Don't (fully) exclude me, it's not necessary! Causal inference with semi-IVs

Christophe Bruneel-Zupanc*
Department of Economics, KU Leuven

February 17, 2025

Abstract

This paper proposes *semi-instrumental variables* (semi-IVs) as an alternative to instrumental variables (IVs) to identify the causal effect of a binary (or discrete) endogenous treatment. A semi-IV is a less restrictive form of instrument: it affects the selection into treatment but is excluded only from one, not necessarily both, potential outcomes. Having two semi-IVs, one excluded from the potential outcome under treatment and the other from the potential outcome under control, is sufficient to nonparametrically point identify local average treatment effect (LATE) and marginal treatment effect (MTE) parameters. In practice, semi-IVs provide a solution to the challenge of finding valid IVs because they are easier to find: many selection-specific shocks, policies, costs, or benefits are valid semi-IVs. As an application, I estimate the returns to working in the manufacturing sector using sector-specific semi-IVs.

Keywords: instrumental variable, treatment effect, exclusion restriction, identification, selection, policy evaluation.

*E-mail address: christophe.bruneel@gmail.com. I am especially grateful to Jad Beyhum, Geert Dhaene, and Alexandre Gaillard for their tremendous help and to Gregory Veramendi for his help with the data of an earlier version. I also thank Jaap Abbring, Otilia Boldea, Christian Bontemps, Laurens Cherchye, Edoardo Ciscato, Ben Deaner, Sebastian Fleitas, Ferre de Graeve, Olivier de Groote, Vincent Han, Xavier d'Haultfoeuille, Peter Hull, Toru Kitagawa, Tobias Klein, Denis Kojevnikov, Soonwoo Kwon, Louise Laage, Pascal Lavergne, Arthur Lewbel, Jan de Loecker, Iris Kesternich, Sebastian Maes, Thierry Magnac, Matt Masten, Arnaud Maurel, Costas Meghir, Geert Meesters, Bob Miller, Erwin Ooghe, Aureo de Paula, Krishna Pendakur, Christian Proebsting, Bram de Rock, Adam Rosen, Silvia Sarpietro, Sami Stouli, Alexandros Theloudis, Jo Van Biesebroeck, Frank Verboven, Frederic Vermeulen, Ao Wang, Martin Weidner, Frank Windmeijer, Nese Yildiz, Andrei Zeleneev, Mariana Zerpa, and participants in seminars and conferences in Antwerp, Duke, Leuven, Rotterdam, Bristol, Toulouse, Manchester, San Antonio, Tilburg, Oxford, Barcelona, and Thessaloniki for helpful discussions, suggestions, and comments. I acknowledge financial support from KU Leuven grant STG/21/040 and FWO grant G018725N.

1 Introduction

Identifying the causal effect of a treatment on an outcome is a central challenge in economics, particularly when the selection of treatment is endogenous. A common approach is to use instrumental variables (IVs), which must satisfy three conditions to be valid: they must affect treatment selection (relevance), be independent of unobserved factors (exogeneity), and have no direct effect on the potential outcomes (exclusion restriction). The IV approach is appealing because it can identify causal effects under relatively mild additional assumptions. For instance, when IVs influence treatment selection uniformly in the same direction for all individuals – a condition known as monotonicity (Imbens and Angrist, 1994) – they enable the nonparametric identification of average treatment effects for individuals whose treatment choice is influenced by the IVs. These include local average treatment effect (LATE) (Imbens and Angrist, 1994) and marginal treatment effect (MTE) (Heckman and Vytlacil, 2005) parameters, which are widely estimated in applied research. However, finding valid IVs that meet these conditions is challenging. In particular, the exclusion restriction is a strong and often controversial assumption, even for commonly used instruments, because it is difficult to find a variable that affects the selection without affecting any potential outcome.

Meanwhile, in many settings with a discrete treatment, there are numerous variables that, while not excluded from *all* potential outcomes, are credibly excluded from *some*. Think, for instance, of treatment-specific characteristics – such as costs, prices, benefits, amenities, or policies – that affect the outcomes of individuals only if they select that specific treatment but not if they select another one. I call such variables *semi-instrumental variables* (semi-IVs): variables relevant for treatment selection and excluded from some, but not necessarily all, potential outcomes. This raises a critical question: is causal inference possible with semi-IVs alone? Remarkably, yes, this paper shows that semi-IVs can serve as substitutes for IVs in addressing endogeneity, even in general models with heterogenous treatment effects.

When the treatment is binary, two semi-IVs – one excluded from the potential outcome under treatment and the other from the potential outcome under control – are sufficient to nonparametrically identify LATE and MTE parameters, just as an IV would. Crucially, this identification does not require any additional assumption beyond those in the standard IV framework (monotonicity of the selection), nor any restriction on how the semi-IVs affect the potential outcomes from which they are not excluded. The identification is *IV-like* and stems directly from the observable variations induced by the semi-IVs. By expanding the variation available for causal inference, semi-IVs may offer a solution when valid IVs are difficult to find. Since semi-IVs are often more readily available in applied settings, they significantly broaden researchers' toolkit and open up new possibilities for identifying causal effects.

To illustrate this, I present several examples of semi-IVs in occupation and location choice models, with various applications in labor, education, and industrial organization, among others (see Section 2.2). As the main running example, which also corresponds to the empirical application in Section 5, I focus on the identification of the earnings returns to working in the manufacturing sector (Heckman and Sedlacek, 1985, 1990). Workers select into sectors based on their unobservable, sector-specific skills, leading to endogenous selection. As a result, the comparison of earnings between manufacturing and nonmanufacturing workers is confounded. Ideally, we would like to find an IV to address the selection on unobservables, but finding one that influences the sector choice without also influencing the earnings is particularly challenging. Fortunately, we have access to many sector-specific characteristics, with observable variations across time and space, at the geographical market (e.g., state) level, which can be leveraged for identification. For instance, for each market, we observe historical measures of a sector's size, productivity, or growth, such as (lagged) shocks to the sector-specific size (e.g., GDP). These are promising semi-IV candidates. Indeed, they are likely to affect workers' sector choices: all else equal, the more attractive a sector, the more likely a worker is to select it. Such aggregate shocks are also independent of the workers' unobserved sector-specific skills. Unfortunately, they are not valid IVs because past shocks to the size of a sector may, in part, be passed through to the potential earnings of workers in that sector (e.g., due to underlying changes in productivity, costs, or economies of scale): the sector-specific semi-IV is *included* in the corresponding sector-specific potential outcome. Crucially, however, conditional on the size of the nonmanufacturing sector, (shocks to) the size of the manufacturing sector should not directly affect the potential earnings of nonmanufacturing workers, and vice versa. This partial exclusion is the key property that makes sector-specific characteristics – and, more generally, treatment-specific variables – valid semi-IVs. Once characteristics specific to the treated are controlled for, similar characteristics for the untreated do not affect the potential outcomes under treatment, and vice versa.

It is important to note what is *not* required for semi-IVs to be valid. First, the two semi-IVs may be correlated, as is often the case. Their validity only requires that, conditional on the included semi-IV, the other one is relevant and rightfully excluded. In a sense, any effect of the excluded semi-IV on a given potential outcome must be subsumed by the effect of the included one. Second, the exclusion restriction applies to individuals' (latent) potential outcomes, but each semi-IV may still indirectly influence the observable mean outcome of the group from which it is excluded via selection effects. For instance, holding other factors constant, an increase in the size of the manufacturing sector encourages more workers (the compliers) to select into manufacturing, which, in turn, alters the observed average earnings of nonmanufacturing workers solely through changes in the workforce composition.

In fact, the observable changes in the average nonmanufacturing earnings and the selection probability, both induced by changes in the excluded manufacturing sector size (as semi-IV), are precisely the variations that identify the mean nonmanufacturing potential earnings for the compliers who select into manufacturing jobs as a result of the change in the semi-IV. Similarly, shifts in the nonmanufacturing sector size identify the mean manufacturing potential earnings for the compliers resulting from these shifts. This mirrors the standard IV nonparametric identification arguments but applied separately to the outcomes of the treated and untreated subpopulations instead of directly to all outcomes.

Such a decomposition hints at why semi-IVs are sufficient to identify treatment effects. LATE and MTE parameters compare the potential outcomes under treatment (e.g., manufacturing) and under control (e.g., nonmanufacturing) for specific subpopulations of (marginal) compliers. An IV is excluded from both potential outcomes, so shifting it identifies the mean of both potential outcomes for the same set of compliers at once, and directly yields the treatment effect. While convenient, this is not necessary. Indeed, as previously described, two distinct semi-IVs can separately identify the means of the potential outcomes under treatment and control for specific subpopulations. Then, to reconstruct treatment effects, one needs to align the changes in each semi-IV that induce the same flow of compliers into treatment. This alignment is straightforward under the usual LATE monotonicity assumption. Indeed, as with multiple instruments (Carneiro et al., 2011), monotonicity ensures that the treatment probability summarizes the effect of all semi-IVs on selection into a single measure. At the margin, the marginal complier induced into treatment by an infinitesimal increase in the manufacturing sector size is the same as the one induced into treatment by an infinitesimal decrease in the nonmanufacturing sector size. More generally, any two semi-IV shifts causing the same observable shift in treatment probability induce the same set of compliers.

Overall, the primary trade-off of the semi-IV approach is that researchers must find two semi-IVs, rather than a single IV, to identify similar treatment effects.[1] In many cases, valid semi-IVs naturally mirror each other in the sense that they are alternative-specific versions of the same type of variable (e.g., sector-specific size). As a result, researchers often only need to identify a single type of semi-IV that can be adapted to all potential outcomes. The additional cost of implementing the semi-IV approach is, therefore, relatively modest, especially in comparison to its key advantage: relaxing the stringent full exclusion restriction

[1]The two semi-IVs also need to be capable of inducing similar flows of compliers. This is always possible at the margin if both semi-IVs are continuous and relevant. So, to identify MTE parameters, there is no extra requirement on the semi-IVs compared to standard IVs, which are also required to be continuously distributed. However, LATE parameters can be identified using only discrete (or even binary) IVs. Exactly matching compliers with two discrete semi-IVs (and thus a limited support of the observable treatment probabilities) is theoretically possible but may be complicated in practice. To avoid these concerns, having at least one continuous semi-IV is recommended. This is a mild additional requirement.

of IVs. This makes semi-IVs a convenient alternative to IVs for causal inference.

Building on constructive identification arguments, this paper develops both nonparametric and semi-parametric methods for estimating MTEs (and thus LATEs) with semi-IVs. In particular, the semi-parametric method adapts the local IV MTE estimation (Heckman and Vytlacil, 1999; Carneiro et al., 2011; Andresen, 2018) to the semi-IV framework. The estimator is fast, easy-to-implement, and performs well, as illustrated by simulations in Online Appendix E. A user-friendly implementation is available in the companion R package `semiIVreg` (Bruneel-Zupanc, 2024).[2]

To illustrate the method, I estimate the (marginal) earnings returns to working in the manufacturing sector in the US and their evolution from 1999 to 2018 using semi-IVs. Following the previous discussion, I use one-year lagged sector-specific GDPs, taken at the local market (state) level, as semi-IVs. The analysis focuses on young white men with only high school education, the demographic group most affected by the decline of the manufacturing sector over the period considered (Pierce and Schott, 2016). I find significant heterogeneity in returns: in terms of potential earnings, some workers would earn twice as much in manufacturing jobs, while others would earn 50% more in nonmanufacturing jobs. Moreover, I find that the returns to working in manufacturing declined over the 20-year period, closely mirroring the decline in manufacturing employment. This occurs despite an increasing observable earnings gap in favor of manufacturing over nonmanufacturing workers (as estimated by naive OLS), highlighting the importance of controlling for the endogenous change in sector composition. Perhaps most strikingly, I find that the potential earnings in both sectors declined over the period. This decline highlights a marked deterioration in the economic prospects of young, less-educated white males, regardless of their choice of sector.

The main focus of this paper is on the effect of a binary treatment. This case provides the most transparent intuition for why semi-IVs can effectively substitute for IVs. However, the semi-IV approach naturally extends to the more general case of a discrete treatment with $J > 2$ distinct alternatives. In this setting, having one alternative-specific semi-IV – which must be excluded from all but one potential outcome – for each potential outcome (giving a total of J distinct semi-IVs) is sufficient to nonparametrically identify discrete treatment effects, such as the effect of one treatment relative to the best alternative (Heckman et al., 2006; Heckman and Vytlacil, 2007a,b) or margin-specific MTEs (Mountjoy, 2022). Notably, identifying these treatment effects with standard IV methods typically requires $J-1$ special "alternative-specific" IVs, which must affect the latent utility of only one alternative but must be excluded from all potential outcomes, unlike semi-IVs. Thus, semi-IVs provide a

[2]The package is available for download on Github. See the associated vignette at https://cbruneelzupanc.github.io/semiIVreg/ for user guidelines and replicable simulated examples.

natural generalization by conveniently allowing the alternative-specific instruments to also impact their corresponding potential outcomes, hence enlarging the available variation to identify discrete treatment effects. This gain comes at the modest cost of requiring J semi-IVs instead of $J-1$ IVs. The formal extension of the discrete treatment identification results with IVs to semi-IVs is provided in Online Appendix F. The semi-IV approach does not extend to continuous treatments as it would require infinitely many semi-IVs.

Related Literature. Exploiting targeted exclusion restrictions from certain potential outcomes to achieve identification has been used in part of the literature on Roy models (Heckman and Sedlacek, 1985, 1990; Heckman and Honoré, 1990; Heckman, 1990; Heckman and Vytlacil, 2007a,b; Bayer et al., 2011; French and Taber, 2011; D'Haultfœuille and Maurel, 2013). For instance, the sector-specific variables in Heckman and Sedlacek (1985, 1990) and D'Haultfœuille and Maurel (2013) can be interpreted as semi-IVs, excluded from the potential outcomes of workers selecting into other sectors. Roy models can be *parametrically* identified using distributional assumptions on the unobserved sector skills, possibly combined with sector-specific variables (Heckman and Honoré, 1990).

However, *nonparametric* identification results using only sector-specific variables (semi-IVs) are scarce and limited to models with specific selection rules. For *Roy models* – where the sector selection depends solely on the difference between the potential earnings, i.e., where workers select the sector giving them the highest potential earnings – Heckman and Honoré (1990) and French and Taber (2011) show that nonparametric identification of the treatment effects can be obtained with sector-specific variables if these variables have an additively separable effect with respect to the unobservables affecting the outcomes. The identification exploits the Roy model structure, which implies that the semi-IVs affect the first stage solely through their effect on the potential outcomes. Thus, by observing changes in sector choice probabilities induced by variations in semi-IVs, one can infer their effects on the corresponding outcomes and subsequently identify certain treatment effects. However, this reasoning is specific to Roy models.[3] For *extended Roy models* – where the selection rule is a function of the differences between the potential earnings and a cost depending only on observable covariates, with no additional unobserved shocks beyond those affecting earnings – D'Haultfœuille and Maurel (2013) show that nonparametric identification of the distribution of the treatment effects is also possible if the sector-specific variables have an additively separable effect on the outcomes.[4] Their identification results are more general

[3]Under the Roy model specification, the full model (i.e., the joint distribution of the potential outcomes) can be identified, but its identification relies on identification-at-infinity arguments, requiring sector-specific variables (semi-IVs) to shift the propensity score to zero or one, which is rarely satisfied in practice.

[4]In both, Roy and extended Roy models, additive separability in the outcome translates to the selection

than those in Roy models but still rely on the extended Roy model structure, in particular, the known link between the residuals in the outcome and the selection equations.

In contrast, for the more widely applied *generalized Roy models* – where the selection rule is more general and includes additional unobserved shocks (e.g., individual unobserved preferences) that may correlate with those affecting earnings – nonparametric identification of treatment effects has so far relied on IVs combined with the LATE monotonicity assumption (Heckman and Vytlacil, 2005, 2007a,b; French and Taber, 2011).[5] Even with IVs, the average treatment effect (ATE) for the entire population is generally not identified unless there exists a subpopulation for whom the probability of treatment is zero or one (identification-at-infinity). Fortunately, Imbens and Angrist (1994) showed that, even without this hard-to-satisfy identification-at-infinity requirement, an IV can still identify treatment effects of interest for subpopulations of compliers induced into treatment by varying the IV: the LATEs. Heckman and Vytlacil (2005) further extended this result and showed that marginal variations of continuous IVs identify the MTEs for marginal compliers.

This paper shows that, under the same monotonicity assumption as in Imbens and Angrist (1994), semi-IVs can also nonparametrically identify treatment effects for subpopulations of compliers induced into treatment by varying the semi-IVs (LATEs and MTEs). The semi-IV identification argument parallels that for an IV, but applied separately to each potential outcome. It does not rely on a specific selection rule (apart from monotonicity), in contrast to existing identification results in Roy and extended Roy models. Thus, semi-IVs provide an alternative to IVs for identifying LATE and MTE parameters in the most general models (i.e., generalized Roy models). This is particularly advantageous in empirical applications, where we want to avoid structural assumptions and where semi-IVs – such as occupation or location-specific characteristics – are often readily available but have so far only been used as control variables, neglecting their identification power.

This paper shows how semi-IVs can substitute for IVs in the LATE framework (Imbens and Angrist, 1994; Angrist et al., 1996), i.e., under the LATE monotonicity assumption. A companion paper (Bruneel-Zupanc and Beyhum, 2025) shows that semi-IVs can also substitute for IVs in the IV-quantile regression (IVQR) framework (Chernozhukov and Hansen, 2005) to identify quantile treatment effects (QTE) under the assumption of rank invariance on the potential outcomes instead of LATE monotonicity. The IVQR and LATE frameworks encompass most models used in applied research. Since semi-IVs achieve similar identifica-

equation, yielding weak separability in the first stage and imposing a stronger form of LATE monotonicity.

[5]The "LATE model" (Imbens and Angrist, 1994) is, implicitly, a generalized Roy model, in the sense that there is no restriction on the potential outcomes (allowing for heterogenous treatment effects) and no restriction on the selection, apart from monotonicity (see Vytlacil, 2002). See Heckman and Vytlacil (2007a) for a clear-cut distinction between Roy, extended Roy, and generalized Roy models.

tion results in both frameworks, they can substitute for IVs in many empirical applications.[6]

More broadly, this paper contributes to the literature on identifying treatment effects by relaxing key IV assumptions, particularly the exclusion restriction. A more detailed discussion of this literature is in Online Appendix A. For a recent survey, see Lewbel (2019).

Outline. Section 2 shows how the canonical program evaluation problem with a binary treatment and IVs (Angrist et al., 1996) adapts to semi-IVs. It outlines the theoretical properties of semi-IVs and includes detailed examples. Section 3 establishes the nonparametric identification of treatment effects (MTEs and LATEs) using semi-IVs. Section 4 describes the nonparametric estimation of these treatment effects, and shows how local IV estimation of MTEs can be extended to semi-IVs, offering a straightforward semi-parametric procedure for estimating MTEs. Finally, Section 5 applies this procedure to estimate the returns to working in the manufacturing sector in the US.

2 Framework

First, this section shows how the canonical program evaluation problem of identifying the (possibly heterogenous) effect of an endogenous binary treatment on an outcome using an IV (Imbens and Angrist, 1994; Angrist et al., 1996; Heckman and Vytlacil, 2005) naturally extends to semi-IVs. Then, I discuss examples of semi-IVs in a variety of applications.

2.1 The semi-IV model

Potential outcomes and endogenous treatment. We want to evaluate the effect of a binary treatment $D \in \mathcal{D} = \{0, 1\}$ on a scalar, real-valued outcome Y. Let Y_d be the latent potential outcome under treatment state d, for $d \in \mathcal{D}$. The researcher only observes the realized outcome, Y, corresponding to the potential outcome of the selected alternative D. That is,

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

and the potential outcome corresponding to the non-selected treatment is latent. The challenge faced by economists is that D may be endogenous with respect to the potential outcomes Y_0 and Y_1, so we cannot simply compare the observable distributions of Y given $D = 1$ and $D = 0$ in order to identify the causal effect of D on Y. It is well known that a valid IV, combined with a monotonicity assumption, allows us to nonparametrically recover some

[6] This paper focuses on nonparametric *point* identification results. Extensions of the IV nonparametric *set* identification results (e.g., Manski and Pepper, 2000; Chesher, 2010; Chesher and Rosen, 2017) to semi-IVs are left for future research.

causal effects of interest: the LATE (Imbens and Angrist, 1994) and the MTE (Heckman and Vytlacil, 1999, 2005). I will show that valid semi-IVs allow us to recover these parameters as well. The intuition behind why semi-IVs can substitute for IVs can be gleaned from Equation (1). The IV approach requires a variable that is fully excluded from Y. In the semi-IV approach, it suffices to find two variables (the semi-IVs), one excluded from Y_1 to serve as an instrument for the effect of D on $DY = DY_1$, and another one excluded from Y_0 to serve as an instrument for the effect of D on $(1-D)Y = (1-D)Y_0$.

Semi-IVs. Let Z_0 and Z_1 be the two (sets of) complementary semi-IVs, excluded from Y_1 and Y_0, respectively, with supports $\mathcal{Z}_0$ and $\mathcal{Z}_1$. Both semi-IVs are observed regardless of the selected treatment. Let $Z = (Z_0, Z_1) \in \mathcal{Z}$ be the set of all semi-IVs, with support $\mathcal{Z}$. Following Imbens and Angrist (1994), denote by $D(z)$ the "potential treatment" choice if the semi-IVs are exogenously set to $Z = z$. The observed treatment is thus $D = D(Z)$.[7] The validity of the semi-IVs holds conditional on a given set of control variables X, with support $\mathcal{X}$, which have no identifying power (X may be neither excluded, exogenous, nor relevant).

Assumption 1 (Valid semi-IVs)**.** (Z_0, Z_1) *satisfies the following conditions:*

A1.1 (Exogeneity and Exclusion) For all $z \in \mathcal{Z}$*, (i)* $D(z) \perp\!\!\!\perp Z|X$*, and (ii) for* $d = 0, 1$*,* $(Y_d, D(z)) \perp\!\!\!\perp Z_{1-d}|(Z_d, X)$*, where* $\perp\!\!\!\perp$ *denotes conditional independence.*

A1.2 (Relevance) For $d = 0, 1$*, the propensity score function,* $P(z_0, z_1, x) = \mathbb{E}[D|Z_0 = z_0, Z_1 = z_1, X = x]$ *is a nontrivial function of* z_d *for all* $(z_{1-d}, x) \in \mathcal{Z}_{1-d} \times \mathcal{X}$*.*

By definition, (Z_0, Z_1) is a set of valid complementary semi-IVs if Assumption 1 holds. A valid IV, $\tilde{Z}$, would be relevant for the selection into treatment and satisfy the stronger conditional independence assumption $(Y_0, Y_1, D(\tilde{z})) \perp\!\!\!\perp \tilde{Z}|X$ (Imbens and Angrist, 1994). In contrast, each semi-IV is relevant (A1.2), but excluded from only one of the two potential outcomes, not both (A1.1). Hence, we say that an IV is *fully* excluded while a semi-IV is only *partially* excluded from the outcome. Conditional on Z_1 and X, shifting Z_0 enables shifts in the selection probability without directly affecting Y_1: Z_0 acts as an IV for the effect of D on $DY = DY_1$. Yet, Z_0 is not excluded from Y_0 and may affect it in an unrestricted manner, like any covariate X. The reverse holds for Z_1. Together, Z_0 and Z_1 provide an exclusion restriction for each potential outcome, making them *complementary*.

In the remainder of the paper, I refer to the conditional independence Assumption A1.1(ii) as the *(partial) exclusion restriction*, and call Z_{1-d} *excluded* from Y_d given (Z_d, X). This

[7] The stable unit treatment value assumption (SUTVA) is implicitly made, meaning each individual's potential outcomes and treatment are unaffected by the others' treatment status or semi-IVs.

terminology is a simplification, as, similar to IVs (Angrist et al., 1996), satisfying $Y_d \perp\!\!\!\perp Z_{1-d}|(Z_d, X)$ actually requires two conditions: exclusion *and* exogeneity. To clarify the distinction between the two, denote by $Y_d(z, x)$ the potential outcome when D is exogenously set to d, Z to z, and X to x. Thus, the potential outcomes defined earlier are $Y_d = Y_d(Z, X)$, and the observed outcome is $Y = Y_D = Y_D(Z, X)$. Then, the exclusion restriction is that, for all $x \in \mathcal{X}$, for all (z_0, z_1), (z_0, z_1'), and $(z_0', z_1) \in \mathcal{Z}$, and for each $d = 0, 1$,

$$Y_d(z_d, z_{1-d}, x) = Y_d(z_d, z_{1-d}', x) =: Y_d(z_d, x).$$

That is, there is no direct effect on Y_d of exogenously switching Z_{1-d} from z_{1-d} to z_{1-d}', once we already control for $Z_d = z_d$ and $X = x$. Furthermore, exogeneity requires that Z_{1-d} be independent of the potential outcomes, i.e., $Y_d(z_d, x) \perp\!\!\!\perp Z_{1-d}|(Z_d, X)$.[8] Notice that $Y_d(z_d, x)$ is not necessarily independent of Z_d nor X. This is because Z_d and X may be correlated with the unobservables affecting $Y_d(z_d, x)$. The distinction between exclusion and exogeneity may be clearer when expressing the potential outcomes as latent variables in a generalized Roy model. The reformulation of the semi-IV model as a generalized Roy model, analogous to the IV framework of Heckman and Vytlacil (2005), is detailed in Online Appendix B.

Selection into treatment (monotonicity). As in the literature on IVs, the identification of causal effects also requires a restriction on how the semi-IVs affect the selection into treatment. I impose on the semi-IVs the same monotonicity/no-defier condition that Imbens and Angrist (1994) impose on IVs.[9,10]

[8] A stronger version of exogeneity in A1.1 is to assume that $(Y_0(z_0, x), Y_1(z_1, x), D(z)) \perp\!\!\!\perp (Z_0, Z_1)|X$ for all $(z_0, z_1) \in \mathcal{Z}$ and $x \in \mathcal{X}$. This is still very general and only restricts how the semi-IVs affect their corresponding potential outcomes: through a direct effect rather than via their correlation with unobservables affecting the potential outcomes (see Online Appendix B). This assumption enables a clearer causal interpretation of the effect of exogenous changes in semi-IVs on outcomes but is stronger than necessary for identification.

[9] This is exactly the same monotonicity assumption as Imbens and Angrist (1994) but imposed on multiple semi-IVs. From Mogstad et al. (2021), we know that Imbens and Angrist (1994)'s monotonicity assumption is stronger when applied to multiple IVs. Semi-IVs are not necessarily comparable to multiple unrelated IVs, however. Indeed, in most examples, the semi-IVs are alternative-specific versions of similar variables (e.g., sector-specific GDP). Thus, it is often the comparison between two alternatives (e.g., the difference, the ratio, or some other utility comparison) that affects selection, rather than two unrelated semi-IVs. Therefore, the monotonicity assumption can be considered as an assumption imposed on the comparison between the semi-IVs rather than on each of the two semi-IVs separately. This implies that individuals uniformly value the comparison of the semi-IVs, which is a reasonable assumption in many discrete choice models for selection (as in Online Appendix F).

[10] It is possible to relax the monotonicity to a weaker (unordered) *partial monotonicity condition* (as in Mogstad et al., 2021; Mountjoy, 2022) applied to each semi-IV separately (or to the vector monotonicity of Goff, 2024). Then, the mean potential outcomes/marginal treatment responses would still be identified but possibly for different sets of compliers for each semi-IV. Additional manipulations would be required to align these sets of compliers and obtain the marginal treatment effects. For instance, point identification could be achieved using a "comparable compliers" assumption (Mountjoy, 2022); see Online Appendix F.

Assumption 2 (Monotonicity)**.** *Conditional on X, for all $z, z' \in \mathcal{Z}$, either $D_i(z) \geq D_i(z')$ for all individuals i, or $D_i(z) \leq D_i(z')$ for all individuals i.*

Assumption 2 states that, conditional on X, an exogenous change in the semi-IVs Z from z to z' either weakly increases or weakly decreases selection into treatment for all individuals. This rules out non-uniform effects of Z on treatment choice: if some individuals opt into treatment when Z changes from z to z' ($D_i(z') = 1 > D_i(z) = 0$ for some i), then no one opts out of the treatment because of this change ($D_i(z') = 0 < D_i(z) = 1$ for no i). Conversely, if some opt out, no one opts in. That is, there are only compliers and no defiers.

Equivalent selection model with latent resistance to treatment. For IVs, the monotonicity condition is equivalent to the existence of a weakly separable rule for the selection into treatment (Vytlacil, 2002). The same holds for semi-IVs: if the semi-IVs are valid, then the monotonicity assumption is equivalent to assuming that the treatment is determined by a weakly separable rule of the form

$$D = \mathbb{1}\{P(Z, X) - V \geq 0\} \tag{2}$$

for some function P and a random variable V that is continuously distributed given X.

Under the selection rule (2), Assumption A1.1 can be rewritten as follows.

Assumption A1.1'. *(i) $V \perp\!\!\!\perp Z|X$, and (ii) for $d = 0, 1$, $(Y_d, V) \perp\!\!\!\perp Z_{1-d}|(Z_d, X)$.*

As is standard in the IV literature, since V is continuously distributed given $X = x$ for all x, one may normalize $V|X \sim \text{Unif}(0, 1)$ without loss of generality.[11] Given the independence of the semi-IVs and V (Assumption A1.1'), it follows that $V|Z = z, X = x \sim \text{Unif}(0, 1)$ for all $(z, x) \in \mathcal{Z} \times \mathcal{X}$. Under this normalization, the function $P(z, x)$ is the propensity score function, i.e., $P(z, x) \equiv \mathbb{E}[D|Z = z, X = x]$, and $P = P(Z, X)$ is the propensity score random variable with support $\mathcal{P}$.

The normalized V can be understood as the latent *resistance to treatment* (Cornelissen et al., 2016), ranked from the lowest ($V = 0$) to the highest ($V = 1$). Individuals with a low V exhibit minimal resistance to treatment, opting in even when the propensity score P is low, meaning they require fewer incentives to select into treatment. Conversely, those with a high V have a strong resistance to treatment and only opt in if P is high.

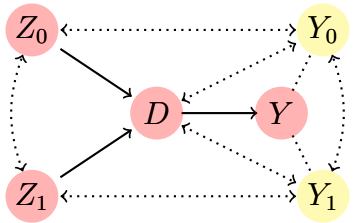
[11] This normalization is standard and innocuous given the model assumptions. Indeed, if the selection rule is $D = \mathbb{1}\{\nu(Z, X) - \eta \geq 0\}$ where η is a general continuous random variable, the model can always be reparametrized with $P(Z, X) = F_{\eta|X}(\nu(Z, X))$ and $V = F_{\eta|X}(\eta)$, ensuring $V|X \sim \text{Unif}(0, 1)$, which corresponds to Equation (2) under this normalization.

