minimum Distance from Independence Estimation of Nonseparable Instrumental Variables Models," *Journal of Econometrics*, 199, 35–48. 84

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge university press. 87

Ura, T. (2018): "Heterogeneous Treatment Effects with Mismeasured Endogenous Treatment," *Quantitative Economics*, 9, 1335–1370. 78

Ura, T. and L. Zhang (2024): "Policy Relevant Treatment Effects with Multidimensional Unobserved Heterogeneity," . 78

Uysal, S. D. (2011): "Three Essays on Doubly Robust Estimation Methods," . 52, 54

van 't Hoff, N., A. Lewbel, and G. Mellace (2024): "Limited Monotonicity and the Combined Compliers LATE," *Boston College Working Papers in Economics*. 77

Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, 33, 127–169. 64, 87

Vohra, V. and J. Goldin (2024): "Identifying the Cumulative Causal Effect of a Non-Binary Treatment from a Binary Instrument," . 80

Vuong, Q. and H. Xu (2017): "Counterfactual Mapping and Individual Treatment Effects in Nonseparable

Models with Binary Endogeneity," *Quantitative Economics*, 8, 589–610. 93

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 5, 11, 12, 21, 42, 79

——— (2006): "Ordered Discrete-Choice Selection Models and Local Average Treatment Effect Assumptions: Equivalence, Nonequivalence, and Representation Results," *The Review of Economics and Statistics*, 88, 578–581. 32, 79, 80, 85

Wald, A. (1940): "The Fitting of Straight Lines If Both Variables Are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300. 20

Walters, C. R. (2018): "The Demand for Effective Charter Schools," *Journal of Political Economy*, 126, 2179–2223. 88

Wang, W., T. Ida, and H. Shimada (2020): "Default Effect versus Active Decision: Evidence from a Field Experiment in Los Alamos," *European Economic Review*, 128, 103498.

Westphal, M., D. A. Kamhöfer, and H. Schmitz (2022): "Marginal College Wage Premiums Under Selection Into Employment," *The Economic Journal*, 132, 2231–2272.

Wilding, A., L. Munford, and M. Sutton (2023): "Estimating the Heterogeneous Health and Well-Being Returns to Social Participation," *Health Economics*, 32, 1921–1940.

Willis, R. J. and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87, S7–S36. 7, 10

Windmeijer, F. (2019): "Two-Stage Least Squares as Minimum Distance," *The Econometrics Journal*, 22, 1–9. 17

Wooldridge, J. M. (1997): "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters*, 56, 129–133. 85

——— (2003): "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters*, 79, 185–191. 85

——— (2008): "Instrumental Variables Estimation of the Average Treatment Effect in Correlated Random Coefficient Models," in *Modeling and Evaluating Treatment Effects in Econometrics*, ed. by D. Millimet, J. Smith, and E. Vytlacil, Else. 85

——— (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT press. 7, 45, 79

——— (2015): "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445. 64, 65, 87

Xie, H. (2024): "Efficient and Robust Estimation of the Generalized LATE Model," *Journal of Business & Economic Statistics*, 42, 1053–1065. 87

Yau, L. H. Y. and R. J. Little (2001): "Inference for the Complier-Average Causal Effect From Longitudinal Data Subject to Noncompliance and Missing Data, With Application to a Job Training Assessment for the Unemployed," *Journal of the American Statistical Association*, 96, 1232–1244. 52

Zeng, S., F. Li, and P. Ding (2020): "Is Being an Only Child Harmful to Psychological Health?: Evidence from an Instrumental Variable Analysis of China's One-Child Policy," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183, 1615–1635.