Regularity conditions. Following Heckman and Vytlacil (2005), identifying treatment effects also requires two additional mild regularity conditions: (i) $\mathbb{E}|Y_1|$ and $\mathbb{E}|Y_0|$ are finite (ensuring that the treatment parameters are well defined), and (ii) $0 < \Pr(D = 1|X = x) < 1$ for all $x \in \mathcal{X}$, ensuring the existence of a treated and an untreated population for every x.

Discussion. Assumptions 1 and 2, along with the regularity conditions, define the semi-IV model. The flexibility of this model is best understood by considering what is *not* imposed (Mogstad et al., 2018). First, similar to the IV model, I do not restrict the dependence between D and the potential outcomes. This allows for rich forms of observed and unobserved heterogeneity, where the causal effect of D on Y can vary with observed covariates, semi-IVs, and unobservables. Second, despite the monotonicity assumption, the selection rule remains general. The model does not specify the reasons behind the individuals' treatment choices, in contrast to Roy models, where $D = \mathbb{1}\{Y_1 > Y_0\}$, or extended Roy models, where $D = \mathbb{1}\{Y_1 - Y_0 + \nu(Z, X) > 0\}$ for some function ν of the observables Z and X. This flexibility accommodates various decision-making processes, including partial selection on gains $(Y_1 - Y_0)$. Third, the effect of the semi-IVs on the potential outcome from which they are not excluded is unrestricted and possibly heterogenous, varying with factors like latent resistance to treatment, V. Finally, the semi-IVs can (and generally will) be correlated. This is not a problem since the exclusion restriction holds conditional on the included semi-IV (and covariates). The only cost of the correlation between the semi-IVs is that it may weaken their relevance by limiting their joint variation.[12]

Causal graph representation. Figure 1 provides a causal graph representation of the model (Assumption 1) conditional on a specific $X = x$ (which is not displayed). The exclusion is visible by the absence of a link between Z_{1-d} and Y_d for each d.

Figure 1: Causal graph of the semi-IV model (binary D)



● = observed, ● = potential outcome Y_d is observed only if $D = d$. Solid arrows indicate causal effects. Dotted arrows indicate possible causal effects. Double-edged dotted arrows indicate a possible correlation.

[12] Assumption A1.2 rules out the degenerate case where Z_0 and Z_1 are perfectly correlated, as each semi-IV must be relevant for selection into treatment conditional on the other semi-IV (and covariates).

Properties of the model. Conditional on $X = x$, if an individual selects $D = 1$ when $Z = z$, one can conclude that $V \leq P(z, x)$, using (2) and the independence of V and Z given X. Conversely, $D = 0$ when $V > P(z, x)$. Therefore, for any $z' \neq z$, if $P(z, x) = P(z', x)$, then $D_i(z', x) = D_i(z, x)$ for all individuals. The propensity score serves as a *sufficient statistic* for the effect of the semi-IVs on selection: any two values of the semi-IVs yielding the same propensity score select the same individuals into treatment and control (because the semi-IVs are independent of V). This allows one to use multiple IVs (in our case, semi-IVs) and summarize their effect through the propensity score (see Carneiro et al., 2011, for an example of estimation of the MTE of college on earnings using multiple IVs).

Moreover, Assumption A1.1' (or A1.1) implies that

$$\mathbb{E}[Y_d|V, Z_d, Z_{1-d}, X] = \mathbb{E}[Y_d|V, Z_d, X] \quad \text{for } d = 0, 1. \tag{3}$$

Intuitively, conditional on X, Z_d, and V, a shift in the excluded semi-IV Z_{1-d} has no effect on the *latent* average potential outcome Y_d. However, this shift affects the propensity score, thereby altering the composition of the treated population and, as a result, the *observed* outcomes. Indeed, for the treated population, for all $z = (z_0, z_1) \in \mathcal{Z}$ and $x \in \mathcal{X}$, we have

$$\begin{aligned}
\mathbb{E}[DY|Z = z, X = x] &= \mathbb{E}[Y_1|D = 1, Z = z, X = x] \times \mathbb{E}[D|Z = z, X = x] \\
&= \mathbb{E}[Y_1|V \leq P(z, x), Z = z, X = x] \times P(z, x) \\
&= \mathbb{E}[Y_1|V \leq P(z, x), Z_1 = z_1, X = x] \times P(z, x),
\end{aligned}$$

where the first equality follows from the law of total expectation, the second from $D = 1$ being equivalent to $V \leq P(z, x)$, and the third from $(Y_1, V) \perp\!\!\!\perp Z_0|Z_1, X$. Similarly, for the untreated population, we have

$$\mathbb{E}[(1 - D)Y|Z = z, X = x] = \mathbb{E}[Y_0|V > P(z, x), Z_0 = z_0, X = x] \times (1 - P(z, x)).$$

This makes it clear that for each $d \in \{0, 1\}$ the only effect of Z_{1-d} on the conditional mean of $Y\mathbb{1}\{D = d\}$ given Z_d and X is through the propensity score $P(Z, X)$. Since $P(Z, X)$ is identified from data on D, Z, and X, we can condition directly on P instead of Z_{1-d}, yielding

$$\mathbb{E}[Y\mathbb{1}\{D = d\}|Z = z, X = x] = \mathbb{E}[Y\mathbb{1}\{D = d\}|P = P(z, x), Z_d = z_d, X = x], \tag{4}$$

for $d = 0, 1$ and for all $z = (z_0, z_1) \in \mathcal{Z}$ and $x \in \mathcal{X}$. Similar to an IV, Z_0 (resp. Z_1) affects DY (resp. $(1 - D)Y$) only through its effect on the selection propensity P. The excluded semi-IV (Z_{1-d}) can shift the selection probability and alter the composition of the treated and control groups while holding the included semi-IV (Z_d) and covariates (X) fixed. Therefore, conditional on Z_d and X, Z_{1-d} acts as an IV for the effect of D on $Y\mathbb{1}\{D = d\}$.

Comparison with IVs. Each semi-IV acts as an IV for the observable outcomes of either the treated or the untreated subpopulation. An IV is stronger, excluded from the outcome Y, meaning it is excluded from both DY and $(1-D)Y$. Since the conditions for a valid semi-IV are strictly weaker than those for an IV, a valid semi-IV is easier to find in practice. Indeed, as shown in the examples below, there are many selection-specific variables that are invalid IVs because they directly affect their alternative-specific potential outcome; but they are valid semi-IVs because they are credibly excluded from the other potential outcomes.

While a single valid semi-IV is strictly easier to find than a valid IV, it cannot serve as a substitute for a valid IV on its own. To replace an IV, at least two complementary semi-IVs are needed, with at least one excluded per potential outcome. Thus, there are two possible approaches to address endogeneity: finding one IV or two complementary semi-IVs. The full exclusion restriction is often the main obstacle to finding a valid IV. Moreover, as discussed in the examples below, finding two complementary semi-IVs is often not much harder than finding one, since treatment-specific versions of a similar variable can be used as semi-IVs. Therefore, complementary semi-IVs should often be more accessible than an IV.

In addition, and perhaps surprisingly, the possible non-exclusion of the semi-IVs from some potential outcomes has another upside: if individuals select into treatment partly based on their gains, $Y_1 - Y_0$, a variable affecting the potential outcomes should be relevant for the selection by construction. This should make it easier to find "non-weak" semi-IVs.

2.2 Examples and discussion

Example 1 (Occupation choice). Consider the important question of identifying the effects of occupation choice (D) on earnings (Y). The question relates to decisions to work in the manufacturing sector (Heckman and Sedlacek, 1985, 1990), agriculture (Lagakos and Waugh, 2013), or a white-collar, blue-collar, or military occupation (Keane and Wolpin, 1997). Similar questions arise about joining a union (Lee, 1978), attending college (Angrist and Krueger, 1991; Card, 1995, 2001; Carneiro et al., 2011), or selecting a specific college major (Arcidiacono, 2004; Arcidiacono et al., 2020). One needs to address endogeneity due to the self-selection of individuals into occupations based on their unobserved occupation-specific abilities. To do so, occupation-specific characteristics such as the occupation-specific wage rate, unemployment rate, task content, firm size, revenue, productivity, profits, or subsidies can serve as valid semi-IVs. More precisely, identification can be achieved by leveraging observable variations of these semi-IVs across time and geographic markets, and the corresponding variations in observed outcomes (Y) and occupation choices (D).

Returns to working in manufacturing. The main application in this paper (Section 5) focuses on the effect of the manufacturing ($D = 1$) versus the nonmanufacturing ($D = 0$)

sector choice on earnings (Y). I use the one-year lagged manufacturing GDP (Z_1) and non-manufacturing GDP (Z_0) of the market (state of residence of the worker) as semi-IVs. First, these are likely to be relevant variables for sectoral choice because, all else equal, a larger or more active sector is more attractive to new (young) workers. Second, the aggregate sector-specific GDPs are likely to be exogenous with respect to the workers' unobserved sector-specific skills and their unobserved resistance to treatment. Exogeneity requires that, aside from differences in the semi-IVs, the various markets represent otherwise comparable environments. To ensure this, I control for time trends and pre-existing differences across markets by including year and market (state) fixed effects in X. This approach addresses concerns about the self-selection of high-skilled workers into more favorable markets and accounts for the global evolution of worker skills. Consequently, the identifying variation consists of deviations of sector-specific GDPs from these fixed effects, which I will occasionally refer to as "shocks to" the variables.[13] Third, (shocks to) the sector GDPs are not valid IVs because they may influence the potential earnings of workers in that sector (e.g., through changes in productivity, costs, or economies of scale): the sector-specific semi-IV is *included* in the corresponding sector-specific potential earnings, i.e., Z_d influences Y_d. However, conditional on the GDP of the manufacturing sector, shocks to the GDP of the nonmanufacturing sector should not directly affect the potential earnings of manufacturing workers, and vice versa. In other words, Y_d is independent of Z_{1-d} conditional on Z_d and X. Once the included manufacturing characteristics are controlled for, similar characteristics of the other sectors do not affect the manufacturing potential earnings, and vice versa.

Note that, in a given market, the shocks to the manufacturing and nonmanufacturing GDP are likely to be correlated. As long as they are not perfectly correlated, so that some locations experience relatively larger shocks to their manufacturing sectors compared to their nonmanufacturing sectors than other locations, this is not a problem for identification.

As is standard with IVs, we rule out general equilibrium effects that could pose a threat to the partial exclusion restrictions. For instance, an increase in manufacturing GDP might alter worker composition, thereby affecting equilibrium wage rates in both sectors and indirectly influencing nonmanufacturing workers' earnings. To address this, we assume a partial equilibrium framework or a specific timing, so that wage rates are fixed or determined in the previous period. Fortunately, and unlike the IV case, general equilibrium concerns are largely mitigated here. Indeed, the exclusion of Z_{1-d} from Y_d holds conditional on Z_d, and Z_{1-d} can be correlated with Z_d. Thus, to a large extent, general equilibrium effects of Z_{1-d} on Y_d are captured by the effects of Z_{1-d} on Z_d. Controlling for Z_d absorbs most equilibrium

[13] Instead of fixed effects, one could via X control for the permanent levels of the semi-IVs in each market, as done by Carneiro et al. (2011) for some of their IVs, or use shocks to these variables directly as semi-IVs.

effects since sector-specific wage rates are largely determined by the sector's own characteristics (e.g., productivity, size). Thus, general equilibrium effects of the excluded semi-IV Z_{1-d} mostly impact Y_d through Z_d. This property is most salient when Z_0 and Z_1 are the sector-specific wage rates: controlling for the equilibrium wage rate in sector d, the wage rate in the other sector has no effect on Y_d. In fact, by specifying some additional structure, one could even identify general equilibrium models with semi-IVs. For example, Heckman and Sedlacek (1985) propose a general equilibrium framework where sector-specific total labor compensation (closely linked to GDP) could serve as semi-IVs. Alternatively, one could directly use sector-specific wage rates for specific tasks across markets as semi-IVs to identify general equilibrium models, provided such data are available.[14]

Example 2 (Location/migration choice). Consider the related problem of identifying the marginal earnings returns to endogenous location/migration choices of workers (Borjas, 1987; Dahl, 2002; Chiquiar and Hanson, 2005; Bayer et al., 2011; Abramitzky et al., 2012, 2014; Roca and Puga, 2017) or firms (Combes et al., 2012). Location-specific (variations in) expected earnings (Kennan and Walker, 2011), welfare benefits (Borjas, 1999; Kennan and Walker, 2010; Abramitzky, 2009; Agersnap et al., 2020), tax rates (Kleven et al., 2013, 2014, 2020; Akcigit et al., 2016; Muñoz, 2021), or more generally, any place-based policy (Neumark and Simpson, 2015) or measures of local amenities, may serve as valid semi-IVs. Again, these are (i) relevant for the choice to migrate, (ii) exogenous with respect to the individuals' unobserved characteristics, and (iii) the partial exclusion holds because, conditional on one's location-specific shocks, the shocks to the other locations are excluded from one's potential outcome. One can use temporal variation in these location-specific semi-IVs for identification.

Example 3 (Demand estimation). Semi-IVs can be used to estimate the demand for different products. For instance, consider the effect of car type – gasoline ($D = 0$) or diesel ($D = 1$) – on subsequent energy consumption, gasoline (Y_0) or diesel (Y_1). Endogeneity arises because individuals with different unobserved preferences (e.g., for higher mileage) select different cars. Estimating demand for gasoline/diesel cars is crucial for designing optimal energy transition policies, such as taxation or subsidies (Verboven, 2002; Grigolon et

[14]In this example, another threat to the validity of the semi-IVs could be monopsony power with strategic interactions between sectors, such as manufacturing firms adjusting wages based on the size of the nonmanufacturing sector. However, evidence suggests within-sector strategic responses are small, with Staiger et al. (2010) estimating an elasticity of only 0.128 in the case of VA (Veteran Affairs) and non-VA hospitals. Between-sector interactions are likely to be even smaller and negligible, as sectors do not explicitly compete. Given that I focus on large (state-level) markets with many firms in each sector, individual firms are relatively small and primarily compete within their own sector, further diminishing the risk of significant cross-sector strategic interaction. In short, I assume that firms are price takers without monopsony power.

al., 2018). Local market energy-specific prices serve as semi-IVs: they are relevant for the choice of type of car but do not affect the energy consumption of owners of the other car type (e.g., gas prices do not affect the energy consumption of diesel-car owners).[15] More broadly, energy-specific (shocks to) prices or subsidies can serve as semi-IVs to estimate the demand for energy based on appliance choice, such as solar panel adoption (De Groote and Verboven, 2019; Feger et al., 2022) or electric versus non-electric home appliance purchases (Dubin and McFadden, 1984).

Example 4 (Measuring effectiveness). Semi-IVs can also be used to measure the effectiveness (or value-added) of specific types of institutions or products, net of selection effects. For instance, in education, consider the estimation of the effect of charter schools ($D = 1$) (Angrist et al., 2010; Abdulkadiroğlu et al., 2011; Angrist et al., 2013; Abdulkadiroğlu et al., 2016; Walters, 2018; Dobbie and Fryer, 2020), boarding schools (Behaghel et al., 2017), catholic schools (Altonji et al., 2005), or private schools (Angrist et al., 2002; Carneiro et al., 2022) compared to public schools ($D = 0$), on academic achievements or future labor market outcomes (Y). In health, consider the estimation of the effectiveness of different insurance plans (e.g., private vs public) (Abaluck et al., 2021; Finkelstein et al., 2019a), type of hospitals (Gowrisankaran and Town, 1999; Geweke et al., 2003; Hull, 2018), or nursing homes (Einav et al., 2022) on health outcomes. In all these examples, variations in institution-specific prices, subsidies (Finkelstein et al., 2019b), size, funding, or measures of quality (number of teachers/practitioners per student/patient) can serve as semi-IVs.

Randomized experiments with semi-IVs. Even in randomized experiments, a randomly assigned intervention is not necessarily a valid IV. While randomization ensures exogeneity with respect to ex ante unobservables, the intervention itself may not be excluded from the outcome of interest (Angrist et al., 1996). For instance, conditional cash transfers, such as vouchers for private schooling in Colombia (Angrist et al., 2002), likely affect students' achievements directly through income effects for families who receive them if the student attends a private school. As a result, only the intention-to-treat effect of the policy can be estimated, not the treatment effect of attending private school. However, the voucher assignment has no impact on students attending public schools, who receive no money regardless of their assignment, making it a valid semi-IV (Z_1), excluded from outcomes associated with attending a public school (Y_0).[16] To identify the (local) treatment effect of private schools,

[15] To address the potential endogeneity of energy prices with respect to unobserved individual characteristics, one can use supply side shocks to the prices (e.g., cost shifters) or exogenous energy-specific subsidies/tax changes as semi-IVs instead of the prices. Additionally, one can control for market fixed effects.

[16] The same kind of reasoning applies to the Vietnam lottery draft (Angrist, 1990). Non-veterans who

the Colombian government could have designed the experiment differently by assigning two random monetary amounts: one for public schools (Z_0) and one for private schools (Z_1). Both could be nonzero since Z_0 is excluded from private school outcomes (Y_1) and Z_1 from public school outcomes (Y_0). By randomly assigning the monetary amounts, this design would facilitate manipulating treatment probabilities. Moreover, it would allow policymakers to study two effects simultaneously: the treatment effect of private school attendance and the policy/semi-IV effect of income increases on achievement. This approach also addresses ethical concerns by ensuring all families receive some support, not just a select few.

Discussion. With the exception of a few education-related examples, credible IVs are rarely available to address endogeneity in the settings and questions discussed above. Identification has so far mainly relied on distributional assumptions, functional form restrictions, structural models, or on assuming selection only on observables. Semi-IVs expand the researchers' toolkit by enabling IV-like identification with readily available variables, allowing them to credibly tackle a broader range of questions, even when credible IVs are hard to find.

3 Identification

Similar to IVs, semi-IVs identify average treatment effects for *compliers* induced into treatment by their shifts: the LATE and the MTE. More precisely, each semi-IV separately identifies mean potential outcomes under treatment and control at specific unobserved V, which are then combined to obtain treatment effects. In this section, I show the identification of the MTE, followed by the LATE. I also describe the identification of the *direct effects* of the semi-IVs on the potential outcomes from which they are not excluded.

3.1 Marginal treatment effects and responses

3.1.1 Treatment parameters definition

Marginal treatment effect (MTE). The marginal treatment effect (MTE) was introduced by Björklund and Moffitt (1987) and generalized to the nonparametric case by Heckman and Vytlacil (1999, 2005, 2007a,b). With semi-IVs, the MTE is naturally defined by

$$\text{MTE}(v, z_0, z_1, x) = \mathbb{E}[Y_1 - Y_0 | V = v, Z_0 = z_0, Z_1 = z_1, X = x]. \tag{5}$$

were drafted probably had to behave differently or face the consequence of their non-compliance (e.g., jail for conscientious objectors) compared to non-veterans who were not drafted. Thus, the draft is only excluded from the outcomes (subsequent earnings) of veterans, but not from the outcomes of non-veterans. Identifying the effect of being a veteran would require a complementary semi-IV excluded from the outcome of non-veterans, for example differential local tax rates/benefits for veterans at the time of enrollment.

The marginal treatment effect $\text{MTE}(v, z_0, z_1, x)$ is the average causal effect of D on Y for individuals with unobserved resistance to treatment $V = v$ and observed characteristics $X = x, Z_0 = z_0, Z_1 = z_1$. It is called "marginal" treatment effect because it is the effect of treatment for individuals at the margin of taking up treatment when $P = v$.

The main difference with the MTE using IVs is that the MTE also depends on the semi-IVs here. Since Z_0 may affect Y_0, and Z_1 may affect Y_1, we need to condition on (Z_0, Z_1), just as we condition on X. Fortunately, this conditioning does not impede identification.

Marginal treatment responses (MTR_d). Rather than identifying the MTE directly, we focus on the two *marginal treatment responses* (MTR) functions (see for example Mogstad et al., 2018; Mogstad and Torgovitsky, 2018) instead. With semi-IVs, these are defined by

$$m_0(v, z_0, x) = \mathbb{E}[Y_0 | V = v, Z_0 = z_0, X = x], \tag{MTR_0}$$

$$\text{and} \quad m_1(v, z_1, x) = \mathbb{E}[Y_1 | V = v, Z_1 = z_1, X = x]. \tag{MTR_1}$$

For each d, the MTR_d is the *mean potential outcome* Y_d for individuals with unobserved resistance to treatment $V = v$ and observed characteristics $Z_d = z_d$ and $X = x$. With the semi-IV exclusion property, m_d only depends on Z_d but not on Z_{1-d}. One may refer to m_1 and m_0 as the *mean potential outcomes under treatment* and *under control*, respectively.

The MTE is equal to the difference of MTRs,

$$\begin{aligned} \text{MTE}(v, z_0, z_1, x) = &\ \mathbb{E}[Y_1 | V = v, Z = (z_0, z_1), X = x] - \mathbb{E}[Y_0 | V = v, Z = (z_0, z_1), X = x] \\ = &\ m_1(v, z_1, x) - m_0(v, z_0, x). \end{aligned} \tag{6}$$

The second equality holds because we can remove conditioning on the excluded semi-IV using the mean independence equation (3), i.e., using the partial exclusion of the semi-IVs.

3.1.2 Identification

For notational convenience, omit X in the notation and proceed conditional on $X = x$.

In the data, we observe (Y, D, Z_0, Z_1). Since we observe (Z_0, Z_1) and D, we also identify the propensity score $P = P(Z_0, Z_1) = \Pr(D = 1 | Z_0, Z_1)$. Therefore, consider that we observe Y, D, Z_0, Z_1, and P. From there, the identification of each MTR with semi-IVs is similar to the identification of MTRs with a single IV (Mogstad and Torgovitsky, 2018). The only difference is that it is not the same source of variation (not the same semi-IV) that identifies both MTRs at once. Instead, m_1 is identified by using Z_0 as an instrument for the effect of D on DY given Z_1 and X, while m_0 is identified by using Z_1 as an instrument for the effect of D on $(1-D)Y$ given Z_0 and X. Z_0 and Z_1 play, in turn, the role of an instrument (for m_1 and m_0, respectively) and the role of a covariate (for m_0 and m_1, respectively) in the proof.

Once both MTRs have been separately identified, one can recover the MTE using Equation (6). Note that the monotonicity of the selection implicitly plays a key role: although each MTR is identified by shifting a different semi-IV, any change in a semi-IV can be mapped to its corresponding marginal compliers, characterized by V, through the associated propensity score. First, I show the identification of each MTR separately, then that of the MTE.

Identification of MTR$_1$. Let us proceed conditional on a specific $Z_1 = z_1 \in \mathcal{Z}_1$. To identify the mean potential outcome under treatment (m_1), use Z_0 as an IV excluded from Y_1. Y_1 is not observed for the untreated individuals, but recall that $Y = DY + (1-D)Y$ and focus on the observable treated population outcomes, $DY = DY_1$. Any shift in Z_0 affects the outcome of interest (DY here) only through its effect on the propensity score, $P = P(Z_0, z_1)$. This enables us to identify the MTR$_1$ by marginally shifting P at fixed $Z_1 = z_1$ using Z_0.

Assume that the semi-IV excluded from Y_1, Z_0, is continuous and relevant for all $Z_0 = z_0 \in \mathcal{Z}_0$, conditional on $Z_1 = z_1$, i.e., that $\partial P(z_0, z_1)/\partial z_0 \neq 0$. In this case, P is continuously distributed conditional on $Z_1 = z_1$. Using the selection equation (2), we have that $D = 1$ when $V \leq P$. So, for any $P = p$ in the interior of the support of P given $Z_1 = z_1$, we observe

$$\begin{aligned}
\mathbb{E}[DY|Z_1 = z_1, P = p] &= \mathbb{E}[Y_1|D = 1, Z_1 = z_1, P = p] \times p \\
&= \mathbb{E}[Y_1|V \leq p, Z_1 = z_1] \times p \\
&= \int_0^p \frac{\mathbb{E}[Y_1|V = v, Z_1 = z_1]}{p} dv \times p \quad \text{since } V \sim \mathcal{U}(0,1) \perp\!\!\!\perp Z_1 \\
&= \int_0^p m_1(v, z_1) dv. \qquad (7)
\end{aligned}$$

Thus, taking the derivative with respect to the propensity score, we have

$$m_1(v, z_1) = \frac{\partial}{\partial p} \mathbb{E}[DY|P = p, Z_1 = z_1]\Big|_{p=v}, \qquad (8)$$

where $\mathbb{E}[DY|P(z_1, Z_0) = p, Z_1 = z_1]$ is observed and so is its derivative if v is in the interior of the support of P given $Z_1 = z_1$. Therefore, for any z_1 and any v in the interior of the support of P given $Z_1 = z_1$, the MTR$_1$ given $Z_1 = z_1$ and $V = v$ is identified by (8).

The intuition behind why the derivative gives the MTR can be explained as follows. Fix $Z_1 = z_1$. When $P = p$, individuals with $V \leq p$ are treated, and those with $V = p$ are at the margin of the treatment, meaning they are indifferent between being treated or not. Now, consider a marginal increase in P from p to $p' = p + dp$ by shifting Z_0 while holding $Z_1 = z_1$ fixed. This shift moves the marginal individuals with $V = p$ into treatment. Their expected Y_1 is $\mathbb{E}[Y_1|V = p, Z_1 = z_1] = m_1(p, z_1)$. Since these individuals are now treated, the average observable outcome DY changes by the proportion of individuals entering treatment times

their average Y_1, i.e., $d(\mathbb{E}[DY|Z_1 = z_1, P = p]) = dp \times \mathbb{E}[Y_1|V = p, Z_1 = z_1]$. Dividing both sides by dp normalizes this change in DY and yields $m_1(p, z_1)$. Thus, the derivative of the average DY with respect to the propensity score identifies the mean potential outcome under treatment for individuals who are indifferent at $V = p$ (and $Z_1 = z_1$).

The key to identification is having a source of variation that can shift the propensity score at fixed Z_1 without directly affecting DY. This is precisely the role of Z_0, as visible in Equation (4). Implicitly, the derivative in Equation (8) is identified by the local IV regression of YD on D, instrumented by Z_0, at a specific value $\tilde{z}_0$ of Z_0 such that $P(\tilde{z}_0, z_1) = v$, i.e.,

$$\frac{\partial}{\partial p}\mathbb{E}[DY|P = p, Z_1 = z_1]\Big|_{p=v} = \frac{\partial \mathbb{E}[DY|P = P(z_0, z_1), Z_1 = z_1]/\partial z_0}{\partial P(z_0, z_1)/\partial z_0}\Big|_{z_0=\tilde{z}_0}.$$

In order to marginally shift P by shifting Z_0, a relevant and continuously distributed Z_0 is needed, similar to the continuous IV requirement for MTE identification with IVs.

Identification of MTR_0. Similarly, for any $Z_0 = z_0 \in \mathcal{Z}_0$, $m_0(v, z_0)$ is identified for any v in the interior of the support of the propensity score P given $Z_0 = z_0$ by[17]

$$m_0(v, z_0) = -\frac{\partial}{\partial p}\mathbb{E}[(1 - D)Y|P = p, Z_0 = z_0]\Big|_{p=v}. \tag{9}$$

Holding $Z_0 = z_0$ fixed, we shift the propensity score by shifting the other semi-IV, Z_1. Again, obtaining marginal shifts in P requires Z_1 to be continuously distributed and relevant.

Identification of the MTE. Given $Z_1 = z_1$, m_1 is identified for all v in the interior of the support of the propensity score given $Z_1 = z_1$. Given $Z_0 = z_0$, m_0 is identified for all v in the interior of the support of the propensity score given $Z_0 = z_0$. As a consequence, for any $(Z_0, Z_1) = (z_0, z_1) \in \mathcal{Z}$, the MTE is identified by Equation (6), as

$$MTE(v, z_0, z_1) = m_1(v, z_1) - m_0(v, z_0),$$

for all v in the interior of both, the support of P given $Z_1 = z_1$ and the support of P given $Z_0 = z_0$, i.e., the interior of the intersection of these two conditional supports.

Link with local IV (Heckman and Vytlacil, 1999)**.** With an IV, the MTE would be identified directly by the local IV regression of Y on D, i.e., using

$$\frac{\partial}{\partial p}\mathbb{E}\big[Y|P = p, Z_0 = z_0, Z_1 = z_1\big]\Big|_{p=v} = \text{MTE}(v, z_0, z_1). \tag{10}$$

The problem with semi-IVs is that it is not possible to shift P while fixing *both* Z_0 and Z_1.

[17]The minus sign comes from the fact that, $(1 - D) = 1$ when $V > P$, so $\mathbb{E}[(1 - D)Y|Z_0 = z_0, P = p] = \int_p^1 m_0(v, z_0)dv$, and the derivative with respect to p at $p = v$ is applied to the lower bound of the integral.

This is why we use Property (1) to split Y in two subsamples, and rewrite (10) as

$$\begin{aligned}
&\frac{\partial}{\partial p}\mathbb{E}\big[Y|P=p,Z_0=z_0,Z_1=z_1\big]\Big|_{p=v}\\
=\;& \frac{\partial}{\partial p}\mathbb{E}\big[YD|P=p,Z_0=z_0,Z_1=z_1\big]\Big|_{p=v}+\frac{\partial}{\partial p}\mathbb{E}\big[Y(1-D)|P=p,Z_0=z_0,Z_1=z_1\big]\Big|_{p=v}\\
=\;& \frac{\partial}{\partial p}\mathbb{E}\big[YD|P=p,Z_1=z_1\big]\Big|_{p=v}+\frac{\partial}{\partial p}\mathbb{E}\big[Y(1-D)|P=p,Z_0=z_0\big]\Big|_{p=v}\\
=\;& m_1(v,z_1)-m_0(v,z_0).
\end{aligned}$$

For each d, the conditioning on Z_{1-d} in the conditional average of $Y\mathbb{1}\{D=d\}$ was removed using Property (4), which follows from the exclusion of the semi-IVs. This is the key allowing us to identify m_0 and m_1 *separately* and still recover the MTE afterwards. This is the main difference from the standard local IV identification: with semi-IVs, the shifts in P used to identify m_1 and m_0 stem from different underlying semi-IV shifts. Yet monotonicity ensures they are comparable since they correspond to the same underlying $V=v$.

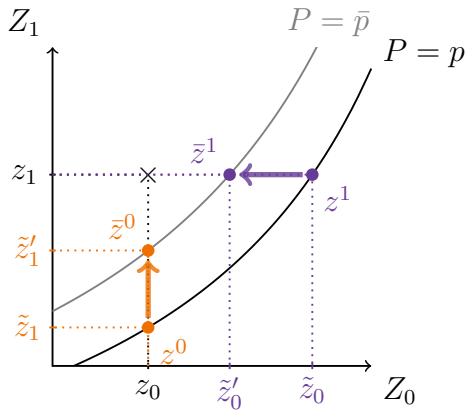


Figure 2: Semi-IV shifts identifying $\text{MTE}(p,z_0,z_1)$ (if $p'=p+dp$) and $\text{LATE}(p,p',z_0,z_1)$

Visual representation. Figure 2 illustrates the identification of the MTE (and LATE). It shows two distinct shifts in semi-IVs that separately identify m_1 and m_0. The black and grey curves represent all the combinations of semi-IVs (Z_0,Z_1) yielding a propensity score of $P=p$ and $P=\bar{p}$, respectively. These are iso-probability curves, which are identified directly from the data. Imagine $\bar{p}=p+dp$, where dp is a marginal change in the propensity score. At a given unobserved resistance to treatment $V=p$, to identify $m_1(p,z_1)$, we use the shift represented by the purple arrow, from $Z=z^1$ to $Z=\bar{z}^1$. This effectively shifts the selection probability from p to $\bar{p}$ by exogenously changing Z_0 from $\tilde{z}_0$ to $\tilde{z}'_0$ while holding $Z_1=z_1$ fixed. Similarly, to identify $m_0(p,z_0)$, we use the shift represented by the orange arrow, from $Z=z^0$ to $Z=\bar{z}^0$, which shifts the selection probability from p to $\bar{p}$ by exogenously changing Z_1 from $\tilde{z}_1$ to $\tilde{z}'_1$ while holding $Z_0=z_0$ fixed. Since it is not possible to move the probability while

holding both Z_0 and Z_1 fixed, the two shifts identifying the marginal treatment responses will always differ. Note that we generally do not use the point (z_0, z_1) for identification, and it is not necessary that $P(z_0, z_1) = p$ to identify $m_0(p, z_0)$ or $m_1(p, z_1)$.[18]

Overall, the identification of both MTRs, and of the MTE, at a given $V = p$ and $(Z_0, Z_1) = (z_0, z_1)$, requires that there exists four points as illustrated in Figure 2: for each d, there must be two points, $Z = z^d$ and $\bar{z}^d$, which both have $Z_d = z_d$ and are such that $P(z^d) = p$ and $P(\bar{z}^d) = \bar{p}$ in order to identify $m_d(p, z_d)$. In other words, we need points with $Z_1 = z_1$, and other points with $Z_0 = z_0$, on both isocurves of probability p and $\bar{p}$, allowing the propensity score to shift from p to $\bar{p}$ while holding either $Z_1 = z_1$ or $Z_0 = z_0$ fixed.

When we focus on the MTRs and MTE with continuous semi-IVs, if z^1 with $Z_1 = z_1$ exists on the isocurve of probability p, the existence of $\bar{z}^1$ is guaranteed by the fact that $V = p$ is in the interior of the support of the propensity score given $Z_1 = z_1$ (conversely for z^0 and $\bar{z}^0$), hence the previous conditions for the identification of the MTRs.

In general, to identify the average treatment effects on a larger set of compliers with $V \in [p, \bar{p}]$ for any general $\bar{p} > p$ (i.e., not only marginal changes in P), the requirement can be directly stated in terms of the existence of these four points. This corresponds to the identification of the general LATE$(p, \bar{p}, z_0, z_1)$ that is developed in the next subsection.

To contextualize this graph, in the application of earnings returns to manufacturing sector choice, identifying returns for workers with resistance $V = v$ in an environment with manufacturing GDP $Z_1 = z_1$ and nonmanufacturing GDP $Z_0 = z_0$ requires observing two environments with different relative sector GDPs but the same propensity to work in manufacturing, $P = v$. The mean manufacturing potential earnings at $V = v$ and $Z_1 = z_1$ are identified by marginally decreasing nonmanufacturing GDP (from $\tilde{z}_0$ to $\tilde{z}_0'$), which increases the propensity score from p to $\bar{p}$ while holding $Z_1 = z_1$ fixed. The resulting change in average DY is due solely to the entry of the marginal compliers with $V = v$ into treatment, identifying their mean manufacturing earnings. A converse approach identifies the mean nonmanufacturing potential earnings. The difference between these two mean potential earnings gives the MTE, i.e., the returns to working in manufacturing at $V = v$ given $Z_1 = z_1, Z_0 = z_0$. Overall, identification relies on comparing four otherwise similar environments that differ in their semi-IV combinations, thus providing the required variation in treatment incentives.

[18] One could define the MTE at $Z = z = (z_0, z_1)$, with $P(z) = p$, in an alternative notation as $\widetilde{\text{MTE}}(z)$, equal to MTE$(P(z), z_0, z_1)$ in the general notation. Then for each d, one can identify the newly defined MTR, $\tilde{m}_d(z) = m_d(p, z_d)$ by marginally shifting Z_{1-d} while holding $Z_d = z_d$ fixed. These treatment parameters are tied to specific values of the semi-IVs and are thus less general than the main definition. One advantage, however, is that these parameters can be identified while relaxing the monotonicity assumption. Indeed, since we focus on marginal shifts, unordered partial monotonicity combined with a comparable complier assumption (Mountjoy, 2022) are sufficient to identify $m_d(z)$ and MTE(z). See Online Appendix F.

3.2 Local average treatment effect (and responses)

3.2.1 Treatment parameters definition

Following Heckman and Vytlacil (2005), define the generalized local average treatment effect (LATE) parameters as a function of the set of compliers they correspond to (in terms of V):

$$\text{LATE}(v, v', z_0, z_1, x) = \mathbb{E}[Y_1 - Y_0 | v \leq V < v', Z_0 = z_0, Z_1 = z_1, X = x]. \tag{11}$$

The $\text{LATE}(v, v', z_0, z_1, x)$ is the average causal effects of D on Y for individuals with unobserved resistance to treatment $V \in [v, v')$ and observed characteristics $X = x, Z_0 = z_0, Z_1 = z_1$. It can also be interpreted as the mean gains for the individuals (compliers) who would be induced to switch into treatment if the probability of treatment changed from v to v', at fixed $X = x, Z_0 = z_0, Z_1 = z_1$. This interpretation allows detaching the definition of the LATE parameter from the specific semi-IV shift that identifies it, which is especially convenient with semi-IVs. For a more standard definition and identification of the $\text{LATE}(z, z', x)$ tied to specific semi-IV shifts from $Z = z$ to z', à la Imbens and Angrist (1994), see Online Appendix C. Since Imbens and Angrist (1994)'s LATE can always be mapped to the general LATE (11), I work with the latter for convenience due to its direct link with the MTE.

Following the split of MTE in MTRs, define the *local average treatment responses* (LATR):

$$\text{LATR}_d(v, v', z_d, x) = \mathbb{E}[Y_d | v \leq V < v', Z_d = z_d, X = x] \text{ for } d = 0, 1. \tag{12}$$

For each d, the LATR_d represents the mean potential outcome Y_d for individuals with $V \in [v, v')$ and covariates $X = x, Z_d = z_d$. Because of the exclusion of the semi-IVs, the LATR_d does not depend on Z_{1-d}, and this is what we exploit for identification.

Similar to the MTE-MTR link in (6), the LATE and LATRs are related by

$$\text{LATE}(v, v', z_0, z_1, x) = \text{LATR}_1(v, v', z_1, x) - \text{LATR}_0(v, v', z_0, x). \tag{13}$$

Thus, the LATE is identified if one can identify both LATR_d for compliers with $V \in [v, v')$.

3.2.2 Identification

Again, for notational convenience, omit X and proceed conditional on $X = x$.

Identification using the MTRs and MTE. By definition, note that the LATR and the LATE are directly related to the MTR and MTE by

$$\text{LATR}_d(v, v', z_d) = \int_v^{v'} \frac{m_d(\tilde{v}, z_d)}{v' - v} d\tilde{v}, \text{ and } \text{LATE}(v, v', z_0, z_1) = \int_v^{v'} \frac{\text{MTE}(\tilde{v}, z_0, z_1)}{v' - v} d\tilde{v}. \tag{14}$$

Therefore, $\text{LATE}(v, v', z_0, z_1)$ is identified if $\text{MTE}(\tilde{v}, z_0, z_1)$ is identified for all $\tilde{v} \in [v, v')$, i.e.,

if for each $d = 0, 1$, $\text{MTR}_d(\tilde{v}, z_d)$ is identified for all $\tilde{v} \in [v, v')$.

Direct identification (Figure 2)**.** The LATE identification via the MTE is not possible if continuously distributed semi-IVs are unavailable to generate continuous variation in both the propensity score given $Z_1 = z_1$ and given $Z_0 = z_0$ over the entire range $P \in [v, v')$. Fortunately, following the intuition from Figure 2, the identification of this general LATE can be done directly using only four points (i.e., two semi-IV shifts) without identifying all the intermediary MTRs. The intuition for identifying the MTE extends directly to larger shifts in the propensity score, from $P = v$ to $P = v'$. Using (7), we have:

$$\mathbb{E}[DY|Z_1 = z_1, P = v'] - \mathbb{E}[DY|Z_1 = z_1, P = v] = \int_v^{v'} m_1(\tilde{v}, z_1) d\tilde{v}.$$

So, if there exists $z^1 = (\tilde{z}_0, z_1)$ with $P(z^1) = v$ and $\bar{z}^1 = (\tilde{z}_0', z_1)$ with $P(\bar{z}^1) = v'$, then $\text{LATR}_1(v, v', z_1)$ is identified by

$$\begin{aligned} \text{LATR}_1(v, v', w_1) &= \frac{\mathbb{E}[DY|Z = \bar{z}^1] - \mathbb{E}[DY|Z = z^1]}{v' - v} \\ &= \frac{\mathbb{E}[DY|Z_1 = z_1, P = v'] - \mathbb{E}[DY|Z_1 = z_1, P = v]}{v' - v}. \end{aligned} \tag{15}$$

This follows the standard IV interpretation, applied to the outcome DY: holding $Z_1 = z_1$ fixed, shifts in Z_0 affect DY only indirectly through their effects on selection probabilities (see Property (4)). At fixed Z_1, the semi-IV Z_0 serves as an IV for the effect of D on DY. Thus, observable differences between $\mathbb{E}[DY|Z_1 = z_1, Z_0 = \tilde{z}_0']$ and $\mathbb{E}[DY|Z_1 = z_1, Z_0 = \tilde{z}_0]$ stem solely from the effect of shifting Z_0 from $\tilde{z}_0$ to $\tilde{z}_0'$ on selection and the corresponding set of compliers, $V \in [v, v')$. This shift corresponds to the purple arrow in Figure 2.

Similarly, Z_1 serves as an IV for the effect of D on $(1-D)Y$ given Z_0. Thus, if $z^0 = (z_0, \tilde{z}_1)$ with $P(z^0) = v$ and $\bar{z}^0 = (z_0, \tilde{z}_1')$ with $P(\bar{z}^0) = v'$ exist, $\text{LATR}_0(v, v', z_0)$ is identified by[19]

$$\text{LATR}_0(v, v', w_0) = -\frac{\mathbb{E}[(1-D)Y|Z = \bar{z}^0] - \mathbb{E}[(1-D)Y|Z = z^0]}{v' - v}. \tag{16}$$

This identifying semi-IV shift corresponds to the orange arrow in Figure 2.

Therefore, for any $v, v' \in [0, 1]$, $\text{LATE}(v, v', z_0, z_1)$ is identified by Equation (13), i.e., by

$$\text{LATE}(v, v', z_0, z_1) = \text{LATR}_1(v, v', z_1) - \text{LATR}_0(v, v', z_0),$$

provided that (at least) four semi-IVs combinations exist: (i) $z^1 = (\tilde{z}_0, z_1)$ with $P(z^1) = v$ and $\bar{z}^1 = (\tilde{z}_0', z_1)$ with $P(\bar{z}_1) = v'$, ensuring that $\text{LATR}_1(v, v', z_1)$ is identified by (15), and (ii) $z^0 = (z_0, \tilde{z}_1)$ with $P(z^0) = v$ and $\bar{z}^0 = (z_0, \tilde{z}_1')$ with $P(\bar{z}) = v'$, ensuring that $\text{LATR}_0(v, v', z_0)$ is identified by (16). Again, the monotonicity assumption plays a key role: it ensures that,

[19] The minus sign comes from the same reason as in (9), see footnote 17.

although LATR_1 and LATR_0 are identified from two different semi-IV shifts, both shifts induce comparable sets of compliers, with $V \in [v, v')$, because they result in the same observed change in the propensity score, from v to v'. Therefore, one can take the difference between these two LATRs to identify the LATE for $V \in [v, v')$.

Discussion. Even without continuous semi-IVs, some LATE for compliers with $V \in [v, v')$ at $Z_1 = z_1, Z_0 = z_0$ can be identified if, for both d, there exist two semi-IV combinations with $Z_d = z_d$, one yielding $P = v$ and the other $P = v'$ (see Figure 2). With two discrete semi-IVs, or better, a discrete and a continuous one, a broad set of LATE is likely identifiable. With two discrete semi-IVs, even if the observable combinations of semi-IVs do not align exactly with the same propensity scores, treatment effects may still be tightly bounded.

In the extreme case where the researcher only has two binary semi-IVs ($Z_d \in \{0, 1\}$), identifying any LATE requires specific conditions that the effect of the semi-IVs on the selection probabilities compensate each other exactly.[20] In most cases this will not hold, but the condition is easily testable and may hold in specific contexts.

Finally, note that the environment with $Z = (z_0, z_1)$ is generally not used to identify $\text{LATE}(v, v', z_0, z_1)$. Contrary to the standard LATE identification of Imbens and Angrist (1994) (see Online Appendix C), the identification of the parameters is detached from the specific value of the semi-IVs. However, in the special case where $P(z_0, z_1) = v$ (resp. v'), then identification is simplified because it only requires two additional points belonging to the isocurve of probability v' (resp. v): one with $Z_1 = z_1$ and another one with $Z_0 = z_0$.

3.3 Direct effect of the semi-IVs (targeted policy evaluation)

In general, for each $d = 0, 1$, an exogenous change[21] in Z_d from z_d to z'_d has two effects on the observed average outcome Y: (i) a *selection effect* since it alters the composition of the treated and untreated population by shifting the treatment incentives (P), and (ii) a *direct effect* since, holding the composition fixed, the change in Z_d directly affects the potential outcomes Y_d. The MTR and LATR capture these selection effects, which are identified by varying P at fixed Z_d using the excluded Z_{1-d}. In many cases, the direct effects of the semi-IVs on their respective potential outcomes are also of interest. For instance, when Z_d is a d-specific targeted policy – such as a subsidy or a tax cut/increase for a given industry

[20] More precisely, if $P(0,0) = P(1,1) = p$, then $\text{LATE}(p, P(1,0), 0, 1)$ and $\text{LATE}(p, P(0,1), 1, 0)$ are identified, while if $P(0,1) = P(1,0) = p$ then $\text{LATE}(p, P(0,0), 1, 1)$ and $\text{LATE}(p, P(1,1), 0, 0)$ are identified.

[21] To discuss the effects of *exogenous changes* in the semi-IVs, one must adopt the stronger version of the exogeneity assumption A1.1 outlined in footnote 8, i.e., $(Y_0(z_0), Y_1(z_1), D(z)) \perp\!\!\!\perp (Z_0, Z_1)|X$. This ensures that the semi-IVs are exogenous, influencing the potential outcomes only through general direct effects (because they are not excluded), without being correlated with d-specific unobservables. This stronger assumption is perfectly reasonable in most applications.

– evaluating the direct effect of Z_d on Y_d is itself valuable.[22] In the manufacturing example, the direct effect corresponds to the pass-through from manufacturing and nonmanufacturing GDPs to the earnings of the workers in the corresponding sectors.

To identify these direct effects, net of the selection effects, the general intuition is to vary Z_d while holding P fixed. Holding P fixed requires using the other semi-IV, Z_{1-d}. For example, conditional on $X = x$ left implicit as before, for each d, define the *average direct effect* of Z_d on the $D = d$ group (i.e., the *treated* or *untreated* population) when $P = p$ by

$$\Delta^{D=d}_{Z_d}(z_d, z'_d, p) = \quad \mathbb{E}[Y_d|D = d, Z_d = z'_d, P = p] - \mathbb{E}[Y_d|D = d, Z_d = z_d, P = p]. \tag{17}$$

This is the average effect of exogenously varying Z_d from z_d to z'_d on the individuals who choose $D = d$ when $P = p$, i.e., those with unobserved resistance to treatment $V \leq p$ if $D = 1$ and $V > p$ if $D = 0$. These average targeted policy effects of Z_d on Y_d, conditional on $D = d$, may depend on the underlying selected population V and thus vary with P. They are also identified using the excluded semi-IV as an IV. Specifically, for any policy change from $Z_d = z_d$ to z'_d, if there exists two distinct points, $\tilde{z} = (z_0, z_1)$ and $\tilde{z}' = (z'_0, z'_1)$, such that $P(\tilde{z}) = P(\tilde{z}') = p$, then the direct policy effect $\Delta^{D=d}_{Z_d}(z_d, z'_d, p)$ is identified by

$$\Delta^{D=d}_{Z_d}(z_d, z'_d, p) = \mathbb{E}[Y|D = d, Z = \tilde{z}'] - \mathbb{E}[Y|D = d, Z = \tilde{z}] \quad \text{for each } d = 0, 1.$$

In words, the policy effect of changing Z_d from z_d to z'_d is identified if implemented in environments with various Z_{1-d}, allowing $P = p$ to be observed with both values of Z_d.

The direct effects of Z_d for individuals with $V = v$ are also identified. Define

$$\Delta_{Z_d}(z_d, z'_d, v) = \mathbb{E}[Y_d|Z_d = z'_d, V = v] - \mathbb{E}[Y_d|Z_d = z_d, V = v], \tag{18}$$

the average effect of changing Z_d from z_d to z'_d on Y_d for individuals with unobserved resistance to treatment $V = v$. This is directly identified by

$$\Delta_{Z_d}(z_d, z'_d, v) = m_d(v, z'_d) - m_d(v, z_d),$$

if the corresponding MTR_d are identified at (v, z_d) and (v, z'_d).[23]

The direct effects of the semi-IVs are generally heterogenous with respect to V. Since these effects can be identified at different values of the unobserved V, one can test their homogeneity. Moreover, testing whether they are homogenous and null – i.e., if $\Delta_{Z_d}(z_d, z'_d, v) = 0$ for all $z_d, z'_d \in \mathcal{Z}_d$ and $v \in [0, 1]$ – provides a test of whether Z_d is a valid IV. Indeed, if Z_d has no effect on Y_d, it is excluded from both potential outcomes, making it a valid IV.

[22] This is akin to a mediation analysis for targeted policies, decomposing the overall effect of an exogenous change in Z_d on Y into the direct effect of Z_d on Y_d and the indirect effect mediated by selection D.

[23] Similarly, the policy effect on a set of compliers with $V \in [v, v')$ is identified using the LATR_d as $\Delta_{Z_d}(z_d, z'_d, v, v') = \mathbb{E}[Y_d|Z_d = z'_d, v \leq V < v'] - \mathbb{E}[Y_d|Z_d = z_d, v \leq V < v'] = \text{LATR}_d(v, v', z'_d) - \text{LATR}_d(v, v', z_d)$.

3.4 Other policy relevant treatment effects (PRTE)

Since the MTRs are identified, any parameter expressed as a weighted function of the MTRs is also identified. The generalized LATE is already a specific policy-relevant treatment effect (PRTE) with semi-IVs (with unit weights). More broadly, Mogstad et al. (2018) (Table 1) provide a list of parameters expressed as weighted functions of the MTRs, along with their corresponding weights with IVs. Many of these parameters and weights can be adapted to identify their counterparts using semi-IVs.

Similarly, when the semi-IVs have limited support – restricting the observable propensity score and the range of unobserved V for which MTRs are identified – treatment effects can be extrapolated using approaches analogous to those in Mogstad et al. (2018) for IVs.

4 Estimation

We observe a sample of $\{Y_i, D_i, Z_{0i}, Z_{1i}, X_i\}_{i=1}^N$. Each Z_d may include multiple valid, continuously distributed semi-IVs. Building on constructive identification arguments, we estimate the MTR_d functions separately to obtain the MTE (and LATE).[24] The procedure involves estimating three key objects: the propensity score $P(Z_0, Z_1, X)$ and the two conditional expectations $\mathbb{E}[DY|P, Z_1, X]$ and $\mathbb{E}[(1-D)Y|P, Z_0, X]$, which are used to estimate m_1 and m_0 as in Equations (8) and (9). The MTE is then obtained as their difference (Equation (6)) on the support where both MTRs are identified.

I propose three estimation approaches. First, a fully nonparametric approach. Second, a more practical semi-parametric approach, which is the semi-IV counterpart to the typical MTE estimation with IVs (Carneiro et al., 2011; Andresen, 2018). Finally, I present a revisited 2SLS for homogenous treatment effect models with semi-IVs.

4.1 Nonparametric estimation

I estimate the three objects of interest using local linear regressions, similar to Mountjoy (2022), who studies a discrete treatment problem with multiple IVs. For practicality, assume that the functions of interest are globally linear in X to avoid the curse of dimensionality.

First, estimate the propensity score P via local linear regression of D on Z_0, Z_1 and X, using a bivariate kernel on (Z_0, Z_1) for every values of $(z_0, z_1) \in \mathcal{Z}$.[25]

Next, use the fitted $\hat{P}_i$ as a generated covariate in the second stage to estimate the conditional expectations of the outcome for $d = 0$ and $d = 1$ by

[24] This section focuses on estimating the MTE and MTR but can be adapted to the estimation of the LATE and LATR with only discrete semi-IVs using a similar control function approach for P.

[25] Assume Z_0 and Z_1 each contain a single semi-IV; otherwise, estimation quickly becomes untractable.

$$\begin{pmatrix} \hat{\beta}_{Z_d}^{Y_d}(z_d,p) \\ \hat{\beta}_{P}^{Y_d}(z_d,p) \\ \hat{\beta}_{X}^{Y_d}(z_d,p) \end{pmatrix} = \underset{\beta_{Z_d},\beta_P,\beta_X}{\operatorname{argmin}} \sum_{i=1}^{N} K\left(\frac{Z_{di}-z_d}{h_d}, \frac{\hat{P}_i - p}{h_P}\right)\left(Y_i \mathbb{1}\{D_i = d\} - \beta_{Z_d} Z_{di} - \beta_P \hat{P}_i - X_i'\beta_X\right)^2.$$

For all values of z_d and p, X includes a constant, and $K(\cdot)$ is a bivariate kernel with bandwidths h_d for each Z_d and h_P for the fitted $\hat{P}$.

Local linear specifications are attractive because, or any observable (z_d, p), the local coefficient on P directly estimates MTR_d:

$$\hat{m}_d(p, z_d) = \hat{\beta}_P^{Y_d}(z_d, p) = \partial \mathbb{E}[Y_d \mathbb{1}\{D = d\} | Z_d = z_d, \hat{P} = p] / \partial \hat{P} \quad \text{for } d = 0, 1.$$

Then, given any (z_0, z_1, p) for which both MTRs are identified and estimated, the MTE is estimated as their difference, using the empirical counterpart of Equation (6):

$$\text{MTE}(p, z_0, z_1) = \hat{m}_1(p, z_1) - \hat{m}_0(p, z_0).$$

For inference, nonparametric bootstrap is used to obtain proper standard errors accounting for the estimation of $\hat{P}$ and its use as a generated regressor in the second stage.

4.2 Semi-parametric estimation

I show how the local IV estimation (Heckman and Vytlacil, 1999), particularly the approach estimating both MTRs separately (Andresen, 2018) with IVs, adapts to semi-IVs.

The semi-parametric model. The treatment is still determined by the weakly separable selection rule (2). The potential outcomes are given by

$$Y_d = \mu_d(X, Z_d) + U_d = X\beta_d + Z_d\delta_d + U_d, \quad \text{for } d = 0, 1, \tag{19}$$

where U_d is a d-specific unobserved shocks affecting the potential outcomes, and μ_d is a d-specific parametric function, that is specified as linear for simplicity.[26] The joint distribution of (U_0, U_1, V) is unrestricted, allowing for endogeneity between Y_d and D. Under this model, the partial exclusion of the semi-IVs can be written as $(U_d, V) \perp\!\!\!\perp Z_{1-d} | (Z_d, X)$ for each $d = 0, 1$. Property (3) can be expressed in terms of U_d by

$$\mathbb{E}[U_d | V, Z_d, Z_{1-d}, X] = \mathbb{E}[U_d | V, Z_d, X] \quad \text{for } d = 0, 1. \tag{20}$$

Separability. In addition, following the empirical literature estimating MTE (e.g., Carneiro

[26] Linearity can be relaxed as long as μ_d follows a known parametric form. Since the dimensions of X and Z_d are unspecified, even a simple linear model can be made more flexible by adding interactions between X and Z_d or incorporating flexible functions of Z_d (e.g., splines or polynomials).

et al., 2011; Maestas et al., 2013; Eisenhauer et al., 2015; Brinch et al., 2017; Cornelissen et al., 2018), and existing estimation packages (Brave and Walstrum, 2014; Andresen, 2018), I impose the following additional separability assumption to simplify the estimation.[27]

Assumption 3 (Separability)**.** *For each* $d \in \{0,1\}$, $\mathbb{E}[U_d|V, Z_0, Z_1, X] = \mathbb{E}[U_d|V]$.

Separability is a stronger assumption than the implication (20) derived from the exclusion restriction in the general model. Importantly, separability does *not* imply that the potential outcome Y_d is independent of Z_d and X. Rather, it restricts their dependence to be fully captured by the conditional mean function, $\mu_d(X, Z_d)$. Combined with the additive separability in (19), it implies that the effect of X, and more importantly, of Z_d on Y_d are homogenous with respect to U_d, and thus, with respect to V. While this restriction is strong, it is commonly assumed (with respect to X) in the empirical literature estimating MTE. Since Z_d behaves like a covariate X for Y_d, it is natural to extend the separability assumption to Z_d. Treatment effects remain generally heterogenous under separability.

Even under this stronger separability assumption, the exclusion of Z_{1-d} from Y_d still holds only conditional on Z_d (and X), i.e.,

$$\mathbb{E}[Y_d|V, Z_d, Z_{1-d}, X] = \mathbb{E}[Y_d|V, Z_d, X].$$

Thus if one does not control for Z_d, one cannot exclude Z_{1-d} from the average Y_d, since Z_{1-d} may indirectly affect the average Y_d through its correlation with Z_d in $\mu_d(X, Z_d)$.

MTRs. Under Assumption 3 and model (19), the MTRs are

$$\begin{aligned} m_d(v, z_d, x) &= \mathbb{E}[Y_d|V = v, Z_d = z_d, X = x] = \mu_d(x, z_d) + \overbrace{\mathbb{E}[U_d|V = v]}^{:= k_d(v)} \\ &= x\beta_d + z_d\delta_d + k_d(v) \qquad \text{for } d = 0, 1. \end{aligned} \tag{21}$$

In other words, the mean potential outcomes, m_d, can be decomposed into two components: the effect of the observables, $\mu_d(x, z_d)$, and the mean unobservable U_d given $V = v$, $k_d(v)$, which captures selection effects. While more restrictive than the general model, this structure has practical advantages. The direct effects of semi-IVs (and covariates) on Y_d are homogenous and more easily interpretable as they do not vary with the unobserved resistance to treatment V. They are also easier to identify.[28] Moreover, since $k_d(v)$ is independent of (X, Z_d), the full support of the propensity score (given $D = d$) can be used for identification,

[27] A stronger version of this separability assumption imposing that $U_d \perp\!\!\!\perp (Z_0, Z_1, X)$ for $d = 0, 1$ is sometimes assumed instead. But it is stronger than necessary, especially for X, because Assumption 3 does not restrict the dependence between X and V. Since $Z_d \perp\!\!\!\perp V$, the distinction is less important for the semi-IVs. See the Section 6.2 of Mogstad and Torgovitsky (2018) for more discussions.

[28] For any z_d, z'_d, x, one can identify $\mu_d(x, z'_d) - \mu_d(x, z_d)$ by controlling for P, for any $p \in \mathcal{P}$. Indeed,

$$\mu_d(x, z'_d) - \mu_d(x, z_d) = \mathbb{E}[Y|D = d, Z_d = z'_d, P = p, X = x] - \mathbb{E}[Y|D = d, Z_d = z_d, P = p, X = x],$$

rather than only its support conditional on specific $Z_d = z_d, X = x$ (and $D = d$).

Partially linear model. Under separability, the observed average outcomes in the treated and untreated subsamples of the semi-parametric model (19) follow a partially linear model:

$$\mathbb{E}[Y|D = d, Z_d, P, X] = \mathbb{E}[Y_d|D = d, Z_d, P, X] = X\beta_d + Z_d\delta_d + \kappa_d(P), \quad \text{for } d = 0, 1, \quad (22)$$

where $\mathbb{E}[U_d|D = d, P, Z_d, X] = \mathbb{E}[U_d|D = d, P] =: \kappa_d(P)$, a nonparametric function of P under separability.[29] Thus, average outcomes in each subsample follow a *partially linear model* with a linear component in covariates and semi-IVs, and an additive nonparametric term in P.

To estimate the MTRs (21), it suffices to consistently estimate the partially linear model (22), which directly provides an estimate for β_d and δ_d, and allows recovering $k_d(v)$ as a deterministic function of $\kappa_d(v)$. Indeed, $\kappa_1(p) = \mathbb{E}[U_1|V \leq p] = \int_0^p \mathbb{E}[U_1|V = v]/p \; dv$ and $\kappa_0(p) = \mathbb{E}[U_0|V > p] = \int_p^1 \mathbb{E}[U_0|V = v]/(1-p)dv$. Differentiating these gives

$$k_1(p) = \frac{\partial \kappa_1(p)p}{\partial p} \quad \text{and} \quad k_0(p) = -\frac{\partial \kappa_0(p)(1-p)}{\partial p}.$$

Estimation. Following the literature (Heckman et al., 1998, 2006), I apply nonparametric Robinson (1988)'s double residual regression to estimate the partially linear models separately for treated and untreated subsamples.[30] The estimation can be decomposed into three steps. In the first stage, the propensity score is estimated. Then, for each subsample of observations with $D = d$ separately, we estimate the effects of the semi-IVs and covariates on Y (thus, on Y_d in the subsample) by controlling for P nonparametrically. More precisely, nonparametrically regress Y, X, Z_d on P, and then run the linear regression of the residuals of Y on the residuals of X and Z_d to estimate $\hat{\beta}_d$ and $\hat{\delta}_d$. Finally, we estimate the remaining selection effects, $k_d(v)$. To do so, nonparametrically regress Y minus the effects of the covariates and semi-IVs, i.e., $Y - \hat{\beta}_d X + \hat{\delta}_d Z_d$, on P. The detailed estimation procedure with

because $\mathbb{E}[U_d|D = d, Z_d, X, P] = \mathbb{E}[U_d|D = d, P]$ is independent of Z_d and X. So the direct effect of changing Z_d from z_d to z'_d (at $X = x$) is identified if one can shift Z_d while holding P fixed (for any p), which is feasible if Z_{1-d} is relevant. With the linear specification, identification is even simpler: it suffices that there exist two $z_d \neq z'_d$ that yield the same $P = p$, for any $p \in \mathcal{P}$. In this case, δ_d is identified by:

$$\delta_d = \Big(\mathbb{E}[Y|D = d, Z_d = z'_d, P = p, X = x] - \mathbb{E}[Y|D = d, Z_d = z_d, P = p, X = x]\Big) \Big/ \Big(z'_d - z_d\Big).$$

[29] The conditioning on $P = p$ remains because $D = 1$ is equivalent to $V \leq p$ and $D = 0$ to $V > p$.

[30] *Alternative estimation.* Robinson (1988)'s double residual regression is the most popular procedure to nonparametrically control for P, but alternatives exist. For instance, $\kappa_d(p)$ can be specified as flexible polynomials or splines, following a nonparametric sieve approach. In this case, after estimating P in a first stage, the model can be estimated by running the regression of Y on X, Z_d, and a flexible function of P in each subsample with $D = d$. If the chosen specification for $\kappa_d(p)$ is sufficiently flexible, this yields consistent estimates of β_d, δ_d, and $\kappa_d(p)$ for values of p in the interior of the support of P given $D = d$.

semi-IVs is described in Online Appendix D.

These three steps parallel those of MTE estimation with IVs (Andresen, 2018).[31] The key difference is the presence of the direct effects of the semi-IVs, δ_d. Identifying these requires being able to fix the propensity score (i.e., to control for P) while shifting Z_d. This is achievable by also shifting the other excluded semi-IV, Z_{1-d}, to offset the effect of Z_d on P. Note that covariates are already incorporated in the standard MTE estimation with IVs (Andresen, 2018). Thus, the effects of the semi-IVs are naturally incorporated as additional coefficients estimated in the second stage. Therefore, MTE estimation with semi-IVs is similar but also requires the relevance of the semi-IVs to provide identifying excluded variation.

Estimated MTR and MTE. For each d, once β_d, δ_d and $k_d(v)$ are estimated, the MTR_d function is obtained by plugging these into (21) for any v in the interior of the support of $\widehat{P}$ given $D = d$, and for any (z_d, x) in the joint support of Z_d and X given $D = d$, i.e.,

$$\widehat{m}_d(v, z_d, x) = x\widehat{\beta}_d + z_d\widehat{\delta}_d + \widehat{k}_d(v).$$

Similarly, the MTE is estimated for any $(z_0, z_1, x) \in \mathcal{Z} \times \mathcal{X}$ and for any v in the *common support* of $\widehat{P}$, i.e., at the intersection of the supports of $\widehat{P}$ given $D = 1$ and given $D = 0$:[32]

$$\widehat{MTE}(v, z_0, z_1, x) = \underbrace{x(\widehat{\beta}_1 - \widehat{\beta}_0) + z_1\widehat{\delta}_1 - z_0\widehat{\delta}_0}_{\text{effects of the observables}} + \underbrace{\widehat{k}_1(v) - \widehat{k}_0(v)}_{\text{effects of the selection}}.$$

Under separability, the MTE can be decomposed into two components: one driven by observables, including the semi-IVs, and another driven by the selection. Note that the MTRs and MTE still depend on the covariates and semi-IVs. Typically, we report MTE/MTR curves at average covariates and semi-IV values (i.e., at $\bar{x}, \bar{z}_0, \bar{z}_1$) across all identified values of v (see Figure 6 for example). Once the MTRs and MTE are estimated, they can be used to estimate more complex parameters such as the LATR and LATE, or other PRTEs.

Inference. As in its IV counterpart, raw analytical standard errors from the final-stage local polynomial estimation of the MTR and MTE are biased, as they ignore that the propensity score and the effect of the semi-IVs and covariates were estimated in previous stages. To account for this, nonparametrically bootstrapping the standard errors is recommended.

[31] For more details, see also Carneiro et al. (2011), Appendix A.2.1. of Heckman et al. (1998), or the online supplement of Heckman et al. (2006).

[32] In practice, there are very few observations at the extreme ends of the propensity score support, leading to poor estimation of the MTRs in those regions. To address this, it is common to trim the supports further to focus on well-estimated interior points (e.g., Brinch et al., 2017). For example, one might restrict P in each subsample to values between the 1% and 99% percentiles of its conditional distribution given $D = d$ and define the common support as the intersection of these trimmed supports.

Performances. The semi-parametric estimator performs well with reasonable sample sizes, as illustrated using Monte Carlo simulations in Online Appendix E.

Practical implementation. The MTE and MTR estimation can be performed using the `semiivreg()` command from the `semiIVreg` package (Bruneel-Zupanc, 2024). The syntax of the function is designed to mimic the `ivreg()` command for IV estimation.

```
semiivreg(y ~ d | z0 + x | z1 + x, data=data)
```

In order to obtain the proper standard errors, the bootstrap is also coded as:

```
semiivreg_boot(y ~ d | z0 + x | z1 + x, Nboot=500, data=data)
```

4.3 Two-Stage Least Squares with semi-IVs

Consider a simpler model with homogenous treatment effects:

$$Y = \alpha + D\delta + DZ_1\delta_1 + (1-D)Z_0\delta_0 + DX\beta_1 + (1-D)X\beta_0 + U, \tag{23}$$

where $\mathbb{E}[U] = 0$ and $\mathbb{E}[U|Z_0, Z_1, X] = 0$ serves as the counterpart to the separability Assumption 3 in this model. The treatment is still determined by Equation (2). This is a model with homogenous treatment effects with respect to the unobservables, meaning that $\text{MTE}(v, z_0, z_1, x)$ is independent of v. However, the average treatment effect remains heterogenous with respect to observables. Indeed

$$\text{ATE}(z_0, z_1, x) = \mathbb{E}[Y_1 - Y_0|X = x, Z_0 = z_0, Z_1 = z_1] = \delta + x(\beta_1 - \beta_0) + z_1\delta_1 - z_0\delta_0.$$

Even when covariates X have the same effects on both potential outcomes ($\beta_1 = \beta_0$), the ATE still depends on the semi-IVs. Thus, we typically report specific ATEs, such as the ATE at the mean, $\text{ATE}(\bar{z}_0, \bar{z}_1, \bar{x})$.

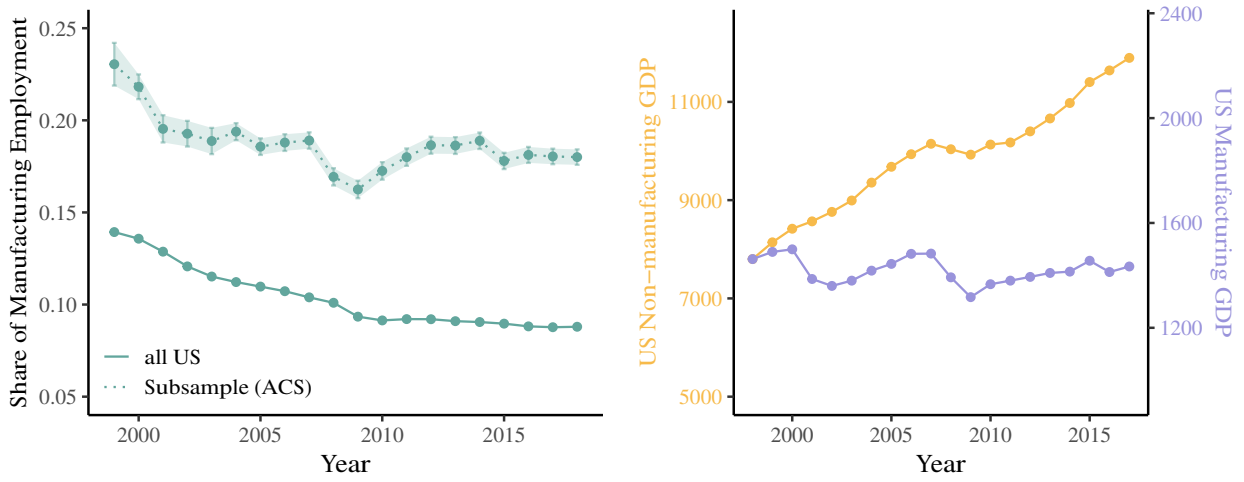
Standard IV-GMM estimation does not directly apply to Model (23) since the number of parameters exceeds the number of instruments. However, the model can be estimated using a modified two-stage least squares (2SLS) approach. In the first stage, estimate $P(Z_0, Z_1, X)$. Then, replace D by the fitted values $\hat{P}$ in the second stage outcome equation (23). This yields consistent estimates of all parameters $\alpha, \delta, \delta_0, \delta_1, \beta_0$, and β_1.[33]

[33] If the true underlying model is a model with heterogenous treatment effects, e.g., Model (19), the question of whether the estimates from this semi-IV-2SLS can be interpreted as convex combinations of positively weighted semi-IV LATE (like standard IV 2SLS) remains open.

5 Application: returns to working in manufacturing

5.1 Context: the manufacturing decline

Figure 3: Evolution of the manufacturing sector 1999 – 2018



(a) Manufacturing employment share

(b) GDP by industry

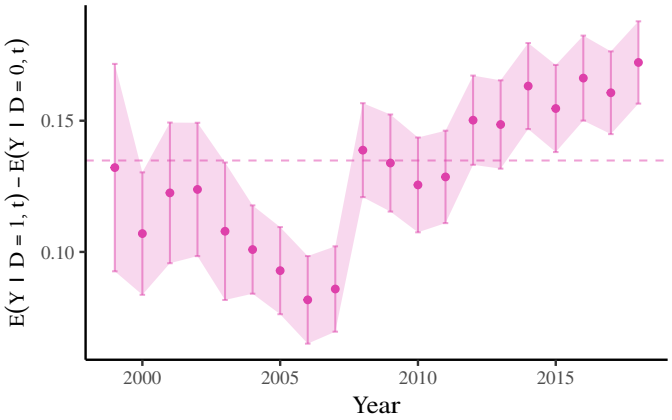
Notes: Left panel shows evolution of the manufacturing sector's employment share. The "all US" curve is computed directly from the Quarterly Census of Employment and Wages (QCEW) data of the Bureau of Labor Statistics, by dividing employment in the US manufacturing supersector (NAICS) by the overall US employment (excluding only the military). The "ACS subsample" curve represents the estimated probability of being employed in manufacturing for young white men (ages 18-30) with only a high school degree, obtained from the American Community Survey (ACS) by regressing a manufacturing dummy on year fixed effects. Right panel displays the evolution of real US GDP by sector (manufacturing vs. nonmanufacturing) in billions of 1999 dollars, using data from the Bureau of Economic Analysis (BEA).

The manufacturing sector's employment share has steadily declined in the US over the past decades (Baily and Bosworth, 2014; Pierce and Schott, 2016; Fort et al., 2018). As shown in Figure 3a, from 1999 to 2018, its share fell from 13.9% to 8.8% of total US employment: a 36.7% relative decline over 20 years. Even among young white men with only a high school education, a group more likely to work in manufacturing, the sector's employment share dropped from 23% to 18% in the same period, a 21.7% relative decline.

This decline in manufacturing coincides with the rise of the nonmanufacturing sector, particularly services (Lee and Wolpin, 2006; Buera and Kaboski, 2012; Autor and Dorn, 2013). As Figure 3b shows, US manufacturing GDP (in base-1999 dollars) barely changed from $1,462$ billion in 1998 to $1,433$ billion in 2017 (a 2% decline), while nonmanufacturing grew by 52.5%, from $7,800$ to $11,895$ billion over the same period.

A naive look at earnings by sector (Figure 4) suggests that (i) manufacturing workers earn 13.5% more on average than their nonmanufacturing counterparts throughout the pe-

Figure 4: Evolution of the difference of average (log) yearly earnings in manufacturing ($D = 1$) and nonmanufacturing sectors ($D = 0$).



Notes: $\mathbb{E}[\log(\text{earnings})|D = 1, \text{year}] - \mathbb{E}[\log(\text{earnings})|D = 0, \text{year}]$ is estimated by OLS, regressing log(earnings) on year fixed effects separately for the manufacturing and nonmanufacturing sectors. The horizontal line represents the 13.5% average earnings premium for manufacturing jobs in the sample.

riod and (ii), perhaps surprisingly, this gap *increased* from 13.2% in 1999 to 17.2% in 2018.

However, this naive comparison is likely biased due to selection on unobservables: even after controlling for demographics (age, education, race), manufacturing workers systematically differ from those in other sectors (Heckman and Sedlacek, 1985). In principle, an IV could address the endogeneity of sector choice, but finding one that affects sector choice without also affecting subsequent earnings is difficult. Fortunately, measures of local sector-specific strength can serve as valid semi-IVs instead, allowing for unbiased estimates of the returns to manufacturing (net of selection) and their evolution over time.[34]

5.2 Model specification and semi-IVs

I estimate the effect of working in manufacturing ($D = 1$) on the logarithm of the yearly earnings of workers (Y), measured in thousands of real 1999 US dollars.[35] I focus on young white male workers aged 18 to 30 with only a high school diploma and no college education, a demographic more likely to work in manufacturing and thus more affected by the sector's decline. To address the endogeneity of sector choice, I use the one-year lagged state-level manufacturing GDP (Z_1) and nonmanufacturing GDP (Z_0), measured in millions of 1999 US dollars, as semi-IVs. These should be valid as discussed in Example 1 of Section 2.2.

[34] I remain agnostic about the reasons behind the evolution of returns to working in manufacturing. The literature highlights two candidate explanations: (i) technological change, which affects the task content of manufacturing jobs (Autor et al., 2003; Goos et al., 2014), and (ii) trade, with rising import competition (Autor et al., 2013). Autor et al. (2015) propose methods to disentangle these two explanations.

[35] The manufacturing sector corresponds to the NAICS supersectors "31-33". The nonmanufacturing sector includes all other industries, private or public, except military jobs.

I estimate the following semi-parametric model for the potential earnings:

$$Y_d = \underbrace{\mu_d^{year} + \mu_d^{\text{state}} + \text{age}\beta_d^{age} + \text{age}^2\beta_d^{age^2}}_{:= X\beta_d} + log(Z_d)\delta_d + U_d, \tag{24}$$

which corresponds to Model (19) with covariates X including age, age^2, and state and year fixed effects. Sector-specific (d-specific) effects are estimated for all covariates, including fixed effects. The sector-specific state and year fixed effects control for intrinsic differences in state-level sectoral markets and national trends, capturing sector-specific costs, technological changes, and trade shocks. The sector-specific time fixed effects are particularly interesting as they track the evolution of earnings in both sectors and, consequently, the returns to working in manufacturing over time. With these fixed effects, the semi-IVs' identifying variations do not come from their levels but from deviations relative to sector-specific permanent state levels and the global time trend.[36] Since potential outcomes Y_d are not directly observed with data on (Y, D, Z, X), Model (24) cannot be estimated directly. Instead, it is estimated following the semi-parametric procedure in Section 4.2.

5.3 Data

The individual-level data on earnings, employment, and sector choice for young white men come from the ACS 1% yearly survey waves (2000-2019), effectively covering earnings from 1999 to 2018.[37] I include all US states except Hawaii, Alaska, and Washington D.C., as these are outliers in manufacturing employment, with some years in the sample having no observed young white male manufacturing workers there. State-level sector-specific GDP data come from the Bureau of Economic Analysis (BEA). All monetary variables are adjusted to real 1999 US dollars for consistency and comparability.

5.4 Results

5.4.1 Propensity score estimation (1$^{\text{st}}$ stage)

Propensity score. I estimate the propensity score using a probit regression of D on $\log(Z_0)$, $\log(Z_1)$, age, age^2, and state and year fixed effects. The results are reported in Column 1 of Table 1. Both semi-IVs are relevant for sector choice D. As expected, all else equal, a larger (lagged) manufacturing GDP increases the probability that young workers select

[36] The semi-IV effects δ_d are assumed to be homogenous across time and place, allowing identification using local variations in the semi-IVs over time. Identification would not be possible with interacted year-state fixed effects or if the semi-IVs had state- and year-specific effects, such as $\delta_{d,state,year}$.

[37] As the name suggests, each survey wave includes a random, representative sample of 1% of the US population. However, from 2000 to 2004, the coverage was smaller, with the 2000-2004 waves containing only 0.13%, 0.43%, 0.38%, 0.42%, and 0.42% of the population, respectively.

Table 1: Direct effect of the semi-IVs on P (first stage) and Y_d

	1^{st} stage	Outcome, log(earnings)		
	D	Y_0	Y_1	$Y_1 - Y_0$
$\log(Z_0)$:				
log(Non-manufacturing GDP)	-0.388^{***}	0.438^{***}		-0.438^{***}
	(0.057)	(0.028)		(0.028)
$\log(Z_1)$:				
log(Manufacturing GDP)	0.163^{***}		0.146^{***}	0.146^{***}
	(0.026)		(0.026)	(0.026)
Observations	476,117	350,146	79,783	

Notes: *p<0.1; **p<0.05; ***p<0.01. Covariates X include a quadratic term in age, along with state and year fixed effects. The first-stage model is a probit. The semi-IVs are measured in millions of 1999 US dollars, and earnings in the log of thousands of 1999 US dollars. I report the number of observations in the main (trimmed) sample. This number varies across bootstrap replications, as trimming in each replication is based on the re-estimated propensity score.

manufacturing jobs, while a larger nonmanufacturing GDP reduces it. Several factors may explain this. For instance, a larger sector is more attractive to young workers. Additionally, higher past GDP also allows the sector to recruit more in subsequent periods, leading to a passthrough effect from lagged GDP to future hires.

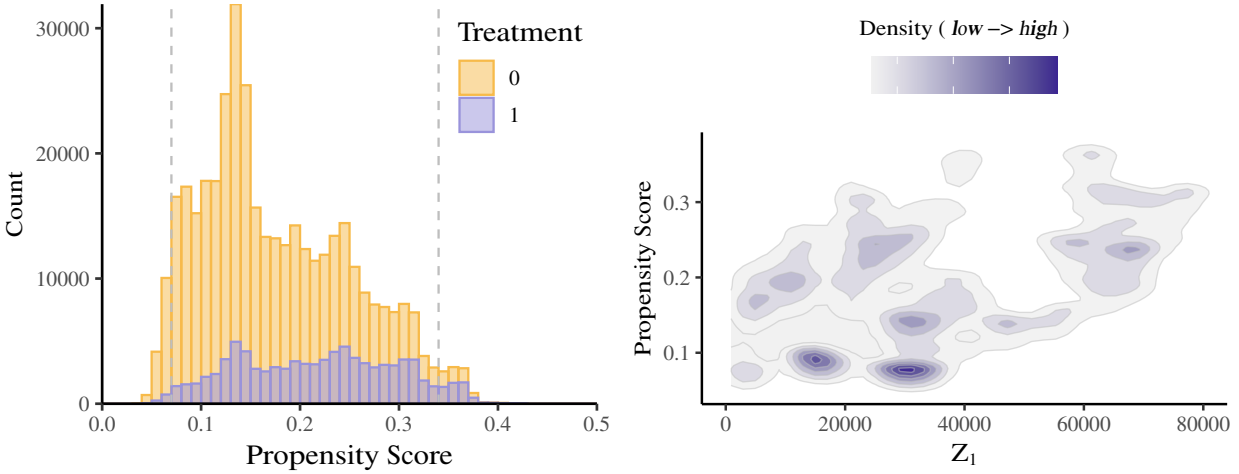
Common Support. Figure 5a reports the distribution of the estimated propensity score in both the manufacturing (treated) and nonmanufacturing (untreated) subsamples. Overall, the propensity score ranges from 3.19% to 42.12%, but the lowest and highest values do not appear in both subsamples. Since the estimation is noisier at the tails, I further trim the data by dropping observations with estimated propensity score below 7.9% or above 33.9%. This corresponds to the common support after trimming the 2.5% lowest and highest values of P in each subsample.[38] As a result, MTE and MTRs are only estimated for unobserved resistance to treatment $V \in \mathcal{P} = [0.079, 0.339]$.[39]

Figure 5b provides a more detailed view of the distribution of P conditional on Z_1. It is visible that Z_0 is relevant because, given any fixed $Z_1 = z_1$, P still exhibits substantial variation. This is the variation used for identification (while further controlling for covariates).

[38] All results are robust to less trimming, e.g., 1%, or even 0%, i.e., restricting to the common support.

[39] Following the applied literature with IVs (e.g. Carneiro et al., 2011), the common support could be expanded by including additional semi-IVs or modeling interacted effects of the semi-IVs and covariates on D. Including more covariates would also widen the support of P, but this would rely on the separability assumption. Instead, we prefer to use variation induced by the semi-IVs for identification.

Figure 5: Propensity score support and distribution



(a) Histogram of $\hat{P}$ for $D = 0$ and $D = 1$

(b) Density of Z_1 and $\hat{P}$

Notes: The propensity score P is estimated using a probit regression of D on $\log(Z_0)$, $\log(Z_1)$, age, age^2, and state and year fixed effects. The semi-IVs are measured in millions of 1999 US dollars. The vertical dashed grey lines in Figure 5a represent the common support $\mathcal{P} = [0.079, 0.339]$. Figure 5b shows the joint distribution of Z_1 and P. The x-axis is truncated for readability, as some US states have larger Z_1.

5.4.2 Effects on earnings (2nd stage)

Given the semi-parametric model (24), the MTR take the form (21) as in Model (19), i.e.,

$$m_d(v, z_d, x) = \mathbb{E}[Y_d | V = v, Z_d = z_d, X = x] = \underbrace{x'\beta_d + log(z_d)\delta_d}_{\text{effect of observables}} + \underbrace{k_d(v)}_{\text{selection effect}}.$$

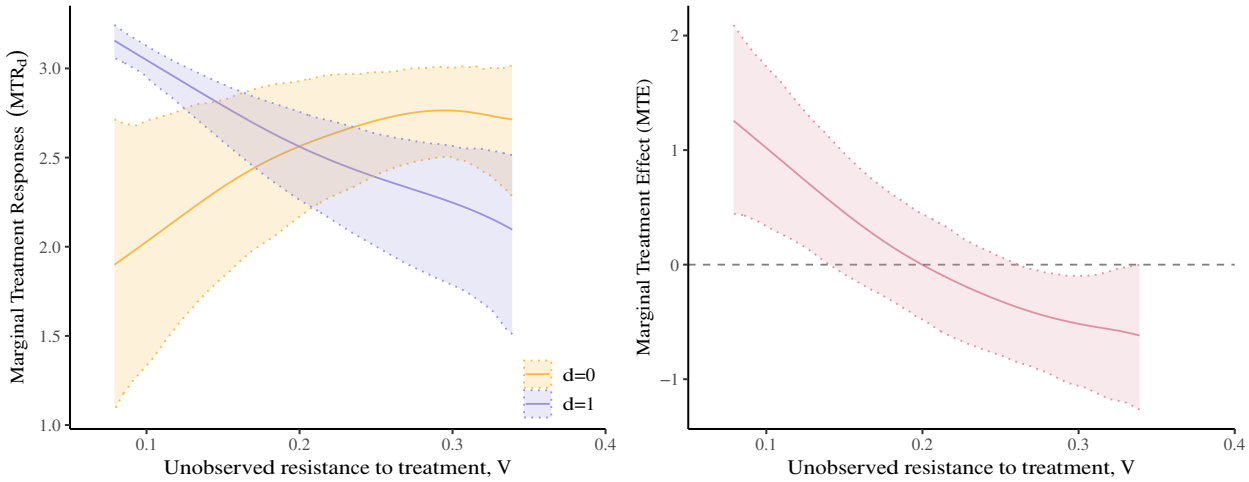
First, I discuss the estimated effects of the semi-IVs (Table 1), then the estimated MTRs and MTE evaluated at average covariates and semi-IVs (Figure 6 and Table 2).

Direct effects of the semi-IVs. Table 1 (Columns 2-4) reports the estimated effects of the semi-IVs on Y_0, Y_1, and $Y_1 - Y_0$, respectively. Each semi-IV significantly affects earnings in its corresponding sector. A 1% increase in nonmanufacturing GDP raises the nonmanufacturing workers' earnings by 0.438% in the next period. The pass-through from lagged manufacturing GDP to manufacturing workers' earnings is significantly smaller: a 1% increase raises next-period manufacturing earnings by 0.146%. Note that the ratio of the effects of the semi-IVs on their respective potential outcomes is close to the ratio of their effects on the first stage. This suggests selection on gains, where the semi-IVs primarily influence the first stage through their effect on Y_d. Thus, selection is mainly driven by the comparison of the semi-IVs' effects rather than by each semi-IV individually.

Table 2: (Some) Local Average Treatment Effects and Responses

	$\textbf{LATR}_0$ $\mathbb{E}[Y_0 \mid V \in \mathcal{S}_V]$	$\textbf{LATR}_1$ $\mathbb{E}[Y_1 \mid V \in \mathcal{S}_V]$	**LATE** $\mathbb{E}[Y_1 - Y_0 \mid V \in \mathcal{S}_V]$
Set of V, $\mathcal{S}_V$:			
Observable common support: $\mathcal{S}_V = [7.9\%, 33.9\%]$	2.499*** (0.207)	2.573*** (0.128)	0.075 (0.241)
Low V: $\mathcal{S}_V = [10\%, 15\%]$	2.184*** (0.313)	2.918*** (0.061)	0.734** (0.316)
High V: $\mathcal{S}_V = [25\%, 30\%]$	2.747*** (0.136)	2.319*** (0.183)	−0.428* (0.229)
Specific value $V = v$: → See Figure 6	MTR_0	MTR_1	MTE
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 6: Estimated Marginal Treatment Effects and Responses



Notes: All estimates are evaluated at the mean covariates ($\bar{x}$) and semi-IVs ($\bar{z}_0, \bar{z}_1$) values. The covariates include age, age^2, and state and year fixed effects. The average age is 25.05. The "average fixed effects" are based on the sample's average year and state distribution and correspond to averaging the state and year dummies. The semi-IVs are measured in millions of 1999 US dollars, and earnings in the log of thousands of 1999 US dollars. The first stage is estimated using a probit model (see Table 1). The partially linear model is estimated via double residual regression (Robinson, 1988), separately for each subsample $d = 0$ and $d = 1$. The first residual regression is performed via local linear regression with a Gaussian kernel and bandwidth 0.02. The second residual regression is performed via local quadratic regression with a Gaussian kernel and bandwidth 0.0482, chosen using the mean-squared error direct plug-in criteria of Fan and Gijbels (1996), as computed by Calonico et al. (2019) (`nprobust` package). Results are almost insensitive to bandwidth choices between 0.02 and 0.10. Standard errors are block bootstrapped (500 replications) at the state level. In Table 2, the LATE/LATR estimates are obtained by integrating the estimated MTE/MTR.

Heterogenous returns to working in manufacturing. Figure 6 reports the estimated MTR_d (left panel) and MTE (right panel) curves at $V = v$ for an average individual with $X = \bar{x}, Z_0 = \bar{z}_0, Z_1 = \bar{z}_1$.[40] These can only be estimated within the common support, i.e., for $V \in [0.079, 0.339]$. Over the entire common support, the estimated average earnings return to working in manufacturing (LATE) is 7.5% and not significantly different from zero (Table 2, first row). However, I find significant heterogeneity in the average potential outcomes (MTR) and treatment effects (MTE) across V (Figure 6). Individuals with low resistance to treatment (small V) have, on average, high manufacturing earnings and low nonmanufacturing earnings. This is consistent with selection on gains: these individuals are the first to select into manufacturing because they benefit the most from it. For them, the returns to working in manufacturing are high: they earn more than twice as much in the manufacturing sector as in the nonmanufacturing sector (MTE > 1). As V increases, average manufacturing earnings decrease while nonmanufacturing earnings increase. Around $V = 20\%$, the returns to working in manufacturing become negative.

This heterogeneity is also reflected in the LATR/LATE estimates (Table 2), obtained by integrating the MTR/MTE. Among individuals with relatively low resistance to treatment ($V \in [10\%, 15\%]$), average log earnings are 2.918 in manufacturing (LATR_1) versus 2.184 in nonmanufacturing (LATR_0). This results in a large positive LATE of 0.734, meaning that their earnings are, on average, 73.4% higher in manufacturing. Conversely, for workers with relatively high V ($V \in [25\%, 30\%]$), working in manufacturing reduces their earnings by 42.8% on average.

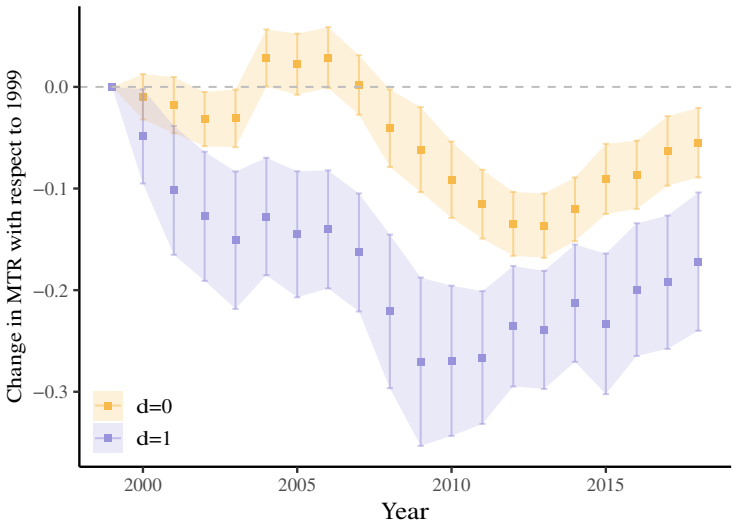
5.4.3 Evolution of the returns to working in manufacturing over time

Figure 8 shows the evolution of returns to working in manufacturing over time, relative to the 1999 baseline. This evolution accounts for time fixed effects, as well as the changes in the semi-IVs over time and their resulting effects on earnings. As shown in Figure 8b, the returns to working in manufacturing declined until 2009, down to -20%, and have slightly rebounded since then, although they remain about 10% lower than in 1999.[41] I also report the evolution of the manufacturing employment probability in Figure 8c and observe that it closely follows the evolution of the MTE. This suggests selection on earnings gains: workers primarily choose to enter manufacturing or not based on their expected earnings in each sector. The manufacturing employment share remains lower than in 1999, largely because

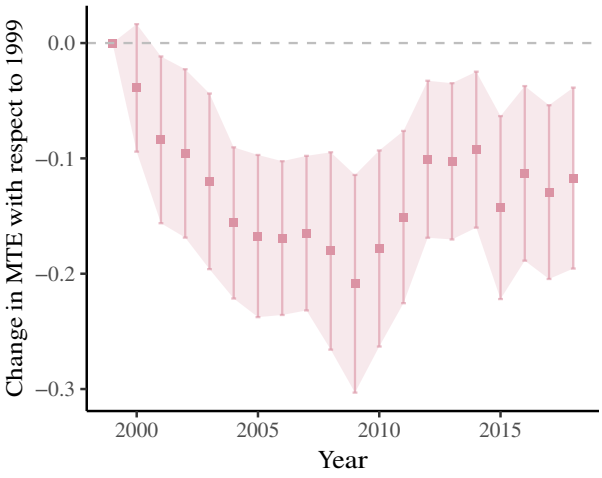
[40] Due to the partially linear model, the MTR and MTE *at the average* are equivalent to the *average MTR and MTE* at $V = v$ in the sample. For the fixed effects, the "average fixed effects" already correspond to the sample's average distribution of state and year.

[41] I interpret these estimates as changes in the MTE, though under the semi-parametric specification, they also correspond to changes in the global average treatment effects.

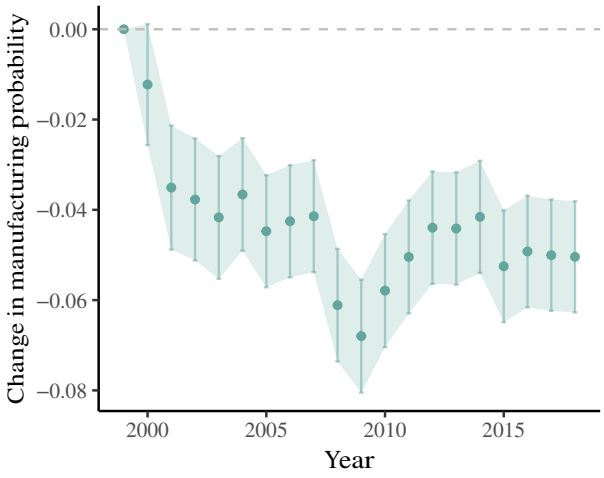
Figure 8: Yearly evolution of the MTR, MTE, and manufacturing employment probability



(a) $\Delta_t \text{MTR}_0$ and $\Delta_t \text{MTR}_1$



(b) $\Delta_t \text{MTE}$



(c) $\Pr(D = 1|\text{Year}= t)$.

Notes: Evolution of the MTR and MTE over time, relative to the baseline year 1999. The time evolution accounts for the evolution of the semi-IVs over time. To ensure their effects are differenced out and do not influence the computations, state and age are fixed at arbitrary values, and comparisons are made at a fixed $V = v$ to remove the influence of time-invariant $k_d(v)$. For each year t, I compute the average evolution of sector GDP (semi-IVs), $log(\bar{z}_{0,t})$ and $log(\bar{z}_{1,t})$, along with the estimated year fixed effect μ_d^t, for each d. The top panel shows the evolution of the MTR_d, $\Delta_{t-1999}^{MTR_d} = \mu_d^t + (log(\bar{z}_{d,t}) - log(\bar{z}_{d,1999}))\delta_d$ for each $d = 0, 1$. The bottom left panel shows the evolution of MTE given by $\Delta_{t-1999}^{MTE}\Delta_{t-1999}^{MTR_1} - \Delta_{t-1999}^{MTR_0}$. The bottom right panel shows the evolution of the manufacturing employment probability, estimated separately using an OLS regression of sector choice D on year fixed effects.

returns to working in manufacturing have not fully recovered. This pattern was undetectable in the naive (OLS) estimation of the returns to working manufacturing (Figure 4), which misleadingly suggests higher returns in 2018 than in 1999. Using the semi-IV approach to correct for sector choice endogeneity provides a clearer picture and helps explain the persistent decline in manufacturing employment.

Perhaps surprisingly, the evolution of both MTRs (Figure 8a) reveals a harsher reality for young, uneducated white men: while their manufacturing earnings (Y_1) have declined, their nonmanufacturing earnings (Y_0) have also fallen, albeit less sharply, since 1999. This highlights a marked deterioration in the economic prospects for this population, regardless of their sector choice.

A practical remark. As highlighted by this last result, the MTRs are often informative on their own. In general, they provide valuable information about the selection process, about the dependence between potential outcomes and resistance to treatment. In this application, we observe a standard scenario: m_1 is high when V is low and decreases with V, while the reverse holds for m_0. However, many other scenarios could yield the same observed MTE. For instance, the MTE could be entirely driven by m_1 while m_0 would remain flat. Reporting only the MTE thus results in a loss of information. Since MTRs can also be estimated with IVs at no additional cost, I recommend always reporting them alongside the MTE.

6 Conclusion

This paper shows that semi-IVs provide a viable alternative to IVs for causal inference, particularly for estimating LATE and MTE. This novel approach is especially useful since semi-IVs are often more readily available, as illustrated by the examples and the returns to working in manufacturing application. Expanding the set of valid identifying variations broadens the empirical researchers' toolkit. This expansion should enable researchers to address important economic questions for which finding a credible IV was previously infeasible.

References

Abaluck, Jason, Mauricio Caceres Bravo, Peter Hull, and Amanda Starc, "Mortality Effects and Choice Across Private Health Insurance Plans," *The Quarterly Journal of Economics*, 2021, *136* (3), 1557–1610.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Peter Hull, and Parag A. Pathak, "Charters without lotteries: Testing takeovers in New Orleans and Boston," *American Economic Review*, 2016, *106* (7), 1878–1920.

Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak, " Accountability and Flexibility in Public Schools: Evidence from Boston's Charters And Pilots," *The Quarterly Journal of Economics*, 2011, *126* (2), 699–748.

Abramitzky, Ran, "The effect of redistribution on migration: Evidence from the Israeli kibbutz," *Journal of Public Economics*, 2009, *93* (3-4), 498–511.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American Economic Review*, 2012, *102* (5), 1832–1856.

_ , _ , and _ , "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration," *Journal of Political Economy*, 2014, *122* (3), 467–506.

Agersnap, Ole, Amalie Jensen, and Henrik Kleven, "The welfare magnet hypothesis: Evidence from an immigrant welfare scheme in Denmark," *American Economic Review: Insights*, 2020, *2* (4), 527–542.

Akcigit, Ufuk, Salomé Baslandze, and Stefanie Stantcheva, "Taxation and the International Mobility of Inventors," *American Economic Review*, October 2016, *106* (10), 2930–81.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Andresen, Martin E and Martin Huber, "Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction," *The Econometrics Journal*, 02 2021, *24* (3), 536–558.

Andresen, Martin Eckhoff, "Exploring marginal treatment effects: Flexible estimation using Stata," *The Stata Journal*, 2018, *18* (1), 118–158.

Angrist, Joshua D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 1990, *80* (3), 313–336.

_ and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 1991, *106* (4), 979–1014.

_ , Guido W. Imbens, and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 1996, *91* (434), 444–455.

_ , Parag A. Pathak, and Christopher R. Walters, "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, October 2013, *5* (4), 1–27.

_ , Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters, "Inputs and Impacts in Charter Schools: KIPP Lynn," *American Economic Review*, May 2010, *100* (2), 239–43.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment," *American economic review*, 2002, *92* (5), 1535–1558.

Arcidiacono, Peter, "Ability sorting and the returns to college major," *Journal of Econometrics*, 2004, *121* (1), 343 – 375. Higher education (Annals issue).

_ , V Joseph Hotz, Arnaud Maurel, and Teresa Romano, "Ex ante returns and occupational choice," *Journal of Political Economy*, 2020, *128* (12), 4475–4522.

Autor, David H. and David Dorn, "The growth of low-skill service jobs and the polarization of the US labor market," *American economic review*, 2013, *103* (5), 1553–1597.

_ , _ , and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, October 2013, *103* (6), 2121–68.

_ , _ , and _ , "Untangling trade and technology: Evidence from local labour markets," *The Economic Journal*, 2015, *125* (584), 621–646.

_ , Frank Levy, and Richard J Murnane, "The skill content of recent technological change: An empirical exploration," *The Quarterly journal of economics*, 2003, *118* (4), 1279–1333.

Baily, Martin Neil and Barry P Bosworth, "US manufacturing: Understanding its past and its potential future," *Journal of Economic Perspectives*, 2014, *28* (1), 3–26.

Bayer, Patrick, Shakeeb Khan, and Christopher Timmins, "Nonparametric Identification and Estimation in a Roy Model With Common Nonpecuniary Returns," *Journal of Business and Economic Statistics*, 2011, *29* (2), 201–215.

Behaghel, Luc, Clément De Chaisemartin, and Marc Gurgand, "Ready for boarding? The effects of a boarding school for disadvantaged students," *American Economic Journal: Applied Economics*, 2017, *9* (1), 140–164.

Björklund, Anders and Robert Moffitt, "The estimation of wage gains and welfare gains in self-selection models," *The Review of Economics and Statistics*, 1987, pp. 42–49.

Borjas, George J., "Self-Selection and the Earnings of Immigrants," *The American Economic Review*, 1987, *77* (4), 531–553.

Borjas, George J, "Immigration and welfare magnets," *Journal of labor economics*, 1999, *17* (4), 607–637.

Brave, Scott and Thomas Walstrum, "Estimating marginal treatment effects using parametric and semiparametric methods," *The Stata Journal*, 2014, *14* (1), 191–217.

Brinch, Christian N, Magne Mogstad, and Matthew Wiswall, "Beyond LATE with a discrete instrument," *Journal of Political Economy*, 2017, *125* (4), 985–1039.

Bruneel-Zupanc, Christophe, *semiIVreg: Semi-Instrumental Variable (semi-IV) Regression* 2024. R package version 1.0.0, https://cbruneelzupanc.github.io/semiIVreg/.

_ and Jad Beyhum, "Identification with possibly invalid IVs," *arXiv preprint arXiv:2401.03990*, 2024.

_ and _ , "Semi-IV Quantile Regression," 2025.

Buera, Francisco J and Joseph P Kaboski, "The rise of the service economy," *American Economic Review*, 2012, *102* (6), 2540–2569.

Caetano, Carolina, Christoph Rothe, and Neşe Yıldız, "A discontinuity test for identification in triangular nonseparable models," *Journal of Econometrics*, 2016, *193* (1), 113–122.

Calonico, Sebastian, Matias D Cattaneo, and Max H Farrell, "nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference," *arXiv preprint arXiv:1906.00198*, 2019.

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, 1995.

_ , "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 2001, *69* (5), 1127–1160.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil, "Estimating Marginal Returns to Education," *The American Economic Review*, 2011, *101* (6), 2754–2781.

_ , Jishnu Das, and Hugo Reis, "The Value of Private Schools: Evidence from Pakistan," *The Review of Economics and Statistics*, 09 2022, pp. 1–45.

Chernozhukov, Victor and Christian Hansen, "An IV Model of Quantile Treatment Effects," *Econometrica*, 2005, *73* (1), 245–261.

Chesher, Andrew, "Instrumental Variable Models for Discrete Outcomes," *Econometrica*, 2010, *78* (2), 575–601.

_ and Adam M. Rosen, "Generalized Instrumental Variable models," *Econometrica*, 2017, *85* (3), 959–

989.

Chiquiar, Daniel and Gordon H. Hanson, "International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States," *Journal of political Economy*, 2005, *113* (2), 239–281.

Combes, Pierre-Philippe, Gilles Duranton, Laurent Gobillon, Diego Puga, and Sébastien Roux, "The productivity advantages of large cities: Distinguishing agglomeration from firm selection," *Econometrica*, 2012, *80* (6), 2543–2594.

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg, "From LATE to MTE: Alternative methods for the evaluation of policy interventions," *Labour Economics*, 2016, *41*, 47–60.

_ , _ , _ , and _ , "Who benefits from universal child care? Estimating marginal returns to early child care attendance," *Journal of Political Economy*, 2018, *126* (6), 2356–2409.

Dahl, Gordon B., "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 2002, *70* (6), 2367–2420.

de Chaisemartin, Clément, "Tolerating defiance? Local average treatment effects without monotonicity," *Quantitative Economics*, 2017, *8* (2), 367–396.

D'Haultfœuille, Xavier and Philippe Février, "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 2015, *83* (3), 1199–1210.

Dobbie, Will and Roland G. Fryer, "Charter Schools and Labor Market Outcomes," *Journal of Labor Economics*, 2020, *38* (4), 915–957.

Dubin, Jeffrey and Daniel L. McFadden, "An econometric analysis of residential electric appliance holdings and consumption," *Econometrica*, 1984, *52* (2), 345–362.

D'Haultfœuille, Xavier and Arnaud Maurel, "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, 2013, *174* (2), 95–106.

_ , _ , and Yichong Zhang, "Extremal quantile regressions for selection models and the black–white wage gap," *Journal of Econometrics*, 2018, *203* (1), 129–142.

_ , Stefan Hoderlein, and Yuya Sasaki, "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*, 2021.

Einav, Liran, Amy Finkelstein, and Neale Mahoney, "Producing Health: Measuring Value Added of Nursing Homes," Technical Report, National Bureau of Economic Research 2022.

Eisenhauer, Philipp, James J Heckman, and Edward J Vytlacil, "The generalized Roy model and the cost-benefit analysis of social programs," *Journal of Political Economy*, 2015, *123* (2), 413–443.

Fan, Jianqing. and Irène Gijbels, *Local polynomial modelling and its applications* Monographs on statistics and applied probability ; 66, 1st ed. ed., Boca Raton, Fla: Chapman and Hall/CRC, 1996.

Feger, Fabian, Nicola Pavanini, and Doina Radulescu, "Welfare and Redistribution in Residential Electricity Markets with Solar Power," *The Review of Economic Studies*, 02 2022, *89* (6), 3267–3302.

Finkelstein, Amy, Nathaniel Hendren, and Erzo F. P. Luttmer, "The Value of Medicaid: Interpreting Results from the Oregon Health Insurance Experiment," *Journal of Political Economy*, 2019, *127* (6), 2836–2874.

_ , _ , and Mark Shepard, "Subsidizing health insurance for low-income adults: Evidence from Massachusetts," *American Economic Review*, 2019, *109* (4), 1530–1567.

Fort, Teresa C, Justin R Pierce, and Peter K Schott, "New perspectives on the decline of US manufacturing employment," *Journal of Economic Perspectives*, 2018, *32* (2), 47–72.

French, Eric and Christopher Taber, "Chapter 6 - Identification of Models of the Labor Market," in Orley Ashenfelter and David Card, eds., *Orley Ashenfelter and David Card, eds.*, Vol. 4 of *Handbook of Labor Economics*, Elsevier, 2011, pp. 537–617.

Gao, Wayne Yuan and Rui Wang, "IV Regressions without Exclusion Restrictions," 2023.

Geweke, John, Gautam Gowrisankaran, and Robert J. Town, "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 2003, *71* (4), 1215–1238.

Goff, Leonard, "A vector monotonicity assumption for multiple instruments," *Journal of Econometrics*, 2024, *241* (1), 105735.

Goos, Maarten, Alan Manning, and Anna Salomons, "Explaining job polarization: Routine-biased technological change and offshoring," *American economic review*, 2014, *104* (8), 2509–2526.

Gowrisankaran, Gautam and Robert J Town, "Estimating the quality of care in hospitals using instrumental variables," *Journal of health economics*, 1999, *18* (6), 747–767.

Grigolon, Laura, Mathias Reynaert, and Frank Verboven, "Consumer valuation of fuel costs and tax policy: Evidence from the European car market," *American Economic Journal: Economic Policy*, 2018, *10* (3), 193–225.

Groote, Olivier De and Frank Verboven, "Subsidies and Time Discounting in New Technology Adoption: Evidence from Solar Photovoltaic Systems," *American Economic Review*, June 2019, *109* (6), 2137–72.

Heckman, James, "Varieties of Selection Bias," *The American Economic Review*, 1990, *80* (2), 313–318.

_ , Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *66* (5), 1017–1098.

Heckman, James J. and Bo E. Honoré, "The Empirical Content of the Roy Model," *Econometrica*, 1990, *58* (5), 1121–1149.

_ and Edward J. Vytlacil, "Local instrumental variables and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 1999, *96* (8), 4730–4734.

_ and _ , "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

_ and _ , "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in James J. Heckman and Edward E. Leamer, eds., *James J. Heckman and Edward E. Leamer, eds.*, Vol. 6 of *Handbook of Econometrics*, Elsevier, 2007, pp. 4779–4874.

_ and _ , "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 James J. Heckman and Edward E. Leamer, eds., *James J. Heckman and Edward E. Leamer, eds.*, Vol. 6 of *Handbook of Econometrics*, Elsevier, 2007, pp. 4875–5143.

_ and Guilherme L Sedlacek, "Heterogeneity, aggregation, and market wage functions: an empirical model of self-selection in the labor market," *Journal of political Economy*, 1985, *93* (6), 1077–1125.

_ and _ , "Self-selection and the distribution of hourly wages," *Journal of Labor Economics*, 1990, *8* (1, Part 2), S329–S363.

_ and Rodrigo Pinto, "Unordered monotonicity," *Econometrica*, 2018, *86* (1), 1–35.

_ , Sergio Urzua, and Edward J. Vytlacil, "Understanding instrumental variables in models with essential heterogeneity," *The review of economics and statistics*, 2006, *88* (3), 389–432.

_ , _ , and _ , "Instrumental Variables in Models with Multiple Outcomes: the General Unordered Case," *Annales d'Économie et de Statistique*, 2008, (91/92), 151–174.

Hull, Peter, "Estimating hospital quality with quasi-experimental data," *Available at SSRN 3118358*, 2018.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Keane, Michael P and Kenneth I Wolpin, "The career decisions of young men," *Journal of political Economy*, 1997, *105* (3), 473–522.

Kennan, John and James R Walker, "Wages, welfare benefits and migration," *Journal of Econometrics*, 2010, *156* (1), 229–238.

_ and _ , "The effect of expected income on individual migration decisions," *Econometrica*, 2011, *79* (1),

211–251.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad, "Field of Study, Earnings, and Self-Selection," *The Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

_ , "The identification region of the potential outcome distributions under instrument independence," *Journal of Econometrics*, 2021, *225* (2), 231–253. Themed Issue: Treatment Effect 1.

Klein, Roger and Francis Vella, "Estimating a class of triangular simultaneous equations models without exclusion restrictions," *Journal of Econometrics*, 2010, *154* (2), 154–164.

Kleven, Henrik, Camille Landais, Mathilde Munoz, and Stefanie Stantcheva, "Taxation and migration: Evidence and policy implications," *Journal of Economic Perspectives*, 2020, *34* (2), 119–142.

Kleven, Henrik Jacobsen, Camille Landais, and Emmanuel Saez, "Taxation and international migration of superstars: Evidence from the European football market," *American economic review*, 2013, *103* (5), 1892–1924.

_ , _ , _ , and Esben Schultz, "Migration and wage effects of taxing top earners: Evidence from the foreigners' tax scheme in Denmark," *The Quarterly Journal of Economics*, 2014, *129* (1), 333–378.

Kline, Patrick and Christopher R Walters, "Evaluating public programs with close substitutes: The case of Head Start," *The Quarterly Journal of Economics*, 2016, *131* (4), 1795–1848.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens, "Identification and Inference With Many Invalid Instruments," *Journal of Business & Economic Statistics*, 2015, *33* (4), 474–484.

Kwon, Soonwoo and Jonathan Roth, "Testing Mechanisms," 2024.

Lagakos, David and Michael E. Waugh, "Selection, Agriculture, and Cross-Country Productivity Differences," *American Economic Review*, April 2013, *103* (2), 948–80.

Lee, Donghoon and Kenneth I Wolpin, "Intersectoral labor mobility and the growth of the service sector," *Econometrica*, 2006, *74* (1), 1–46.

Lee, Lung-Fei, "Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables," *International economic review*, 1978, pp. 415–433.

Lee, Sokbae and Bernard Salanié, "Identifying effects of multivalued treatments," *Econometrica*, 2018, *86* (6), 1939–1963.

Lewbel, Arthur, "Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," *Journal of Business and Economic Statistics*, 2012, *30* (1), 67–80.

_ , "Identification and estimation using heteroscedasticity without instruments: The binary endogenous regressor case," *Economics Letters*, 2018, *165*, 10–12.

_ , "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, 2019, *57* (4), 835–903.

_ , Susanne M. Schennach, and Linqi Zhang, "Identification of a Triangular Two Equation System Without Instruments," *Journal of Business & Economic Statistics*, 2023, *0* (0), 1–12.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *American economic review*, 2013, *103* (5), 1797–1829.

Manski, Charles F. and John V. Pepper, "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 2000, *68* (4), 997–1010.

Masten, Matthew A. and Alexandre Poirier, "Identification of Treatment Effects Under Conditional Partial Independence," *Econometrica*, 2018, *86* (1), 317–351.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters, "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, November

2021, *111* (11), 3663–98.

_ and _, "Identification and extrapolation of causal effects with instrumental variables," *Annual Review of Economics*, 2018, *10* (1), 577–613.

_, Andres Santos, and Alexander Torgovitsky, "Using instrumental variables for inference about policy relevant treatment parameters," *Econometrica*, 2018, *86* (5), 1589–1619.

Mountjoy, Jack, "Community colleges and upward mobility," *American Economic Review*, 2022, *112* (8), 2580–2630.

Muñoz, Mathilde, "Do European Top Earners React to Labour Taxation Through Migration ?," June 2021. working paper or preprint.

Neumark, David and Helen Simpson, "Place-based policies," in "Handbook of regional and urban economics," Vol. 5, Elsevier, 2015, pp. 1197–1287.

Newey, Whitney and Sami Stouli, "Control variables, discrete instruments, and identification of structural functions," *Journal of Econometrics*, 2021, *222* (1, Part A), 73–88. Annals Issue: Structural Econometrics Honoring Daniel McFadden.

Pierce, Justin R and Peter K Schott, "The surprisingly swift decline of US manufacturing employment," *American Economic Review*, 2016, *106* (7), 1632–1662.

Rigobon, Roberto, "Identification through heteroskedasticity," *Review of Economics and Statistics*, 2003, *85* (4), 777–792.

Robinson, Peter M, "Root-N-consistent semiparametric regression," *Econometrica: Journal of the Econometric Society*, 1988, pp. 931–954.

Roca, Jorge De La and Diego Puga, "Learning by working in big cities," *The Review of Economic Studies*, 2017, *84* (1), 106–142.

Staiger, Douglas O, Joanne Spetz, and Ciaran S Phibbs, "Is there monopsony in the labor market? Evidence from a natural experiment," *Journal of Labor Economics*, 2010, *28* (2), 211–236.

Torgovitsky, Alexander, "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 2015, *83* (3), 1185–1197.

Tsyawo, Emmanuel Selorm, "Feasible IV regression without excluded instruments," *The Econometrics Journal*, 2023, *26* (2), 235–256.

Verboven, Frank, "Quality-based price discrimination and tax incidence: evidence from gasoline and diesel cars," *RAND Journal of Economics*, 2002, pp. 275–297.

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

Walters, Christopher R., "The Demand for Effective Charter Schools," *Journal of Political Economy*, 2018, *126* (6), 2179–2223.

Wang, Rui, "Point Identification of LATE with Two Imperfect Instruments," 2023.

Supplementary materials: online appendices

A Additional Literature Review

Relaxing the exclusion restriction. Given the importance of IVs for causal inference and the restrictiveness of the standard assumptions imposed on them, numerous papers explore identification while relaxing some of these assumptions (D'Haultfœuille and Février, 2015; Torgovitsky, 2015; de Chaisemartin, 2017; Newey and Stouli, 2021; Mogstad et al., 2021; Kitagawa, 2021) or provide tests of their validity (Kitagawa, 2015; Caetano et al., 2016; Andresen and Huber, 2021; D'Haultfœuille et al., 2021; Kwon and Roth, 2024). This paper focuses on relaxing the exclusion restriction and relates to the literature showing that identification can also be achieved without instruments or with included instruments, provided additional conditions hold, e.g., conditional partial independence (Masten and Poirier, 2018), identification-at-infinity (D'Haultfœuille et al., 2018), identification based on heteroskedasticity (Rigobon, 2003; Klein and Vella, 2010; Lewbel, 2012, 2018; Lewbel et al., 2023), or specific functional form or relevance conditions (D'Haultfœuille et al., 2021; Gao and Wang, 2023; Tsyawo, 2023). See Lewbel (2019) for an extensive discussion. Unlike these approaches, I use the *same* general nonparametric nonseparable model – i.e., generalized Roy models with weakly separable selection rule (Imbens and Angrist, 1994; Vytlacil, 2002; Heckman and Vytlacil, 2005) – as in the standard IV literature (extended with included semi-IVs). Crucially, I do not need to compensate for relaxing the exclusion restriction by adding further structure to the model (e.g., specifying the distribution of the error terms or introducing other assumptions). Instead, as with standard IVs, my identification strategy is "variable-based", relying only on the properties of the semi-IVs. Thus, under very general conditions, identification with semi-IVs follows the same logic as with IVs.

Being able to obtain IV-like identification with "complementary invalid IVs" in nonseparable models is novel and is one of the main contributions of this paper and of Bruneel-Zupanc and Beyhum (2024). In the latter, we show that identification can be achieved by decomposing the exclusion and exogeneity properties of standard IVs into two other types of complementary invalid IVs, named "quasi-IVs": one that is excluded but possibly endogenous, and another that is exogenous but possibly included. A similar idea appears in Wang (2023), who proves that the LATE can be identified using two other (complementary) imperfect instruments: one violating the exclusion restriction and another violating monotonicity. A related idea is also found in Kolesár et al. (2015), who show that in linear models, identification can be achieved with many invalid (included) instruments, provided that their direct effects on outcomes are uncorrelated with their effects on the endogenous (continuous)

regressor (which could be interpreted as a specific form of complementarity). Semi-IVs are another class of "complementary invalid IVs", which violate the full exclusion restriction but still satisfy some targeted exclusion restrictions from specific potential outcomes.

B Heckman and Vytlacil (2005) with semi-IVs

In this Appendix, I show how the Framework of Section 2 can alternatively be written as a direct adaptation of the IV Framework of Heckman and Vytlacil (2005). The main expositional difference is that I explicitly write the potential outcome equations here.

The potential outcomes and the selection are modeled as follows

$$Y_d = q_d(X, Z_d, U_d) \qquad \text{for } d \in \mathcal{D}, \tag{25}$$

$$D = \mathbb{1}\{g(X, Z) \geq V\}, \tag{26}$$

where the semi-IVs, $Z = (Z_0, Z_1)$, and the covariates, X, are observed, while the "shocks", (U_0, U_1, V), are unobserved. The observables are Y, D, X, and Z, meaning that all the semi-IVs are observed, regardless of the selected alternative. The scalar random variable V may be correlated with $U = (U_0, U_1)$, yielding the endogeneity of D with respect to the potential outcomes. The potential outcomes are thus determined as general nonseparable functions (q_d) of unrestricted shocks U_d (which may contain several unobservables), which allows for rich forms of observed and unobserved heterogeneity, with (possibly) heterogenous treatment effects, general selection bias, and possible selection on gains. With the nonseparable potential outcome function, the semi-IVs can affect their corresponding potential outcomes in a completely unrestricted manner, also allowing for heterogenous effects of the semi-IVs on the potential outcomes from which they are not excluded.

The model can be described by the following set of assumptions.

Assumption 4 (Valid semi-IV model)**.** *The following conditions hold jointly*

A4.1 The random vectors (U_1, V) are independent of Z_0 conditional on Z_1 and X.
A4.2 The random vectors (U_0, V) are independent of Z_1 conditional on Z_0 and X.
A4.3 The term $g(Z_0, Z_1, X)$ is a nondegenerate random variable conditional on Z_1 and X.
A4.4 The term $g(Z_0, Z_1, X)$ is a nondegenerate random variable conditional on Z_0 and X.
A4.5 The distribution of V is absolutely continuous with respect to Lebesgue measure.
A4.6 $E|Y_1|$ and $E|Y_0|$ are finite.
A4.7 $1 > Pr(D = 1|X) > 0$.

These assumptions are a direct adaptation of the IV assumptions of Heckman and Vytlacil (2005) to the semi-IVs. As with IVs, they are equivalent to the main assumptions in Section

2 (Vytlacil, 2002). Indeed, the exclusion restriction holds because Z_0 is excluded from Y_1 in Equation (25), and the exogeneity holds because of Assumption A4.1. Both of these imply that Z_0 is "excluded" from Y_1 conditional on Z_1 and X (Assumption A1.1): because it does not enter directly through the functional form, nor through the error term U_1 thanks to the independence. Similarly, assumption A4.2 combined with Equation (25) imply that Z_1 is excluded from Y_0 conditional on Z_0 and X.

A stronger, but often natural assumption, is to assume that

$$(U_0, U_1, V) \perp\!\!\!\perp (Z_0, Z_1)|X.$$

This is equivalent to assuming that $(Y_0(z_0, x), Y_1(z_1, x), D(z)) \perp\!\!\!\perp (Z_0, Z_1)|X$, as in Footnote 8. This assumption is slightly stronger than the previous one, as it restricts how the semi-IVs affect their respective potential outcomes: they can only affect them through direct effects and not through their correlation with the unobserved U_d.

The index model (26) for the selection is the same as the index model (2) in Section 2, and is equivalent to the Monotonicity assumption 2 of Imbens and Angrist (1994). Assumptions A4.3 and A4.4 are the relevance assumptions of Z_0 (resp. Z_1) conditional on Z_1 (resp. Z_0) and X. Assumption A4.5 is also assumed in the main text (with a normalization).

Finally, Assumptions A4.6 and A4.7 are merely regularity conditions that ensure that the treatment parameters are well defined and that one observes a control and a treatment group for all X.

C LATE (Imbens and Angrist, 1994) with semi-IVs

Local Average Treatment Effect (LATE). For any $X = x$, for any $z = (z_0, z_1)$ and $z' = (z'_0, z'_1) \in \mathcal{Z}$, with $P(z) = p < P(z') = p'$, define the counterpart of Imbens and Angrist (1994)'s LATE with semi-IVs as

$$\Delta_{LATE}(z, z', x) = \mathbb{E}[\, Y_1(z'_1, x) - Y_0(z'_0, x) \mid D(z', x) = 1, D(z, x) = 0, X = x \,].$$

In other words, this LATE is the mean gain for individuals (compliers) induced to switch from $D = 0$ to $D = 1$ by an exogenous change in Z from z to z'.[42] The LATE is naturally decomposed into two local average treatment responses (LATR_d) defined as

$$\Delta_{LATR_d}(z, z', x) = \mathbb{E}[\, Y_d(z'_d, x) \mid D(z', x) = 1, D(z, x) = 0, X = x \,] \quad \text{for each } d = 0, 1,$$

[42] One could arbitrarily define the LATE differently, by comparing $Y_1(z_1) - Y_0(z_0)$, $Y_1(z'_1) - Y_0(z_0)$ or $Y_1(z_1) - Y_0(z'_0)$ instead. I choose this definition, which represents the treatment effect on the compliers in the new environment after the change in semi-IVs. Under this definition, the initial environment, $Z = z$, only serves to determine the initial propensity score, $P(z)$. This choice is arbitrary and not impactful since all these alternative parameters are identified as well.

such that $\Delta_{LATE}(z,z',x) = \Delta_{LATR_1}(z,z',x) - \Delta_{LATR_0}(z,z',x)$.

Note that in terms of the general LATE and LATR$_d$ definitions of Section 3, we have

$$\Delta_{LATE}(z,z',x) = \text{LATE}(p,p',z_0',z_1',x), \quad \text{and} \quad \Delta_{LATR}(z,z',x) = \text{LATR}(p,p',z_d',x),$$

since individuals with $D(z',x) = 1; D(z,x) = 0$ are the compliers with $p \leq V < p'$. Thus, $\Delta_{LATE}(z,z',x)$ can be identified as we identify LATE(p,p',z_0,z_1',x), by comparing four key points. Let us see how to achieve this result alternatively, by adapting the reasoning of Imbens and Angrist (1994) to the presence of semi-IVs instead.

Identification. We want to identify $\Delta_{LATE}(z,z')$ using data on (D,Y,Z_0,Z_1). For any $z = (z_0,z_1)$ and $z' = (z_0',z_1') \in \mathcal{Z}$, with $P(z) = p < P(z') = p'$, we have

$$\begin{aligned}&\mathbb{E}\big[Y|Z=z'\big] - \mathbb{E}\big[Y|Z=z\big] \\ &= \mathbb{E}[DY|Z=z'] - \mathbb{E}[DY|Z=z] \;+\; \mathbb{E}[(1-D)Y|Z=z'] - \mathbb{E}[(1-D)Y|Z=z]. \end{aligned} \tag{27}$$

Treated $\Delta_{LATR_1}(z,z')$. First, focus on the treated sample (terms in DY). We have:

$$\begin{aligned}&\mathbb{E}[DY|Z=z'] - \mathbb{E}[DY|Z=z] \\ &= \mathbb{E}[Y_1 D(z')|Z=z'] - \mathbb{E}[Y_1 D(z)|Z=z] \\ &= \mathbb{E}[Y_1 D(z')|Z_1=z_1'] - \mathbb{E}[Y_1 D(z)|Z_1=z_1] \\ &= \underbrace{\mathbb{E}[Y_1(D(z')-D(z))|Z_1=z_1']}_{=\text{ selection effect given } Z_1=z_1'} \;+\; \underbrace{\mathbb{E}[Y_1 D(z)|Z_1=z_1'] - \mathbb{E}[Y_1 D(z)|Z_1=z_1]}_{=\text{ direct effect of } Z_1 \text{ from } z_1 \text{ to } z_1' \text{ when } D(z)=1}. \end{aligned} \tag{28}$$

The first equality holds because when $D=1$, $Y=Y_1$, and by rewriting D as the potential treatment $D(z)$. The second equality comes from the exclusion of Z_0 from Y_1 and $D(z)$ given Z_1, and the last equality is simply a rewriting where I added and subtracted $\mathbb{E}[YD(z)|Z_1 = z_1']$. In words, an exogenous change in Z from z to z' affects the average DY for two reasons. First, as it would be the case with IVs, the shift in the semi-IVs yields a selection/composition effect because the set of individuals who select $D=1$ increases (since $p' > p$). There are compliers who choose $D(z')=1$ while they were choosing $D(z)=0$ before the change. Second, because Z_1 changed from z_1 to z_1', we observe a direct effect of the semi-IV change on the outcomes, since Z_1 is not excluded from Y_1. In this case, we need to isolate the effect of Z_1 on the always takers, who choose $D(z)=1$ even before the semi-IV change. This is the main difference with standard IVs: to isolate the selection effect, we need to be able to identify this direct effect separately.

To do so, we need to find a point $\bar{z}^1 = (\tilde{z}_0, z_1')$ such that $P(\bar{z}^1) = P(z) = p$. Then, the

direct effect is identified by the comparison of

$$\begin{aligned}\mathbb{E}[DY|Z=\bar{z}^1]-\mathbb{E}[DY|Z=z] &= \mathbb{E}[Y_1D(z)|Z=\bar{z}^1]-\mathbb{E}[Y_1D(z)|Z=z]\\ &= \mathbb{E}[Y_1D(z)|Z_1=z_1']-\mathbb{E}[Y_1D(z)|Z_1=z_1], \end{aligned} \tag{29}$$

where we can drop the conditioning on Z_0 because of its exclusion from Y_1 and $D(z)$ given Z_1. This corresponds to the purple arrow in Figure 9.

Plugging (29) in (28), we can isolate the selection effect as

$$\begin{aligned}&\mathbb{E}[Y_1(D(z')-D(z))|Z_1=z_1']\\ &= \Big(\mathbb{E}[DY|Z=z']-\mathbb{E}[DY|Z=z]\Big)-\Big(\mathbb{E}[DY|Z=\bar{z}^1]-\mathbb{E}[DY|Z=z]\Big)\\ &= \mathbb{E}[DY|Z=z']-\mathbb{E}[DY|Z=\bar{z}^1]. \end{aligned} \tag{30}$$

In other words, the selection effect for Y_1 (close to $\Delta_{LATR_1}(z,z')$) is obtained by comparing the outcomes in the treated sample given $Z=z'$ to the ones given $Z=\bar{z}^1$. Note that it corresponds exactly to the two points comparison that would be used to identify $\text{LATR}_1(p,p',z_1')$ for the generalized LATR/LATE: a shift in the propensity score P from p to p' by shifting the excluded semi-IV Z_0 from $\tilde{z}_0$ to z_0 while holding $Z_1=z_1'$ fixed.

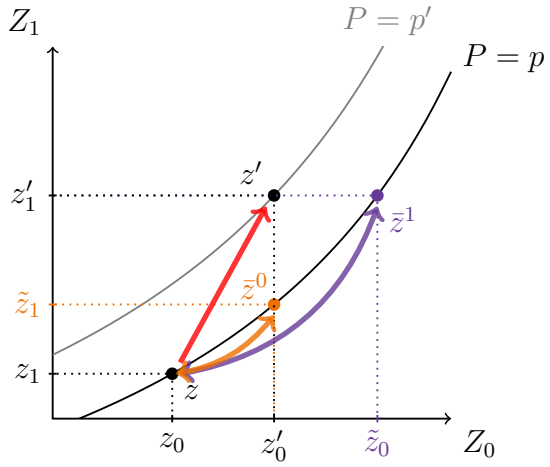


Figure 9: Identification of $\Delta_{LATE}(z,z')$

Untreated $\Delta_{LATR_0}(z,z')$. Similarly, on the untreated sample (terms in $(1-D)Y$), we have

$$\begin{aligned}&\mathbb{E}[(1-D)Y|Z=z']-\mathbb{E}[(1-D)Y|Z=z] \\ &= -\underbrace{\mathbb{E}[Y_0(D(z')-D(z))|Z_0=z_0']}_{:=\text{ selection effect given } Z_0=z_0'}+\underbrace{\mathbb{E}[Y_0(1-D(z))|Z_0=z_0']-\mathbb{E}[Y_0(1-D(z))|Z_0=z_0]}_{:=\text{ direct effect of } Z_0 \text{ from } z_0 \text{ to } z_0' \text{ when } D(z)=0}. \end{aligned} \tag{31}$$

To isolate the direct effect of Z_0, we need to find a point $\bar{z}_0=(z_0',\tilde{z}_1)$ such that $P(\bar{z}_0)=$

$P(z) = p$ (cf the orange arrow in Figure 9), such that,

$$\begin{aligned}&\mathbb{E}[(1-D)Y|Z=\bar{z}_0] - \mathbb{E}[(1-D)Y|Z=z]\\&= \mathbb{E}[Y_0(1-D(z))|Z_0=z_0'] - \mathbb{E}[Y_0(1-D(z))|Z_0=z_0]. \end{aligned} \tag{32}$$

Plugging (32) into (31), we can isolate the selection effect for the untreated as

$$\begin{aligned}&-\mathbb{E}[Y_0(D(z')-D(z))|Z_0=z_0]\\&= \Big(\mathbb{E}[(1-D)Y|Z=z'] - \mathbb{E}[(1-D)Y|Z=z]\Big) - \Big(\mathbb{E}[(1-D)Y|Z=\bar{z}^0] - \mathbb{E}[(1-D)Y|Z=z]\Big)\\&= \quad \mathbb{E}[(1-D)Y|Z=z'] - \mathbb{E}[(1-D)Y|Z=\bar{z}^0]. \end{aligned} \tag{33}$$

In other words, the selection effect for Y_0 (close to minus $\Delta_{LATR_0}(z,z')$) is obtained by comparing the outcomes in the untreated sample given $Z=z'$ to the ones given $Z=\bar{z}^0$. Note that it corresponds exactly to the two points comparison that would be used to identify $\text{LATR}_0(p,p',w_0)$ for the generalized LATR/LATE: a shift in the propensity score P from p to p' by shifting the excluded semi-IV Z_1 from $\tilde{z}_1$ to z_1' while holding $Z_0=z_0'$ fixed.

LATE. To obtain the $\Delta_{LATE}(z,z')$, notice that, because of the exclusion of the semi-IVs,

$$\begin{aligned}&\mathbb{E}[Y_1(D(z')-D(z))|Z_1=z_1'] = \mathbb{E}[Y_1(D(z')-D(z))|Z_1=z_1', Z_0=z_0'] \quad \text{for any } z_0',\\ \text{and}\quad &\mathbb{E}[Y_0(D(z')-D(z))|Z_0=z_0'] = \mathbb{E}[Y_0(D(z')-D(z))|Z_0=z_0', Z_1=z_1'] \quad \text{for any } z_1'.\end{aligned}$$

Thus, we can rewrite and add (30) with (33) to obtain:

$$\begin{aligned}&\mathbb{E}[(Y_1-Y_0)(D(z')-D(z))|Z_0=z_0', Z_1=z_1']\\&= \mathbb{E}[DY|Z=z'] - \mathbb{E}[DY|Z=\bar{z}^1] + \mathbb{E}[(1-D)Y|Z=z'] - \mathbb{E}[(1-D)Y|Z=\bar{z}^0].\end{aligned}$$

Since the first term is equal to $\Delta_{LATE}(z,z')(p'-p)$, the LATE is identified by

$$\begin{aligned}\Delta_{LATE}(z,z') &= \frac{\mathbb{E}[DY|Z=z'] - \mathbb{E}[DY|Z=\bar{z}^1] + \mathbb{E}[(1-D)Y|Z=z'] - \mathbb{E}[(1-D)Y|Z=\bar{z}^0]}{p'-p}\\&= \frac{\mathbb{E}[Y|Z=z'] - \mathbb{E}[DY|Z=\bar{z}^1] - \mathbb{E}[(1-D)Y|Z=\bar{z}^0]}{p'-p},\end{aligned}$$

if both $\bar{z}^1 = (\tilde{z}_0, z_1')$ with $P(z_1)=P(z)$ and $\bar{z}^0 = (z_0', \tilde{z}_1)$ with $P(\bar{z}_0)=P(z)$ exist. This formula corresponds exactly to the same two by two points comparisons that would be used to identify the generalized $\text{LATE}(p,p',z_0',z_1')$. Additionally here, given the specific parameter definition, we compare two different points, $\bar{z}^0$ and $\bar{z}^1$, to the same z', requiring only 3 different points/environments in total. The definition of the LATE and its identification are strongly tied to $Z=z'$. On the contrary, the original point, z, only pins down $p=P(z)$, but its underlying values, (z_0,z_1) are not important otherwise.

D Estimation: Robinson double residual regression

This appendix details Robinson (1988)'s double residual regression procedure estimating the partially linear model of Equation (22) on the treated and untreated subsamples, i.e.,

$$\mathbb{E}[Y|D=d,Z_d,P,X] = X\beta_d + Z_d\delta_d + \kappa_d(P), \qquad \text{for } d=0,1.$$

As the counterpart estimation of MTE with IVs, the estimation consists of three steps.

1^{st} stage: propensity score. Estimate the propensity score $P(Z_0,Z_1,X)$ by regressing D on (Z_0,Z_1,X). I use a probit model for this stage, but linear probability models, logit, or nonparametric alternatives are applicable. One can check here whether the semi-IVs are relevant or not. If they are, one can proceed using the fitted propensity score, $\hat{P}$, as an observed variable in the following stages.

2^{nd} stage: direct effect of the semi-IVs (and covariates). On the subsample with $D=d$, estimate $\widehat{\mathbb{E}}[Y|D=d,\hat{P}]$, $\widehat{\mathbb{E}}[X|D=d,\hat{P}]$ and $\widehat{\mathbb{E}}[Z_d|D=d,\hat{P}]$ for all X and Z_d variables, using local (linear) regressions of Y,X, and Z_d on $\hat{P}$. Then, create the residuals, $r^d_{variable} = variable - \widehat{\mathbb{E}}[variable|D=d,\hat{P}]$ for each variable Y,Z_d, and X. Notice that, according to (19), we have

$$\begin{aligned} &Y_d - \mathbb{E}[Y_d|D=d,P] \\ &= (X-\mathbb{E}[X|D=d,P])\beta_d + (Z_d-\mathbb{E}[Z_d|D=d,P])\delta_d + \overbrace{U_d - \mathbb{E}[U_d|D=d,P]}^{:=\epsilon}, \end{aligned}$$

with $\mathbb{E}[\epsilon|r^d_X, r^d_{Z_d}]=0$ by the separability and independence assumptions. So, on the subsample $D=d$, the linear regression (without intercept) of

$$r^d_Y = r^d_X\beta_d + r^d_{Z_d}\delta_d + \epsilon,$$

provides consistent estimates of the effect of the covariates, $\hat{\beta}_d$, and of the semi-IVs Z_d, $\hat{\delta}_d$, on Y_d. The role of the excluded semi-IV, Z_{1-d} is to avoid perfect colinearity between Z_d and P here, i.e., to have variations in P at fixed Z_d.

3^{rd} stage: $k_d(v)$. Once β_d and δ_d are estimated for both $d=0,1$, construct

$$\tilde{Y}_i = Y_i - (X_i\hat{\beta}_d - Z_{di}\hat{\delta}_d))\mathbb{1}\{D_i=d\},$$

i.e., the outcome net of the effects of the covariates and semi-IVs. If $\hat{\beta}_d$ and $\hat{\delta}_d$ are consistently estimated, by analogy, using (19), we have that

$$\tilde{Y}_i = D_i\widehat{U}_{1i} + (1-D_i)\widehat{U}_{0i},$$

i.e., $\tilde{Y}$ provides consistent estimates of the unobservables U_d when $D = d$. Then, for the treated individuals we have

$$\mathbb{E}[D\tilde{Y}|\hat{P} = p] = p\ \mathbb{E}[U_1|D = 1, \hat{P} = p] = p\ \mathbb{E}[U_1|V \leq p] = p \int_0^p \mathbb{E}[U_1|V = v]\frac{1}{p}dv = \int_0^p k_1(v)dv.$$

Consequently, for any v in the interior of the support of $\hat{P}$ given $D = 1$,

$$\frac{\partial \mathbb{E}[D\tilde{Y}_i|\hat{P} = p]}{\partial p}\Big|_{p=v} = \widehat{k_1}(v).$$

Similarly, for any v in the interior of the support of $\widehat{P}$ given $D = 0$,

$$-\frac{\partial \mathbb{E}[(1-D)\tilde{Y}_i|\hat{P} = p]}{\partial p}\Big|_{p=v} = \widehat{k_0}(v).$$

So to estimate $k_d(v)$, estimate $\mathbb{E}[\tilde{Y}_i \mathbb{1}\{D = d\}|\hat{P} = p]$ and its derivative using local quadratic regressions of $\tilde{Y}D$ and $\tilde{Y}(1-D)$ on $\widehat{P}$.[43]

Notice that the identification and estimation in the separable model (19) is very close to the general identification proof. The main advantage of separability is that $k_d(v)$ does not depend on z_d (or x). As a result, one can identify $k_d(v)$ for any v in the interior of the support of $\widehat{P}$ given $D = d$, instead of only on the support of $\widehat{P}$ given $D = d$ and $Z_d = z_d$ and $X = x$ as in the general proof.[44]

[43] Following Fan and Gijbels (1996), in order to estimate the first order derivative I specify a local polynomial of the order of the derivative +1, hence a local quadratic regression. The `semiIVreg` package includes several options for the kernel and bandwidth choices. The default option is to use the Gaussian kernel and to select the bandwidth according to the mean-squared error direct plug-in criteria of Fan and Gijbels (1996), as computed by Calonico et al. (2019) (`nprobust` package).

[44] In the general identification proof, I never specified that the support of P was taken conditional on $D = d$. This is because, in terms of identification, if $0 < P < 1$, we should always observe some individuals with $D = 1$ and some with $D = 0$. With small samples, this is not necessarily the case, hence the additional precision for the estimation.

E Monte Carlo Simulations

E.1 Heterogenous treatment effect

I simulate a simple generalized Roy model, close to the simulations of Heckman et al. (2006), but including semi-IVs.

E.1.1 Generalized Roy Model

Outcome and selection. I simulate the following generalized Roy model with heterogenous treatment effects:

$$\begin{aligned} Y_d &= \mu_d + Z_d\delta_d + U_d, \\ D &= \mathbb{1}\{\underbrace{\alpha + Z_0\alpha_0 + Z_1\alpha_1}_{:=\, g(Z)} - \tilde{V} \geq 0\}, \end{aligned}$$

with $\tilde{V} \perp\!\!\!\perp (Z_0, Z_1)$, and $\mathbb{E}[U_d|\tilde{V}, Z_0, Z_1] = \mathbb{E}[U_d|\tilde{V}]$ such that the semi-IVs are valid and even satisfy the stronger Separability assumption 3.

Semi-IVs. The semi-IVs are correlated and simulated as

$$\begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{Z_0} \\ \mu_{Z_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Z_0} & \sigma_{Z_0Z_1} \\ \sigma_{Z_0Z_1} & \sigma^2_{Z_1} \end{pmatrix}\right).$$

If $\delta_0 \neq 0$, Z_0 is not excluded from Y_0 and is thus not a valid IV. However, it is still excluded from Y_1 conditional on Z_1. Because of the correlation between the semi-IVs, it is important to condition on Z_1 for the exogeneity to hold. Conversely, Z_1 is not a valid IV if $\delta_1 \neq 0$; it is, however, well excluded from Y_0 conditional on Z_0. Despite its nonexclusion from Y_0, if $\alpha_0 \neq 0$, Z_0 is relevant and thus a valid semi-IV excluded from Y_1. Similarly, if $\alpha_1 \neq 0$, Z_1 is relevant and thus a valid semi-IV excluded from Y_0.

Error distribution. The joint distribution of U_0 and U_1 is given by:

$$\begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{U_0} & \sigma_{U_0U_1} \\ \sigma_{U_0U_1} & \sigma^2_{U_1} \end{pmatrix}\right)$$

Then, the (unnormalized) resistance to treatment, $\tilde{V}$, is defined as a function of U_1, U_0 and some other unobserved utility cost of treatment, C,

$$\tilde{V} = -(U_1 - U_0) + C,$$

where $C \sim \mathcal{N}(0, \sigma^2_{\text{cost}})$ and is independent of (U_0, U_1). The dependence of $\tilde{V}$ on U_1 and U_0 is the reason why the treatment D is endogenous in this model. Here $\tilde{V}$ directly includes the

comparison between U_1 and U_0, implying some *selection on gains*. Fixing U_0 and C, if U_1 is very large, then Y_1 is also large and the resistance to treatment $\tilde{V}$ is very low, meaning that the individual would be one of the first to select into treatment: thus $D = 1$ is very likely. A similar reasoning holds when U_0 and thus Y_0 is very low. Reciprocally, if U_1 is very low or if U_0 is very high, then $\tilde{V}$ is high, and the individual is not likely to select into treatment. If the utility cost of treatment, C, is high, $\tilde{V}$ is also high, and the individual is less likely to select into treatment.

By simple computations, we have that

$$\tilde{V} \sim \mathcal{N}(0, \underbrace{\sigma_{U_0}^2 + \sigma_{U_1}^2 - 2\sigma_{U_0U_1} + \sigma_C^2}_{:= \sigma_{\tilde{V}}^2}).$$

Since $\tilde{V}$ is normally distributed, the first stage is a probit model. As in the main text, normalize $\tilde{V}$ to $V = F_{\tilde{V}}(\tilde{V}) \sim \text{Unif}(0, 1)$, where V represents the ranks of the unobserved resistance to treatment, with a low V meaning a low resistance, and thus higher chances to select $D = 1$.

Theoretical MTR and MTE. Given the model specification, we can derive a closed form for the MTR and MTE. I do not use the knowledge of the true parametric distribution of the errors for the semi-parametric estimation, but it will be helpful to compare the estimation results to the truth. We have

$$k_d(v) = \mathbb{E}[U_d|V = v] = \mathbb{E}[U_d|\tilde{V} = F_{\tilde{V}}^{-1}(v)]$$

Since $(\tilde{V}, U_0, U_1)$ are trivariate normal it yields the closed-form for $k_d(v)$:

$$k_0(v) = \frac{\sigma_{U_0}^2 - \sigma_{U_0U_1}}{\sigma_{\tilde{V}}^2}\Big(F_{\tilde{V}}^{-1}(v) - \mu_{\tilde{V}}\Big),$$
$$\text{and} \quad k_1(v) = \frac{-\sigma_{U_1}^2 + \sigma_{U_0U_1}}{\sigma_{\tilde{V}}^2}\Big(F_{\tilde{V}}^{-1}(v) - \mu_{\tilde{V}}\Big),$$

where $\mu_{\tilde{V}} = 0$. Then, the MTR_d and MTE are given by:

$$m_0(v, z_0) = \mu_0 + z_0\delta_0 + k_0(v),$$
$$m_1(v, z_1) = \mu_1 + z_1\delta_1 + k_1(v),$$
$$\text{and} \quad MTE(v, z_0, z_1) = \mu_1 - \mu_0 + z_1\delta_1 - z_0\delta_0 + k_1(v) - k_0(v).$$

E.1.2 Monte Carlo simulations

Since the semi-IVs are valid, I can estimate the MTRs and MTE applying the semi-parametric estimation procedure described in Section 4.2. Figure 10 reports the estimation results from

Figure 10: Monte-Carlo simulation with heterogenous treatment effects.

Specification with $N = 10,000$ observations.

Parameters

Outcome: $\mu_0 = 3.2$, $\mu_1 = 3.6$, $\delta_0 = 1$, $\delta_1 = 1.3$.

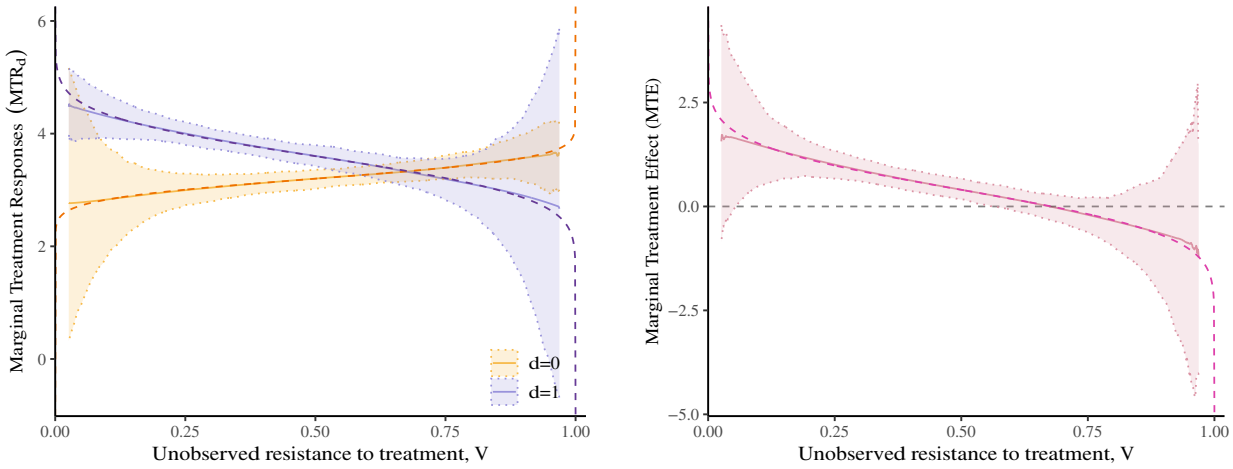
Selection: $\alpha = -0.2$, $\alpha_0 = -1.2$, $\alpha_1 = 1$.

Semi-IVs: $\mu_{Z_0} = 0$, $\mu_{Z_1} = 0$, $\sigma^2_{Z_0} = 1$, $\sigma^2_{Z_1} = 0.8$, $\sigma_{Z_0 Z_1} = 0.3$.

Errors: $\sigma^2_{U_0} = 1$, $\sigma^2_{U_1} = 1.5$, $\sigma_{U_0 U_1} = 0.5$, $\sigma^2_C = 1.5$.

	semi-IV		OLS	
Coefficients	Mean	(SD)	Mean	(SD)
Effects of the semi-IVs.				
Z_0: $\delta_0 = 1$	1.000	(0.022)	0.932	(0.015)
Z_1: $\delta_1 = 1.3$	1.300	(0.022)	1.217	(0.020)
Treatment (OLS).				
Intercept: $\mu_0 = 3.2$			3.395	(0.013)
D: $\mu_1 - \mu_0 = 0.4$			0.625	(0.022)

Notes: The OLS corresponds to the naive a regression of Y on a constant, D, DZ_1 and $(1-D)Z_0$. We do not report estimates of μ_0 nor μ_1 with semi-IVs since we do not estimate it directly.



Notes: Results of the estimation from 500 Monte Carlo simulations. The table provides the estimated coefficients for the linear part of the model (the direct effects of the semi-IVs), while the straight lines in the figures provide the estimated MTRs and MTE, taken at a reference individual with $(Z_0, Z_1) = (0, 0)$ (i.e., their averages). In this special case, $m_d(v, z_d = 0) = \mu_d + k_d(v)$. The colored areas represent the intervals where 95% of the estimates lie. The dashed lines represent the true theoretical MTRs and MTE. We use arbitrary bandwidths of 0.10 for both stages of the double residual regression. The results are robust to any bandwidth choice between 0.05 and 0.20. We trim the support of V to $[0.025, 0.970]$, such that the common support of $\hat{P}$ includes these in almost all simulations.

500 Monte Carlo simulations.

First, for the linear part of the semi-parametric model, the effects of the semi-IVs (δ_d) are precisely estimated. Not surprisingly, running a naive OLS assuming that the error term is independent of D yields biased estimates of both the effects of the semi-IVs and the difference in intercepts (treatment effects).

Second, for the nonparametric part of the model, I report the estimated MTRs and MTE at the mean semi-IVs, i.e., $(Z_0, Z_1) = (0, 0)$. The estimates are reported over the common support of the propensity given $D = 1$ and $D = 0$, which I trim to $[0.025, 0.97]$ such that almost all simulations have observations with $\hat{P}$ in this common support. The estimated MTRs and MTEs are unbiased compared to their theoretical counterpart. They are also precisely estimated with 95% of the estimates being in tight boundaries. This precision is remarkable because, with "only" $10,000$ observations, there are not so many observations in the neighborhood of any given $V = v$ in the treated or untreated subsamples. The semi-parametric estimation works well, even without making any assumption about the shape of the MTRs (especially the $k_d(v)$ functions) or about the distribution of errors. The estimates are noisier at the tails by construction of the estimator.

E.2 Homogenous treatment effect

E.2.1 Constant treatment effect model

Let us now simulate simpler models with homogenous treatment effects. The outcome and selection are given by

$$\begin{aligned} Y_d &= \mu_d + Z_d\delta_d + U, \\ D &= \mathbb{1}\{\alpha + Z_0\alpha_0 + Z_1\alpha_1 - \tilde{V} \geq 0\}, \end{aligned}$$

where $(U, V) \perp\!\!\!\perp (Z_0, Z_1)$ such that the semi-IVs are well excluded and satisfy a stronger form of separability than Assumption 3. Compared to the heterogenous treatment effect model, the main difference is that it is the same U that affects both potential outcomes. Think about it as a single unobservable, e.g., the unobserved ability. As a consequence, the individual treatment effects

$$Y_1 - Y_0 = \mu_1 - \mu_0 + Z_1\delta_1 - Z_0\delta_0,$$

are constant given (Z_0, Z_1), and independent of U. In this case, one can follow (23) and write a single outcome equation to be estimated:

$$Y = \mu_0 + D(\mu_1 - \mu_0) + DZ_1\delta_1 + (1 - D)Z_0\delta_0 + U. \tag{34}$$

Then, the MTE are also independent of V, and equal to the ATE for all v:

$$\text{ATE}(z_0, z_1) = \mathbb{E}[Y_1 - Y_0 | Z_0 = z_0, Z_1 = z_1] = \mu_1 - \mu_0 + z_1\delta_1 - z_0\delta_0.$$

In particular, $ATE(0,0) = \mu_1 - \mu_0$, i.e., the effect of D in Equation (34).

The semi-IV follow a bivariate normal distribution as in Section E.1:

$$\begin{pmatrix} Z_0 \\ Z_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{Z_0} \\ \mu_{Z_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{Z_0} & \sigma_{Z_0 Z_1} \\ \sigma_{Z_0 Z_1} & \sigma^2_{Z_1} \end{pmatrix} \right).$$

About the shocks, there is endogeneity because the outcome shock, U, and the unobservable resistance to treatment, $\tilde{V}$, are correlated as follows:

$$\begin{pmatrix} U \\ \tilde{V} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_U & \sigma_{U\tilde{V}} \\ \sigma_{U\tilde{V}} & \sigma^2_{\tilde{V}} \end{pmatrix} \right).$$

As usual, we normalize $V = F_{\tilde{V}}(\tilde{V})$.

E.2.2 Monte Carlo simulations

This model can be estimated using several approaches with semi-IVs. First, one can use the simple 2SLS approach described in Section 4.3. One can also estimate the model with the general semi-parametric (SP) approach of Section 4.2: in this case, we estimate heterogenous MTEs and MTRs varying with V (at the reference individual with $(Z_0, Z_1) = (0,0)$), and then compute their averages by integrating $\text{MTE}(v,0,0)$ over the support of P (the range of the support should not matter since the true treatment effects are assumed to be homogenous here). The averages of $m_1(v,0), m_0(v,0)$ and $\text{MTE}(v,0,0)$ provide an estimate of μ_1, μ_0 and $\mu_1 - \mu_0 = \text{ATE}(0,0)$, respectively. As for the effects of the semi-IVs, they are already estimated as a single coefficient in the semi-parametric approach. This procedure has the advantage of being robust when treatment effects are heterogenous. Unfortunately, this comes at the cost of some efficiency loss.

A better approach is to impose the homogeneity constraint on the treatment effects in the semi-parametric estimation (SP-H). To do so, note that, since $(U,V) \perp\!\!\!\perp (Z_0, Z_1)$, we have $(U,V) \perp\!\!\!\perp P$ since $P = P(Z)$, and we can write

$$\mathbb{E}[U] = \mathbb{E}[U|P] = \mathbb{E}[U|D=1,P]P + \mathbb{E}[U|D=0,P](1-P).$$

Denote $\kappa_d(P) = \mathbb{E}[U|D=d,P]$. Under constant treatment effects we have[45]

$$\kappa_0(P) = -\kappa_1(P)\frac{P}{1-P}. \tag{35}$$

[45] One can easily show that in this case, $k_1(p) = \kappa_1(p) + p\kappa_1'(p)$ and $k_0(p) = \kappa_0(p) - (1-p)\kappa_0'(p)$ are equal by construction, which is expected since they both represent the same $\mathbb{E}[U|V=p]$.

As a consequence, Equation (22) can be written as

$$\mathbb{E}[Y|D=0,Z_0,P] = \mu_0 + \delta_0 Z_0 + \kappa_0(P) = \mu_0 + \delta_0 Z_0 - \kappa_1(P)\frac{P}{1-P},$$
$$\text{and} \quad \mathbb{E}[Y|D=1,Z_1,P] = \mu_1 + \delta_1 Z_1 + \kappa_1(P).$$

Pooling these two, one can estimate a single model for the outcome variable Y, i.e.,

$$\begin{aligned}\mathbb{E}[Y|D,Z_0,Z_1,P] &= D\mathbb{E}[Y|D=1,Z_1,P] + (1-D)\mathbb{E}[Y|D=0,Z_0,P] \\ &= \mu_0 + D(\mu_1-\mu_0) + DZ_1\delta_1 - (1-D)Z_0\delta_0 + \Big(D - (1-D)P/(1-P)\Big)\kappa_1(P).\end{aligned}$$

So, by regressing Y on a constant, D, DZ_1, $(1-D)Z_0$ and a flexible function of P multiplied by $(D-(1-D)P/(1-P))$, we estimate an homogenous treatment effect model. The *homogeneity constraint* is that the same $\kappa_1(P)$ is estimated for all D (treated or not). I specify this flexible function $\kappa_1(P)$ as a polynomial function, in the spirit of nonparametric sieve estimation. The advantage of the pooling is that the MTRs are estimated using all the information and not just the separate subsamples. Thus, I take full advantage of the knowledge that $k_0(v)$ must be equal to $k_1(v)$ under homogeneity.

Results. Table 3 reports the coefficients from the model estimated using the three different semi-IV methods discussed above, and compares them to the estimates of a naive OLS and of an IV regression using the semi-IVs as valid IVs. I estimate two specifications: the first one is a general semi-IV model with homogenous treatment effect, and the second one is a special case where the semi-IVs are in fact excluded ($\delta_0 = \delta_1 = 0$) and could thus be used as valid IVs. This specification provides an idea of the loss of efficiency one faces by being too cautious and considering a valid IV is only partially excluded (by considering it as a semi-IV).

Overall, all the estimation methods yield unbiased estimates of the parameters. However, using the general approach (SP), which is robust to heterogenous treatment effects, comes at the cost of reduced efficiency. The simple semi-IV 2SLS is more efficient than the general approach. The semi-parametric approach with homogeneity constraint (SP-H) outperforms the two other alternatives. The efficiency gains stem from the fact that we pool all the observations to estimate the control function $\kappa_1(P)$. Perhaps surprisingly, in the case where the semi-IVs are, in fact, valid IVs, using the SP-H approach causes almost no efficiency loss with respect to the standard IV-2SLS using Z_0 and Z_1 as IVs. The minor difference in standard errors is due to the fact that the semi-IV approach has to estimate two additional parameters for each semi-IV. Given how large the bias of the IV regression can be if the semi-IVs are invalid IVs (cf Specification 1), being cautious and considering them as semi-

Table 3: Monte Carlo simulations with homogenous treatment effects

Specification: with $N = 10,000$ observations.

Parameters

Outcome: $\mu_0 = 3.2$, $\mu_1 = 3.6$,
Specification 1: $\delta_0 = 0.8$, $\delta_1 = 0.5$; Specification 2: $\delta_0 = \delta_1 = 0$.

Selection: $\alpha = 0$, $\alpha_0 = -0.7$, $\alpha_1 = 0.7$.

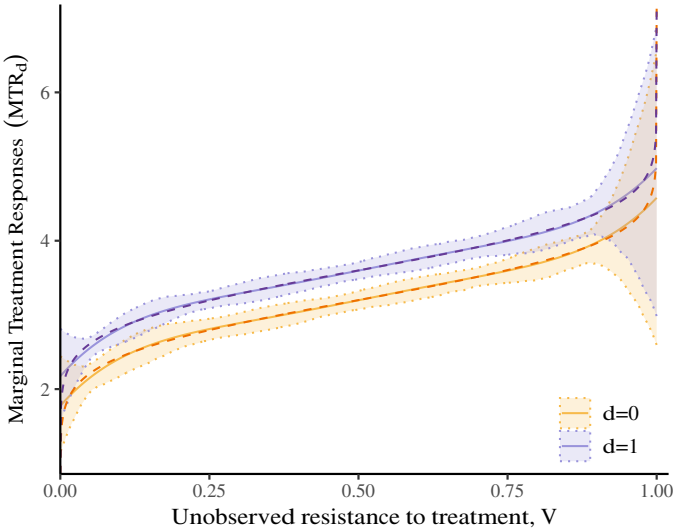
Semi-IVs: $\mu_{Z_0} = 0$, $\mu_{Z_1} = 0$, $\sigma^2_{Z_0} = 1$, $\sigma^2_{Z_1} = 1$, $\sigma_{Z_0 Z_1} = 0.5$.

Errors: $\sigma^2_U = 1$, $\sigma^2_{\tilde{V}} = 1.5$, $\sigma_{U\tilde{V}} = 0.6$.

	True	OLS	IV	semi-IV		
				(2SLS)	(SP)	(SP-H)
Specification 1:						
Intercept: μ_0	3.2	3.614	3.515	3.202	3.196	3.199
		(0.013)	(0.041)	(0.037)	(0.091)	(0.027)
D: $\mu_1 - \mu_0$	0.4	-0.428	0.068	0.402	0.405	0.401
		(0.019)	(0.074)	(0.071)	(0.123)	(0.049)
$Z_0(1-D)$: δ_0	0.8	0.706		0.788	0.800	0.801
		(0.013)		(0.028)	(0.016)	(0.015)
$Z_1 D$: δ_1	0.5	0.595		0.486	0.501	0.500
		(0.013)		(0.028)	(0.016)	(0.015)
Specification 2: $\delta_0 = \delta_1 = 0$						
$\rightarrow$ *semi-IVs are valid IVs.*						
Intercept: μ_0	3.2	3.613	3.200	3.199	3.205	3.199
		(0.013)	(0.023)	(0.036)	(0.090)	(0.026)
D: $\mu_1 - \mu_0$	0.4	-0.427	0.400	0.400	0.393	0.401
		(0.019)	(0.042)	(0.070)	(0.129)	(0.048)
$Z_0(1-D)$: δ_0	0	-0.094		0.000	0.000	0.000
		(0.012)		(0.027)	(0.016)	(0.014)
$Z_1 D$: δ_1	0	0.095		0.000	0.000	0.000
		(0.014)		(0.029)	(0.017)	(0.016)
N = 10,000 observations for every simulations						

Notes: The OLS column corresponds to the estimation of (34) by OLS, i.e., wrongly assuming that D is exogenous such that $\mathbb{E}[U|D, Z_0, Z_1] = 0$. The IV column corresponds to the 2SLS of Y on D, using Z_0 and Z_1 as standard IVs. Then, I implement three different semi-IV estimation procedures. The semi-IV 2SLS corresponds to the simple 2SLS method described in Section 4.3, with a linear probability model of D on Z_0 and Z_1 in the first stage (results are similar if one uses the correct probit instead). The semi-IV "SP" column corresponds to the semi-parametric estimation of Section 4.2 and of Appendix E.1, i.e., I estimate the model with heterogenous treatment effects (bandwidth arbitrarily set to 0.10, and trimming the 1% top and bottom of the support of P for treated and untreated), and then compute the ATE as the average of the $\widehat{MTE}(v, 0, 0)$ and the "intercept" as the average of $\widehat{m}_0(v, 0)$. Finally, the last semi-IV "SP-H" corresponds to the "semi-parametric approach with homogeneity constraint" estimation approach, which imposes that the treatment effects must be homogenous by imposing the relation (35) between κ_0 and κ_1.

Figure 11: Semi-parametric estimation with homogeneity constraint (SP-H) of the MTRs



Notes: Results of the estimation of the MTRs, at $(Z_0, Z_1) = (0, 0)$, from 500 Monte Carlo simulations of the first specification (Table 3), using the semi-parametric estimation with homogeneity constraint on the treatment effect. By construction $m_1(v, 0) = m_0(v, 0) + \mu_1 - \mu_0$. We specify the control function $\kappa_1(P)$ as a polynomial of order 5 function of P. Results are almost insensitive to polynomial orders between 2 and 5.

IVs in case of doubt may be recommended.

In the event where the semi-IVs are valid IVs, note that the semi-IV approaches very precisely estimate that the effects of the semi-IVs on the outcome is null. This suggests that semi-IVs can be used to test the validity of some IVs (although it requires having another complementary valid semi-IV).

Finally, in Figure 11, I report the estimation of the MTRs, at $(Z_0, Z_1) = (0, 0)$, using the semi-parametric estimation approach with homogeneity constraint (SP-H). These estimates are unbiased and very precise. Even with homogenous treatment effects, one should report the estimates of the MTRs. Indeed, it provides crucial information about the correlation between V and U, i.e., about the relation between the unobservables in the outcome and the selection (positively related in the specification of Figure 11).

E.3 Replicability

All the simulations of this section have been implemented using the companion `semiIVreg` package (Bruneel-Zupanc, 2024) and its `simul_data()` function to simulate generalized Roy models. See the package webpage for more details on how to proceed and replicate these simulations.

F (Unordered) Discrete Treatment D

The model with binary treatment can be extended to problems with (unordered) discrete treatment, $D \in \mathcal{D} = \{0, ..., J-1\}$ with $J > 2$ mutually exclusive alternatives. As in the binary case, one can nonparametrically identify similar treatment effects (pooled LATE, margin-specific MTE) with semi-IVs as with IVs. The underlying logic remains the same as in the binary case: by shifting the excluded semi-IVs, one can vary the set of compliers while keeping the included semi-IV fixed.

F.1 The model with discrete D

Potential outcomes. As in the binary case, there is one potential outcome per alternative, Y_d, but one only observes the outcome,

$$Y = Y_D = \sum_{d=0}^{J-1} Y_d \mathbb{1}\{D = d\}.$$

Alternative-specific semi-IVs. We have J different (one per distinct value of $\mathcal{D}$) continuously distributed *alternative d-specific semi-IVs*, Z_d. Denote $Z = (Z_0, Z_1, ..., Z_{J-1})$ the observed set of all semi-IVs. For all $j \in \mathcal{D}$, define D_j a dummy = 1 if $D = j$, 0 otherwise. Define potential treatments as $D_j(Z)$. Similarly, denote $P_d(Z, X) = \mathbb{E}[D_j|Z, X] = \Pr(D = j|Z, X)$, and $\mathbf{P}(Z, X) = (P_0(Z, X), P_1(Z, X), ..., P_{J-1}(Z, X))$ the generalized propensity score, i.e., the set of all d-specific selection probabilities, which sum to 1. The semi-IVs are valid if they satisfy the following generalization of Assumption 1 to J alternatives:[46]

Assumption 5 (Valid semi-IVs, discrete D)**.** *The semi-IVs satisfy the following conditions:*

A5.1 (Exogeneity and Exclusion) For all $z \in \mathcal{Z}$, (i) $D(z) \perp\!\!\!\perp Z|X$, (ii) $(Y_d, D(z)) \perp\!\!\!\perp Z_{-d}|(Z_d, X)$ where Z_{-d} is the set of all the semi-IVs, except Z_d.

A5.2 (Relevance) The generalized propensity score $\mathbf{P}(Z, X)$ is a nondegenerate random variable conditional on (Z_{-d}, X) for all $d \in \mathcal{D}$.

In other words, each semi-IV Z_d may, at most, not be excluded from one of the potential outcome, Y_d. Hence the name *d-specific semi-IVs*. In addition, they must also be relevant, i.e., have an effect on the selection probabilities. Now, we add some further restriction on the selection process.

[46] As in Section 2, I do not write the exclusion and exogeneity separately. The exclusion holds because $Y_d(z, x) = Y_d(z_d, x)$ and does not depend on any z_{-d}, for all $z \in \mathcal{Z}$.

Discrete choice index model. Following the identification of unordered discrete treatment with IVs (Heckman et al., 2006, 2008; Heckman and Vytlacil, 2007b), I extend the monotonicity to the discrete case by imposing a discrete choice index model.[47] The value of choosing $D = d$ for the individual is

$$R_d = u_d(Z_d, X) + \epsilon_d, \tag{36}$$

where u_d is the latent utility of choosing $D = d$, common to every individuals, and ϵ_d are the d-specific continuously distributed unobserved shocks. The ϵ_d can be correlated across alternatives. Note that the semi-IVs are also d-specific in the selection process here: Z_d should only affect u_d and not any other u_k (conditional on Z_k). Note that this d-specificity is also required to identify discrete treatment with d-specific IVs. The advantage is that, with semi-IVs, this d-specificity is a natural extension of the outcome exclusion to the selection and should be seamlessly satisfied by most semi-IVs.

The individual chooses the alternative that yields the highest utility, i.e.,

$$D = \underset{k \in \mathcal{D}}{\text{argmax}}\ R_k.$$

Since only $J - 1$ comparison of utilities matter for the discrete decision, D, we typically normalize the shocks ϵ with respect to ϵ_0, i.e., we define $V_d = \epsilon_d - \epsilon_0$ for $d \in \{1, ..., J-1\}$. Denote $\mathbf{V} = (V_1, ..., V_{J-1}) \in \boldsymbol{\mathcal{V}}$. Under this model, one can re-express Assumption 5 as:

Assumption 5'. *The semi-IVs Z satisfy the following properties*

A5'.1 (Exogeneity and Exclusion) (i) $\mathbf{V} \perp\!\!\!\perp Z|X$. *(ii)* $(Y_d, \mathbf{V}) \perp\!\!\!\perp Z_{-d}|(Z_d, X)$ *where Z_{-d} is the set of all the semi-IVs, except Z_d.*

A5'.2 (Relevance) For all $d \in \mathcal{D}$, $\partial u_d(z_d, x)/\partial Z_d \neq 0$ for all z_d, x.

For expositional simplicity, I further assume that $\partial u_d(z_d, x)/\partial Z_d > 0$ for all d, z_d, x, i.e., that the semi-IVs have a positive effect on their corresponding alternative utilities. In other words, think of Z_d as an incentive, a "push variable" in favour of D.[48]

[47] This discrete index formulation is not necessary to achieve identification with semi-IVs. For example, see how I identify margin-specific treatment effects as in Mountjoy (2022) using semi-IVs under only partial monotonicity with a comparable compliers assumption (Section F.3.5). However, the discrete index model (36) has nice properties and simplifies the exposition.

[48] This is an expositional normalization. If one has a semi-IV that deters participation into D, just focus on $-Z_d$ instead of Z_d. If a semi-IV has a relevant but ambiguous effect (sometimes positive, sometimes negative), the analysis still holds, but it complicates the exposition.

F.2 Example and comparison with discrete D using IVs

Most examples of semi-IVs discussed in Section 2.2 can naturally be extended to discrete choice models. For example, if one were to study occupation/sector choice with multiple sectors, the local market size (e.g., GDP) of each sector could serve as valid d-specific semi-IVs. It is likely that the characteristics of a specific sector only affect the utility of workers working in this sector, so the discrete choice structure is naturally satisfied. A similar reasoning applies more generally to many other discrete models, for example, discrete location choice models using local amenities as semi-IVs.

Note that the counterpart identification of discrete treatment with IVs (Heckman et al., 2006, 2008; Heckman and Vytlacil, 2007b; Kirkeboen et al., 2016; Mountjoy, 2022) generally uses several $(J-1)$ "d-specific IVs" or IVs with a larger support (of size J).[49] This limits the range of applications that identify the effect of discrete treatment with IVs because, even more so than in the binary case, it is difficult to find $J-1$ d-specific IVs, that affect the utility of only one alternative without affecting its corresponding potential outcome or the latent utility of the other alternatives. Fortunately, finding d-specific semi-IVs, which can also affect the d-specific outcome, is considerably easier. To identify the same treatment effects, researchers face a trade-off between finding $J-1$ d-specific IVs or J d-specific semi-IVs. As in the binary case, the semi-IVs are usually the same variable adapted across alternatives (e.g., sector-specific GDP) so finding one more semi-IV is mostly costless, making the semi-IV approach more naturally suited to the discrete treatment case.[50]

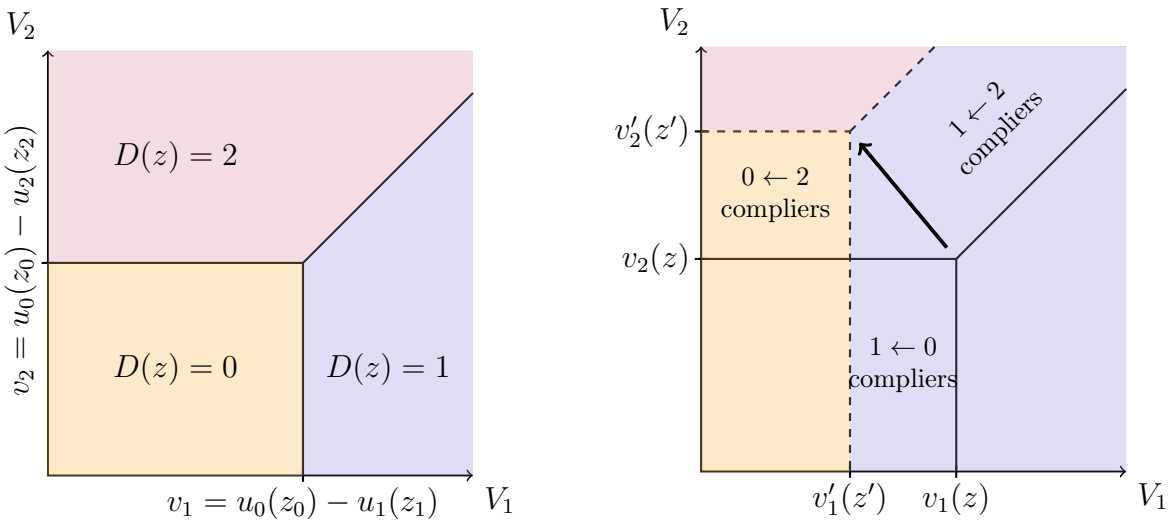
F.3 Identification

For expositional simplicity, omit X in the notation and proceed conditional on $X = x$. We have data on (Y, D, Z) and on the generalized propensity score variable $\mathbf{P} = \mathbf{P}(Z)$, because the function $\mathbf{P}(z)$ is nonparametrically identified for all $z \in \mathcal{Z}$. I first discuss the properties of the discrete choice model, then define the relevant treatment effects of interest with semi-IVs – the counterpart to those with IVs – and finally address their identification.

[49] Kline and Walters (2016) being an exception using a single binary assignment to the head start program as an IV in a discrete treatment setup.

[50] As in the binary case, it is also better if the semi-IVs are continuously distributed, while it is not always necessary for the IVs. However, the identification of marginal treatment effects with IVs also requires continuous IVs anyway (Mountjoy, 2022). Moreover, most examples of semi-IVs are, in fact, continuous.

Figure 12: Treatment choice as a partition



(a) Partition of D at $Z = z$

(b) Change in Z from z to z'

F.3.1 Identifying properties of the model

The multinomial choice model for the selection leads to a clear partitioning of the $J-1$ dimensional space of $\mathbf{V}$ into $D(z) = 0, 1, ..., J-1$, as illustrated in Figure 12 for $J = 3$.[51] From this partition, we have that $P_d(z) = \Pr(D(z) = d)$ which is the total mass of all V in the area with $D(z) = d$ (normalizing the total mass of $\boldsymbol{\mathcal{V}}$ to 1). This partitioning yields several key properties, similar to the ones we had in the binary treatment case, that can be leveraged for identification (even without identifying $u_d(z_d)$). I discuss these properties in the context of $J = 3$, but they hold more generally for any $J \geq 3$.

1. Generalized propensity score sufficient statistics. For any $z, z' \in \mathcal{Z}$ with $z' \neq z$, if $\mathbf{P}(z) = \mathbf{P}(z')$ then $D(z) = D(z')$ for all individuals.[52]

[51] The partitioning comes from the individuals comparing the different alternatives of the treatment and picking the one yielding the highest utility. Assuming ties break in favour of the lowest d, with the discrete index model, we have:

$$
\begin{array}{c|c|c}
D(z) = 0 & D(z) = 1 & D(z) = 2 \\
\Longleftrightarrow & \Longleftrightarrow & \Longleftrightarrow \\
\left\{ \begin{array}{ll} u_0(z_0) + \epsilon_0 & \geq u_1(z_1) + \epsilon_1, \\ u_0(z_0) + \epsilon_0 & \geq u_2(z_2) + \epsilon_2. \end{array} \right. & \left\{ \begin{array}{ll} u_1(z_1) + \epsilon_1 & > u_0(z_0) + \epsilon_0, \\ u_1(z_1) + \epsilon_1 & \geq u_2(z_2) + \epsilon_2. \end{array} \right. & \left\{ \begin{array}{ll} u_1(z_1) + \epsilon_1 & > u_0(z_0) + \epsilon_0, \\ u_1(z_1) + \epsilon_1 & \geq u_2(z_2) + \epsilon_2. \end{array} \right. \\
\Longleftrightarrow & \Longleftrightarrow & \\
\left\{ \begin{array}{ll} V_1 & \leq u_0(z_0) - u_1(z_1), \\ V_2 & \leq u_0(z_0) - u_2(z_2). \end{array} \right. & \left\{ \begin{array}{ll} V_1 & > u_0(z_0) - u_1(z_1), \\ V_2 & \leq u_1(z_1) - u_2(z_2) + V_1. \end{array} \right. & \left\{ \begin{array}{ll} V_2 & > u_0(z_0) - u_2(z_2), \\ V_2 & > u_1(z_1) - u_2(z_2) + V_1. \end{array} \right.
\end{array}
$$

This yields the partitioning observed in Figure 12.

[52] '*Proof*': as visible in Figure 12, the partitioning of $\boldsymbol{\mathcal{V}}$ into D is fully characterized by the two values,

In other words, as in the binary case, the generalized propensity score serves a sufficient statistic to summarize the effect of the semi-IVs on the selection, i.e., on the partition of $\boldsymbol{\mathcal{V}}$ into D. One could denote $D(\mathbf{P})$ instead of $D(Z)$, representing the choices of D given that the propensity score is $\mathbf{P}$. Then the property means that if $\mathbf{P}(z) = \mathbf{p}$, then $D(z) = D(\mathbf{p})$ for all individuals. Importantly, this discrete version of the index sufficiency property requires *all* the probabilities to be equal. Only observing that for some j, $P_j(z) = P_j(z')$, is not sufficient and does not bring useful information about the partitioning. This property is useful in order to isolate the direct effects of the semi-IVs net of the selection, by moving Z while holding $\mathbf{P}$ fixed. As in the binary case

$$\Delta_{Z_d}^{D=d}(z_d, z'_d, \mathbf{p}) = \mathbb{E}[Y_d|D = d, Z_d = z'_d, \mathbf{P} = \mathbf{p}] - \mathbb{E}[Y_d|D = d, Z_d = z_d, \mathbf{P} = \mathbf{p}],$$

is directly identified from the data provided one can move Z_d while holding $\mathbf{P}$ fixed, which is achievable by moving all the other semi-IVs as well.

2. Reachability without shifting Z_d. For any $d \in \mathcal{D}$, provided that all the semi-IVs are sufficiently relevant, for any $z = (z_0, ..., z_{J-1})$ in the interior of $\mathcal{Z}$ with $\mathbf{P}(z) = \mathbf{p}$, one can reach any $\mathbf{p}'$ in the neighbourhood of $\mathbf{p}$ without shifting $Z_d = z_d$, by shifting only the $J-1$ other semi-IVs.[53] In other words, as in the binary case, if the semi-IVs are relevant, one can shift $\mathbf{P}$ while holding a specific Z_d fixed. This permits the use of the excluded semi-IVs, Z_{-d}, as IVs for a specific Y_d, without having to shift Z_d. In general, one needs to shift all the other semi-IVs to mimic the change in a single semi-IV on the propensity score, hence the need for J semi-IVs in the discrete case.

$v_1 = u_0(z_0) - u_1(z_1)$ and $v_2 = u_0(z_0) - u_2(z_2)$. These two values give the coordinates of the "indifferent individual", i.e., the individual indifferent between every treatment alternative. By construction, there is only one point of indifference between the three alternatives in this model. So, the coordinates of the indifferent individual pin down the partitioning of $\mathcal{V}$, and, thus, the generalized propensity score $\mathbf{P}$. In fact, there is a one-to-one mapping between the indifferent individual coordinates and $\mathbf{P}$. Indeed, since all $V \in \boldsymbol{\mathcal{V}}$ have a non-negative mass, going from any (v_1, v_2) to (v'_1, v'_2) always leads to a different $\mathbf{P}$. This is straightforward graphically when $J = 3$: for any V in the interior of $\boldsymbol{\mathcal{V}}$, a shift to the top right necessarily increases P_0, a shift to the bottom right increases P_2, a shift to the top left increases P_1, and a shift to the bottom left reduces P_0. Now, the effects of Z on the partition is fully characterized by their effects on the latent d-specific utilities, and thus, on the coordinates of the indifferent individual. Since there is a one-to-one mapping between this and $\mathbf{P}$, the effect of any Z on the partitioning of D is fully characterized by the observable $\mathbf{P}$ it generates.

[53] '*Proof*': this is visible in Figure 12(b). A shift in Z from z to z' leads to a shift from $v_1(z) = u_0(z_0) - u_1(z_1)$ to $v_1(z') = u_0(z'_0) - u_1(z'_1)$ and from $v_2(z) = u_0(z_0) - u_2(z_2)$ to $v_2(z') = u_0(z'_0) - u_2(z'_2)$. Any $\tilde{z}'$ with $v_1(\tilde{z}') = v_1(z')$ and $v_2(\tilde{z}') = v_2(z')$ would yield the same $\mathbf{P} = \mathbf{p}'$. Imagine one wants to fix $Z_0 = z_0$, i.e., $\tilde{z}' = (z_0, \tilde{z}'_1, \tilde{z}'_2)$. At fixed $Z_0 = z_0$, $u_0(z_0)$ is fixed. But one can still adjust v_1 and v_2 by moving only u_1 and u_2 through moves of Z_1 and Z_2. Typically, picking $\tilde{z}'_1$ such that $u_1(\tilde{z}'_1) = u_0(z_0) - u_0(z'_0) + u_1(z'_1)$ yields $v_1(\tilde{z}') = v_1(z')$, and picking $\tilde{z}'_2$ such that $u_2(\tilde{z}'_2) = u_0(z_0) - u_0(z'_0) + u_1(z'_2)$ yields $v_2(\tilde{z}') = v_2(z')$. Similarly, if one wants to fix Z_1 or Z_2, one can shift only (Z_0, Z_2) or (Z_0, Z_1), respectively. This is feasible because the marginal individual is determined by $\mathbf{V}$, which is of dimension $J-1$, while Z is of dimension J and impacts J different u_d functions. Fixing a specific $Z_d = z_d$ is like normalizing all the u_j with respect to a specific alternative $u_d(z_d)$ baseline. This is feasible with this discrete choice structure with d-specific semi-IVs.

F.3.2 Margin-specific compliers

A fundamental challenge with multivalued treatment is that, as with IVs, a shift in the semi-IVs generates more complex selection patterns than in the binary case, resulting in multiple flows of *margin-specific compliers*. Let us define the "$d' \leftarrow d$ *compliers*" as

$$C_{d' \leftarrow d}(z, z') := Compliers_{d' \leftarrow d}(z, z') = \Big\{ \mathbf{V} \in \boldsymbol{\mathcal{V}} : D(z) = d, D(z') = d' \Big\},$$

i.e., the individuals who shift from $D = d$ when $Z = z$ to $D = d'$ when $Z = z'$. Using the sufficiency property of the propensity score, one can define these margin-specific compliers in terms of the generalized propensity score, $\mathbf{P}$, instead, i.e.,

$$C_{d' \leftarrow d}(\mathbf{p}, \mathbf{p}') = \Big\{ \mathbf{V} \in \boldsymbol{\mathcal{V}} : D(\mathbf{p}) = d, D(\mathbf{p}') = d' \Big\}.$$

Notice that, for any z, z' such that $\mathbf{P}(z) = \mathbf{p}$ and $\mathbf{P}(z') = \mathbf{p}'$, we have

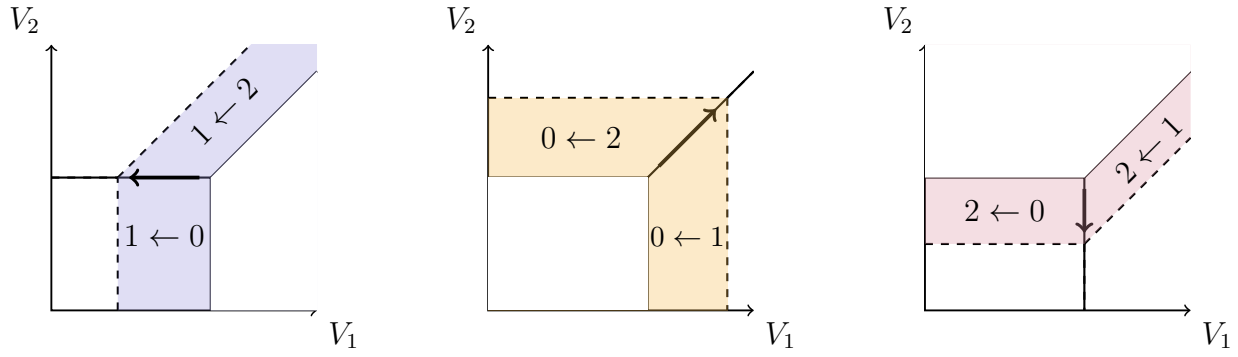
$$C_{d' \leftarrow d}(z, z') = C_{d' \leftarrow d}(\mathbf{P}(z), \mathbf{P}(z')) = C_{d' \leftarrow d}(\mathbf{p}, \mathbf{p}').^{54}$$

These multiple flows of compliers are visible in Figure 12(b) where the shift from z to z' leads to three different flows of compliers: $1 \leftarrow 2$ *compliers*, who switch from $D(z) = 2$ to $D(z') = 1$, $1 \leftarrow 0$ *compliers*, who switch from $D(z) = 0$ to $D(z') = 1$, and $0 \leftarrow 2$ *compliers* who switch from $D(z) = 2$ to $D(z) = 0$. This makes it harder to identify, and even to define, the treatment parameters of interest. For example, for the $1 \leftarrow 0$ compliers, the treatment effect is $Y_1 - Y_0$, but for the $1 \leftarrow 2$ compliers, the relevant margin of treatment is $Y_1 - Y_2$. The analysis is even harder if one focuses on the effect of the shift on Y_0: because the shift from z to z' lead to some individuals shifting into $D(z') = 0$ from $D(z) = 2$ (the $0 \leftarrow 2$ compliers) but also some individuals shifting out of $D(z) = 0$ to $D(z') = 1$ (the $1 \leftarrow 0$ compliers).

An interesting property of the discrete index model is that, as visible in Figure 13, if one increases only one semi-IV, Z_d, then there will only be $J - 1$ flows of compliers from the other alternatives into $D = d$. Thus, I analyse only treatment effects induced by shifts in a single semi-IV at a time, keeping in mind that any general shift in Z can be decomposed into multiple single semi-IV shifts. Next, I define the treatment effects of interest with semi-IVs.

[54] Notice that, in the binary treatment case, one had $Compliers_{1 \leftarrow 0}(p, p') = \{V : p \leq V < p'\}$, i.e., one could characterize the set of compliers in terms of the latent resistance to treatment V. In the discrete treatment case, this is not possible because the mass associated with each $\mathbf{V} = (V_1, V_2)$ point in $\boldsymbol{\mathcal{V}}$ is unknown (because the correlation among the ϵ_d is unrestricted). Therefore, one cannot identify the d-specific utilities (u_d) associated with each alternative d. Thus, one does not know the exact mapping between $\mathbf{P}$, $\mathbf{V}$, and all the u_d. One only knows that there is a one-to-one mapping, which allows us to define the complier areas using the observable shift of the generalized propensity score they correspond to. This complicates the identification of margin-specific treatment effects. Indeed, for instance, without further assumption, one cannot identify the shift in $\mathbf{P}$ that would yield only the $1 \leftarrow 0$ compliers of Figure 12(b) (induced into $D = 1$ by the shift in Z from z to z') to shift back from $D = 0$ into $D = 1$. Fortunately, following Mountjoy (2022), one can proceed by using marginal shifts in d-specific semi-IVs to identify these margin-specific MTEs.

Figure 13: Complier flows with one semi-IV shift ($J = 3$)



(a) $\Delta^+ Z_1$ (b) $\Delta^+ Z_0$ (c) $\Delta^+ Z_2$

F.3.3 Margin-specific LATE

Start with a given set of semi-IVs $Z = z \in \mathcal{Z}$. From here onwards, focus on a shift in only one semi-IV, Z_d, from z_d to z'_d, leading from z to z' with associated propensity scores $\mathbf{P}(z) = \mathbf{p} \neq \mathbf{P}(z') = \mathbf{p}'$ with $p'_d > p_d$ and $p'_k \leq p_k$ for all $k \neq d$. For example, imagine that all else equal, the GDP of the manufacturing sector ($D = d$) increases, making it more attractive and diverting workers from the other sectors into manufacturing. What is the treatment effect of this change for these compliers/diverted workers? It depends on the choice they would have made without the change, i.e., $D(z)$. For complying workers diverted from a sector $k \neq d$, the *margin-specific local average treatment effects* are given by:

$$\text{LATE}_{d \leftarrow k}(z, z') = \mathbb{E}[Y_d - Y_k | \mathbf{V} \in C_{d \leftarrow k}(z, z'), Z_d = z'_d, Z_k = z'_k], \tag{37}$$

with $z'_k = z_k$ since only z_d was shifted. More generally,

$$\begin{aligned} \text{LATE}_{d \leftarrow k}(z, z') &:= \text{LATE}_{d \leftarrow k}(\mathbf{p}, \mathbf{p}', z'_d, z'_k) \\ &= \mathbb{E}[Y_d - Y_k | \mathbf{V} \in C_{d \leftarrow k}(\mathbf{p}, \mathbf{p}'), Z_d = z'_d, Z_k = z'_k]. \end{aligned} \tag{38}$$

One can split these LATE into two LATRs, for the outcomes Y_j with $j \in \{d, k\}$, as

$$\begin{aligned} \text{LATR}^j_{d \leftarrow k}(z, z') &= \mathbb{E}[Y_j | \mathbf{V} \in C_{d \leftarrow k}(z, z'), Z_j = z'_j], \\ &= \mathbb{E}[Y_j | \mathbf{V} \in C_{d \leftarrow k}(\mathbf{p}, \mathbf{p}'), Z_j = z'_j] := \text{LATR}^j_{d \leftarrow k}(\mathbf{p}, \mathbf{p}', z'_j). \end{aligned} \tag{39}$$

(non)-Identification. For a shift in Z_d, the $\text{LATR}^k_{d \leftarrow k}$ are identified for any $k \neq d$, while for Y_d, the $\text{LATR}^d_{d \leftarrow k}$ are not directly identified separately for any k of origin.

First, for any potential outcome Y_j with $j = k \neq d$, the $\text{LATR}^k_{d \leftarrow k}(\mathbf{p}, \mathbf{p}', z_k)$ are directly

identified from the data by

$$\text{LATR}^{k}_{d\leftarrow k}(\mathbf{p},\mathbf{p}',z_k) = -\frac{\mathbb{E}[YD_k|Z=z'] - \mathbb{E}[YD_k|Z=z]}{P_k(z') - P_k(z)}. \tag{40}$$

Indeed, following the reasoning of the binary case and using that $z'_k = z_k$, we have

$$\begin{aligned}
&\mathbb{E}[YD_k|Z=z'] - \mathbb{E}[YD_k|Z=z] \\
&= \mathbb{E}[Y_k D_k(\mathbf{p}')|Z_k = z'_k] - \mathbb{E}[Y_k D_k(\mathbf{p})|Z_k = z_k] \\
&= \mathbb{E}[Y_k (D_k(\mathbf{p}') - D_k(\mathbf{p}))|Z_k = z_k] \\
&= -\mathbb{E}[Y_k|\mathbf{V} \in C_{d\leftarrow k}(\mathbf{p},\mathbf{p}'), Z_k = z_k] \times (P_k(\mathbf{p}') - P_k(\mathbf{p})),
\end{aligned} \tag{41}$$

where $P_k(\mathbf{p})$ is the k element of $\mathbf{p}$ and is equal to $P_k(z)$ if $P(z) = \mathbf{p}$. The minus sign in the last equality comes from the fact that $D_k(\mathbf{p}') - D_k(\mathbf{p})$ can only equals 0 (for the stayers who keep $D = k$) or -1 (for the 'compliers' who flow out of $D = k$). Since we only shifted Z_d to reach $Z = z'$ from $Z = z$, $z'_k = z_k$, and there is no direct effect of the included semi-IV on Y_k. Z_d serves as an IV for Y_k, as in the binary treatment case. More importantly, the reason why we can identify this treatment response if that we can map individuals with $\{D_k(z') = 0; D_k(z) = 1\}$ to the compliers $C_{d\leftarrow k}(z,z')$ or $C_{d\leftarrow k}(\mathbf{p},\mathbf{p}')$. Indeed, under the model assumptions, when one shifts Z_d, there is only an outflow of individuals from $D = k$ to $D = d$. There is no inflow into any $D = k$ with $k \neq d$. So the complying set $d \leftarrow k$ of mass $P_k(z) - P_k(z')$ is identified from the perspective of Y_k, and we can identify $\text{LATR}^{k}_{d\leftarrow k}$ as we would have with an IV. The intuition is that the observable differences in YD_k are driven solely by individuals who moved from $D = k$ to $D = d$ due to the change in semi-IVs.

However, for the potential outcome corresponding to the shifted semi-IV, Z_d, i.e., for Y_j with $j = d$, one cannot identify $\text{LATR}^{d}_{d\leftarrow k}$ for any $k \neq d$. Indeed, even though we observe

$$\mathbb{E}[YD_d|Z=z'] - \mathbb{E}[YD_d|Z=z],$$

this comparison is not enough to identify the margin-specific treatment effects of interest. Because, as in its counterpart with IVs, when shifting Z from z to z' by moving only Z_d, there are $J-1$ inflows of compliers into $D = d$. As a consequence, observing

$$D_d(z) = 0; D_d(z') = 1 \iff \mathbf{V} \in \bigcup_{k\in\mathcal{D};k\neq d} C_{d\leftarrow k}(z,z').$$

Without further assumption, we cannot determine the origin of the compliers selecting $D = d$, and we only observe the total inflow (union of all the $d \leftarrow k$ compliers) of all the compliers selecting $D = d$ due to the change in Z. This is why we can only directly identify a *pooled LATE* across all the margins. In the terminology of Heckman et al. (2006), this is the treatment effect of one choice (d) versus the "*next best alternative*".

F.3.4 Pooled LATE, or LATE versus the next best alternative

Define the *pooled LATE* as the total treatment effect on Y for all the individuals who switch to $D = d$ due to an exogenous shift in Z_d from z_d to z'_d, holding all the other semi-IVs fixed. This pooled LATE is equal to

$$\text{LATE}_d(z, z') = \sum_{k \in \mathcal{D}; k \neq d} \text{LATE}_{d \leftarrow k}(z, z') \times \frac{P_k(z) - P_k(z')}{P_d(z') - P_d(z)}. \tag{42}$$

Plugging (37) and (39) into (42), one can show that this LATE is equal to a sum of $\text{LATR}^j_{d \leftarrow k}$,

$$\begin{aligned} \text{LATE}_d(z, z') = & \overbrace{\sum_{k \in \mathcal{D}; k \neq d} \text{LATR}^d_{d \leftarrow k}(z, z') \frac{P_k(z) - P_k(z')}{P_d(z') - P_d(z)}}^{=: \text{ LATR}_d(z, z')} \\ & - \sum_{k \in \mathcal{D}; k \neq d} \text{LATR}^k_{d \leftarrow k}(z, z') \frac{P_k(z) - P_k(z')}{P_d(z') - P_d(z)}, \end{aligned} \tag{43}$$

where all the $\text{LATR}^k_{d \leftarrow k}$ are identified as described previously, so only the weighted sum of $\text{LATR}^d_{d \leftarrow k}$ over all $k \neq d$ remains to be identified. Since $P_k(z) - P_k(z')$ is simply the mass of the $d \leftarrow k$ compliers, and $P_d(z') - P_d(z)$ the total mass of all the compliers that come to $D = d$, this sum is equal to

$$\begin{aligned} \text{LATR}_d(z, z') &= \mathbb{E}[Y_d | \mathbf{V} \in \bigcup_{k \in \mathcal{D}; k \neq d} C_{d \leftarrow k}(z, z'), Z_d = z'_d] \\ &= \mathbb{E}[Y_d | D_d(z) = 0, D_d(z') = 1, Z_d = z'_d] \\ &= \mathbb{E}[Y_d | D_d(\mathbf{p}) = 0, D_d(\mathbf{p}') = 1, Z_d = z'_d] \\ &=: \text{LATR}_d(\mathbf{p}, \mathbf{p}', z'_d). \end{aligned}$$

With this formulation, one can see that the total LATR_d can be identified using a more standard approach, comparing $YD_d(z')$ and $YD_d(z)$.

Because we have semi-IVs instead of IVs, one additional step remains: the shift from $Z = z$ to z' induces a shift from $Z_d = z_d$ to z'_d, so, as in the binary case, one cannot directly compare $\mathbb{E}[YD_d | Z = z'] - \mathbb{E}[YD_d | Z = z]$ to obtain $\text{LATR}_d(z, z')$, because of the effect of the included semi-IV Z_d on Y_d. Instead, we proceed as in the binary case, i.e., by identifying the LATR by shifting the (generalized) propensity score while holding $Z_d = z_d$ fixed. If there exists a $\tilde{z} = (\tilde{z}_0, ..., \tilde{z}_{J-1})$ with $\tilde{z}_d = z'_d$ and such that $\mathbf{P}(\tilde{z}) = \mathbf{p}$, then we have that[55]

$$\text{LATR}_d(\mathbf{p}, \mathbf{p}', z'_d) = \frac{\mathbb{E}[YD_d | Z = z'] - \mathbb{E}[YD_d | Z = \tilde{z}]}{P_d(z') - P_d(\tilde{z})}. \tag{44}$$

[55]The proof is similar to the one developed in (41), except that individuals with $D_d(z') - D_d(\tilde{z}) = 1$ correspond the union of all $d \leftarrow k$ compliers here.

Indeed, if there exists a $Z = \tilde{z}$ with the same propensity score as z, then a change from $Z = \tilde{z}$ to z' induces the same flow of compliers as a change from z to z', i.e., $C_{d \leftarrow k}(z, z') = C_{d \leftarrow k}(\tilde{z}, z') = C_{d \leftarrow k}(\mathbf{p}, \mathbf{p}')$ for each specific margin $k \neq d$, and thus over all margins as well. So we use this change from $\tilde{z}$ to z', which does not change Z_d, to isolate the selection effect at $Z_d = z'_d$. As in the binary case, a change from $\tilde{z}$ to z' to shift $\mathbf{P}$ without changing Z_d requires a shift in all the other semi-IVs.

Provided all the semi-IVs are sufficiently relevant – which is likely to be the case for small changes in Z_d and in propensity scores – $\tilde{z}$ exists (see the reachability property of Subsection F.3.1), and we can identify $\text{LATR}_d(z, z')$. In this case, the pooled $\text{LATE}_d(z, z')$ is also identified by plugging (44) and (40) into (43). Therefore, in the discrete treatment case, semi-IVs can identify the pooled LATE exactly as IVs.

Similarly, one could identify the MTR and MTE in the discrete case, as the limit case where Z_d varies marginally and so does $P_d(Z)$. The adjustments to identify the pooled MTE are straightforward, so I skip the details here.

F.3.5 Margin-specific MTE

While identifying margin-specific LATE is difficult (with either IVs or semi-IVs), the identification of margin-specific MTE remains possible due to specific properties of the complier sets at the margin. In fact, with IVs, this identification is also achievable while relaxing the discrete choice index model into a partial monotonicity and a comparable complier assumption (Mountjoy, 2022).[56] I show that these IV results also hold with semi-IVs.

Application. I focus on the same question as Mountjoy (2022), while noting that the methodology applies to many other previously mentioned examples. Students make an education choice between no college ($D = 0$), two-year colleges ($D = 1$), or four-year colleges ($D = 2$). We want to identify the effect of these choices on subsequent earnings, at age 30 for example. We have 3 d-specific semi-IVs capturing d-specific costs or benefits. The tuition fees of two and four-year colleges are natural candidates for Z_1 and Z_2, respectively: all else equal, for $d \in \{1, 2\}$, students who do not pick $D = d$ do not pay these tuition fees and do not receive the education associated with these fees, so the fees should have no effect on their outcomes (conditional on the included Z_d). Moreover, the tuition fees should be relevant for the selection into the different types of colleges. Unlike the IV approach, I do not assume that the tuition fees are excluded. For instance, I do not rule out that higher tuition for

[56] It may also be possible to relax the monotonicity assumption in other directions with semi-IVs, as Lee and Salanié (2018) or Heckman and Pinto (2018) do with IVs. However, these are less straightforward than the relaxation of Mountjoy (2022), which naturally applies to semi-IVs.

two-year colleges implies a better quality of education in these two-year colleges and thus a better outcome for individuals who selected $D = 1$. For $D = 0$, Z_0 could be a measure of the local market conditions for the non-college educated individuals (e.g., their unemployment or wage rate). This captures the opportunity cost of going to college and may affect Y_0, but should not affect college attendees ($D = 1$ or 2). Note that the three semi-IVs should be taken at the time of the treatment decision, when the individuals were 17 or 18 years old, while the outcome is taken later (e.g., at age 30). The main difference with Mountjoy (2022) is that I need the third semi-IV here (Z_0). The advantage is that I do not need to find two fully excluded IVs, which is infeasible in many other applications.[57] Because we normalize Z_d to have a positive effect on P_d, we define Z_1 and Z_2 as "minus the tuition fees", so that a marginal *increase* in Z_1 corresponds to a marginal *decrease* in two-year college tuition.

In this Framework, a decrease in two-year college tuition ($\Delta^+ Z_1$) directly affects the outcome of individuals with $D = 1$ and, as in the IV case, induces two distinct flows of compliers for whom we aim to identify the margin-specific MTE: (i) the $1 \leftarrow 0$ compliers, who switch from $D = 0$ to $D = 1$ due to the marginal change in Z_1 (democratization effect), and (ii) the $1 \leftarrow 2$ compliers who switch from $D = 2$ to $D = 1$ (diversion effect). Due to these two flows, the overall effect of D_1 on Y due to the change in Z_1, measured by the pooled MTE, is ambiguous and less informative. Shifting individuals without college into two-year college generally improves their future earnings, whereas diverting students from four-year college may be detrimental. Ideally, we want to identify the margin-specific MTEs on both margins separately, as these are more informative. I show how to do this with semi-IVs.

Margin-specific marginal compliers, MTR and MTE.

Define the *marginal complier* switching from $D = k$ to $D = d$ when Z_d increases marginally at $Z = z$ by $C^{Z_d+}_{d\leftarrow k}(z)$, and the marginal complier switching from $D = k$ to $D = d$ when Z_k decreases marginally at $Z = z$ by $C^{Z_k-}_{d\leftarrow k}(z)$.

Without loss of generality, from here onwards, focus on a marginal increase in Z_1. For each potential outcome Y_d with $d \in \mathcal{D}$, define the $\text{MTR}^d_{1\leftarrow k}$ at $Z = z$ as

$$\text{MTR}^d_{1\leftarrow k}(z) = \mathbb{E}[Y_d | \mathbf{V} \in C^{Z_1+}_{1\leftarrow k}(z), Z_d = z_d] \quad \text{for all } k \neq 1,$$

i.e., the average potential outcome Y_d for marginal compliers who switch from $D = k$ to $D = 1$ due to a marginal increase in Z_1 at $Z = z$.

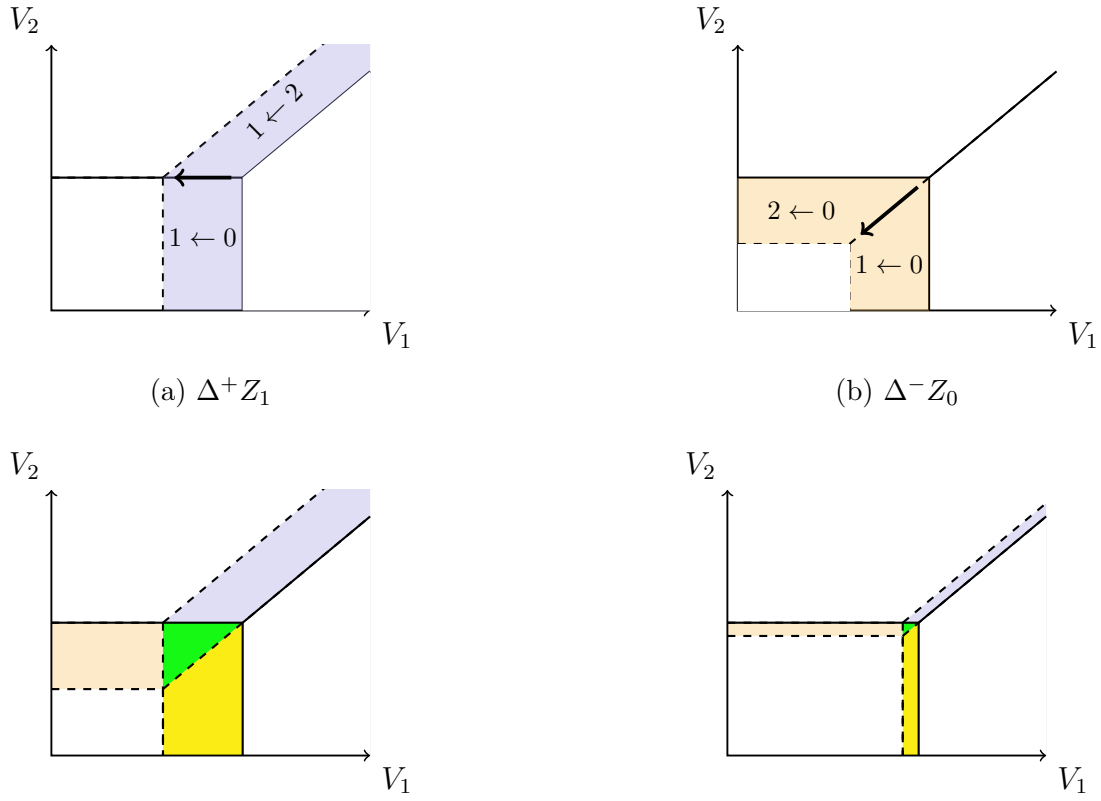
From there, the *margin-specific MTE* can be defined as:

$$\text{MTE}_{1\leftarrow k}(z) = \mathbb{E}[Y_1 - Y_k | \mathbf{V} \in C^{Z_1+}_{1\leftarrow k}(z), Z_k = z_k, Z_1 = z_1] = \text{MTR}^1_{1\leftarrow k}(z) - \text{MTR}^k_{1\leftarrow k}(z).$$

[57] Mountjoy (2022) uses the distance to two and four-year colleges as d-specific IVs.

These are the treatment effects for the marginal compliers who are induced into $D = 1$ from $D = k$ because of the marginal increase in Z_1 when $Z = z$.

Figure 14: Mountjoy (2022)'s identification strategy for $\text{MTE}_{1\leftarrow 0}(z)$



(a) $\Delta^+ Z_1$

(b) $\Delta^- Z_0$

(c) Overlap (yellow)

(d) Overlap for marginal changes

Same marginal compliers property. The key challenge is to disentangle both flows of margin-specific compliers for $D = 1$. Fortunately, under the discrete choice model, if we focus only on marginal changes, one can isolate both margins as illustrated in Figure 14. For any $Z = z$, an increase in Z_1 causes two flows of compliers into $D = 1$: $1 \leftarrow 0$ and $1 \leftarrow 2$. Similarly, a decrease in Z_0 causes two flows of compliers out of $D = 0$: $1 \leftarrow 0$ and $2 \leftarrow 0$. A key insight from Mountjoy (2022) is that, for marginal increases in Z_1 and marginal decreases in Z_0 at $Z = z$, the marginal compliers are the *same*, i.e.,

$$C^{Z_1+}_{1\leftarrow 0}(z) = C^{Z_0-}_{1\leftarrow 0}(z),$$

under the discrete choice index model. This is visible in the panel (d) of Figure 14: the overlap (yellow area) becomes the entire $1 \leftarrow 0$ set at the margin (the green discrepancy

vanishes). Similarly, the marginal compliers are the same for the other margin $1 \leftarrow 2$, i.e.,

$$C^{Z_1+}_{1\leftarrow 2}(z) = C^{Z_2-}_{1\leftarrow 2}(z).$$

Identification. With these "*same marginal compliers*" properties, one can proceed to the identification. As in the binary case, we identify the marginal effect of D on Y_d without shifting Z_d, avoiding confounding direct effects of Z_d on Y_d. This is possible by shifting only the other semi-IVs, Z_{-d}, which are excluded from Y_d and can serve as IVs for Y_d.

For Y_0 and Y_2, using the marginal version of (41), we directly identify the marginal treatment responses at $Z = z$ by marginally increasing the excluded Z_1, i.e.,

$$\text{MTR}^0_{1\leftarrow 0}(z) = \mathbb{E}[Y_0|\mathbf{V} \in C^{Z_1+}_{1\leftarrow 0}(z), Z_0 = z_0] = \frac{\partial \mathbb{E}[Y_0 D_0|Z=z]}{\partial Z_1} \Big/ \frac{\partial P_0(z)}{\partial Z_1},$$
$$\text{and} \quad \text{MTR}^2_{1\leftarrow 2}(z) = \mathbb{E}[Y_2|\mathbf{V} \in C^{Z_1+}_{1\leftarrow 2}(z), Z_2 = z_2] = \frac{\partial \mathbb{E}[Y_2 D_2|Z=z]}{\partial Z_1} \Big/ \frac{\partial P_2(z)}{\partial Z_1}.$$

Then, to identify separately the two margin-specific MTRs for Y_1, one needs to marginally decrease Z_0 and Z_2 and use the "same marginal compliers property" of the discrete index model. More precisely, for $1 \leftarrow k$ compliers, with $k = 0$ or 2, we obtain:

$$\begin{aligned}\text{MTR}^1_{1\leftarrow k}(z) &= \mathbb{E}[Y_1|\mathbf{V} \in C^{Z_1+}_{1\leftarrow k}(z), Z_1 = z_1] \\ &= \mathbb{E}[Y_1|\mathbf{V} \in C^{Z_k-}_{1\leftarrow k}(z), Z_1 = z_1] = \frac{\partial \mathbb{E}[Y_1 D_1|Z=z]}{\partial Z_k^-} \Big/ \frac{\partial P_1(z)}{\partial Z_k^-},\end{aligned}$$

by decreasing Z_k and holding $Z_1 = z_1$ fixed. Notice that $\partial P_1(z)/\partial Z_k^- = -\partial P_k(z)/\partial Z_1^+$ because these are the same set of compliers with the same mass.

Thus, for any $k = 0, 2$, the margin-specific MTE are identified by

$$\text{MTE}_{1\leftarrow k}(z) = \text{MTR}^1_{1\leftarrow k}(z) - \text{MTR}^k_{1\leftarrow k}(z) \quad \text{for any } Z = z \text{ in the interior of } \mathcal{Z}.$$

Incidentally, note that these derivations are very close to the one in Mountjoy (2022), who already uses a second IV to identify the two margins and who also already condition the MTE on $Z = z$. The only slight difference is that I use three different semi-IVs to disentangle the two different margins, while Mountjoy (2022) only uses a second IV to pin down one of the margins, and then uses the difference with the pooled MTE to obtain the remaining margin. Thus, the only minor distinction is that I identify the pooled MTE differently, without shifting Z_1 for MTR^1 in order to avoid the direct effects.

Relaxing the discrete index model. As in Mountjoy (2022), it is straightforward that the previous proof still holds under weaker assumptions than the discrete index model. The necessary ingredients are an (unordered) partial monotonicity assumption combined with *comparable compliers* properties (instead of strictly the 'same' compliers). Both of these are

implied by the discrete index model, which is stronger than necessary. Let us adapt these two assumptions with semi-IVs.

Assumption 6 (Unordered Partial Monotonicity)**.** *For all* $z = (z_0, ..., z_{J-1}) \in \mathcal{Z}$*, define* $z' = (z'_0, ..., z'_{J-1})$ *with* $z'_d > z_d$*, and the other* $z'_k = z_k$ *for all* $k \neq d$*. We have*

$$D_d(z') \geq D_d(z) \quad \text{and} \quad D_k(z') \leq D_k(z),$$

with each inequality holding strictly for at least some individuals (relevance).

Comparable Compliers ($d \leftarrow k$) **property.** For $d, k \in \mathcal{D}$ with $d \neq k$, for all $z = (z_0, ..., z_{J-1}) \in \mathcal{Z}$, define $z' = (z'_0, ..., z'_{J-1})$ with $z'_d > z_d$ and $z'_j = z_j$ for all $j \neq d$, and $\tilde{z} = (\tilde{z}_0, ..., \tilde{z}_{J-1})$ with $\tilde{z}_k < z_k$ and $\tilde{z}_j = z_j$ for all $j \neq k$. We say that the compliers from a marginal increase in Z_d or decrease in Z_k at $Z = z$ are comparable, if for $j \in \{d, k\}$ we have

$$\lim_{\tilde{z}_k \uparrow z_k} \mathbb{E}[Y_j | D(\tilde{z}) = d, D(z) = k, Z_j = z_j] = \lim_{z'_d \downarrow z_d} \mathbb{E}[Y_j | D(z') = d, D(z) = k, Z_j = z_j].$$

Assumption 7 (Comparable Compliers, $J = 3$, Z_1 increase)**.** *The comparable complier properties hold for both* $1 \leftarrow 0$*, and* $1 \leftarrow 2$ *compliers.*

The unordered partial monotonicity (Assumption 6) is sufficient to ensure that shifting only Z_d only induces flows into $D = d$ from all the other alternatives. The comparable compliers property is a weaker form of the "same compliers" property implied by the discrete choice model. The previous identification proof only requires that this property holds for $1 \leftarrow 0$ and $1 \leftarrow 2$ compliers (Assumption 7). Together, assumptions 6 and 7 are sufficient to identify the margin-specific MTE following the previous reasoning, without imposing the stronger discrete index selection model.

Both assumptions are direct adaptations of the ones in Mountjoy (2022) with IVs, and I refer to the original paper for more discussion. The only distinction with semi-IVs is that one needs a comparable compliers assumption around both margins, $1 \leftarrow 2$ and $1 \leftarrow 0$, instead of only one of the two. Fortunately, this is a weak and natural additional assumption to make.

Concluding remark. In summary, for nonparametric identification of margin-specific MTE, one can use J d-specific semi-IVs instead of $J - 1$ d-specific IVs. This is convenient for applied researchers since semi-IVs are easier to find. Moreover, one can also identify the direct effects of the semi-IVs, which may be of interest in their own right. For instance, one can identify the effect of subsidizing two-year college tuitions on the subsequent earnings of two-year college attendees at different margins of compliance.