IsoLATEing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons*

Peter Hull†

December 2015

Abstract

Instrumental variables (IV) estimates of causal effects can be difficult to interpret when the counterfactual to treatment mixes multiple alternatives. I explore identification of multiple counterfactual-specific local average treatment effects from a single quasi-experiment using interactions of an instrument with stratifying controls. I derive the general form of such IV estimands and establish identification under mean-independence of complier treatment effects with respect to the stratification. Under weaker conditional independence assumptions, identification is achieved with a novel non-parametric weighting approach. I use this framework to estimate the returns to GED certification in a sample that includes individuals who would otherwise obtain a traditional high school diploma as well as those who would otherwise drop out. The theoretical results may also offer a strategy to adjust for endogenous attrition in randomized control trials; I illustrate this through a re-analysis of the Oregon Health Insurance Experiment.

*I thank Joshua Angrist, Parag Pathak, Christopher Walters, Amy Finkelstein, Miikka Rokkanen, Sally Hudson, Kathleen Mullen, Christopher Taber, C. Jack Liebersohn, Kirill Borusyak, Mayara Felix, and seminar participants from the University of Calgary and the Bank of Canada for valuable feedback. I gratefully acknowledge financial support from the National Institute on Aging (grant #T32-AG000186)

†MIT Department of Economics. Email: `hull@mit.edu`; website: `http://economics.mit.edu/grad/hull`

1 Introduction

What are the labor market returns to passing a high school equivalency test, such as the U.S. General Educational Development (GED) exam? As with many economic questions, a likely answer is "it depends." In recent years the inherent heterogeneity of causal effects has become a central consideration in applied research. In a seminal contribution, Imbens and Angrist (1994) show that an instrumental variables (IV) regression on a single binary treatment variable may estimate average causal effects for "compliers" – those who are induced into treatment by receipt of the instrument. IV identification of these local average treatment effects (LATEs) is given by a binary instrument that is as good as randomly assigned, monotone in its effect on treatment receipt, and excludable from potential outcome realizations.

When treatment effects vary, it is of natural interest to characterize their heterogeneity. A straightforward characterization stratifies individuals along a dimension that is unaffected by and independent of the instrument. When the Imbens and Angrist (1994) assumptions hold within such strata, stratum-specific LATEs are identified by conditional IV regressions. For example, in a randomized trial in which offers for a particular program are assigned by lottery, stratification on pre-randomization controls can reveal differential effects of the program for compliers with different baseline characteristics.

Often, however, the most important dimensions of heterogeneity are not directly revealed by a baseline stratification. Suppose some individuals are caused to take the GED by a plausibly exogenous decrease in passing standards and subsequently earn different wages in adulthood. Such compliers may be drawn from meaningfully different counterfactual levels of education: for some the alternative to the GED may be to drop out of high school, while others may see an easier GED as a lower-cost substitute to a high school diploma. Under the LATE assumptions, an IV regression with a single GED treatment channel identifies a causally-interpretable weighted average of effects for these two types of individuals, but this parameter may be difficult to interpret *economically*. Namely, if labor markets tend to reward workers for higher levels of educational achievement, the overall LATE may mix together potentially large positive and negative effects. Except in very special situations, baseline measures are unlikely to provide enough information to perfectly separate individuals by their counterfactual educational attainment.

In some settings it may be difficult even to give *causal* interpretation to effects averaged over

different treatment counterfactuals. In an extreme case, the instrument may move compliers from a state in which outcomes, measured by a survey or otherwise voluntarily provided, are completely unobserved to the researcher. This leads to the well-known problem of differential attrition: restricting analyses to individuals with *ex post* valid outcomes is likely to introduce selection bias into an otherwise gold-standard randomized design, while IV estimates in the complete sample do not identify meaningful causal parameters. Here isolating effects for a subset of compliers – those who would contribute outcomes even when untreated – is of first-order concern, yet as in the GED example potential attritors are unlikely to be perfectly identified and removed by baseline characteristics alone.

In this paper I explore ways in which baseline stratifications, while not able to completely separate different complier groups, may nevertheless be useful for disentangling treatment effects by their counterfactual state. The basic strategy is intuitive: if there exists a stratification across which the composition of compliers with different alternatives varies but, on average, causal effects do not, differences in stratum-specific reduced-form effects may be attributed to differences in complier shares in such a way that identifies a LATE for treatment relative to each counterfactual. This intuition motivates an IV regression with multiple endogenous variables identified by interactions of the instrument with stratum indicators. The requirement that average complier treatment effects be mean-independent of the stratification may be too strong in practice, however. In general I show that IV can informs bounds on particular linear combinations of counterfactual-specific effects, and that the approach may be generalized to settings where average cross-stratum heterogeneity is captured by a rich set of controls. In this case, I propose a non-parametric weighting estimator to identify the multiple LATEs.

Interacting an instrument with covariates to identify coefficients on multiple endogenous variables has a long history in economics; the results here extend the usual constant-effects framework to a minimal set of assumptions that allow for treatment effect heterogeneity.[1] In two closely related settings, Behaghel, Crepon, and Gurgand (2013) consider non-parametric identification of multiple causal channels given an independently-assigned instrument for each channel, while Kirkebøen, Leuven, and Mogstad (2016) show that counterfactual-specific LATEs may be recovered by IV regression when a researcher is able to directly observe and stratify on each individual's

[1] Wooldridge (2002) discusses identification of simultaneous equation models by nonlinear transformations of the instrument and predetermined covariates. Abdulkadiroğlu et al. (2016) and Cohodes (2015) are two recent examples of this approach in a treatment effects framework.

most-preferred alternative to treatment. This paper offers an alternative approach for when only a single quasi-experiment is available and the counterfactual treatment status of each individual is unknown. All three techniques fit within the general principal stratification framework of Frangakis and Rubin (2002), which can be thought to extend the three behavioral groups – always-takers, never-takers, and compliers – of the original LATE theorem to allow for multiple causal channels.[2] Finally, in an extension the IV results I propose using covariates and a non-parametric weighting scheme to account for heterogeneity in average causal effects, an approach similar to that of Angrist and Fernandez-Val (2013) and Angrist and Rokkanen (2016) for LATE extrapolation across different quasi-experiments and within regression discontinuity designs, respectively.

I develop the main theoretical results in the context of the motivating GED example and illustrate IV identification in a selection model similar to the one Heckman and Urzúa (2010) use to demonstrate structural identification of counterfactual-specific GED returns. I then apply the theory to two settings. First, I leverage a plausibly exogenous policy change that differentially affected GED passing standards in five U.S. states to replicate the findings of Heckman et al. (2012) that (1) an easier GED exam decreases high school completion rates and (2) non-GED students that are older at the time of the change are more likely to drop out than to finish high school. Leveraging an assumption that an individual's age in the year of reform is not systematically related to her returns-to-schooling profile in adulthood, I use a cohort stratification and instrumented difference-in-differences to jointly estimate average GED wage gains for those who would otherwise drop out from high school and those who would otherwise graduate. Although the extent of identifying variation is relatively modest in this application and the estimation is correspondingly imprecise, the resulting point estimates are remarkably similar to the parameters of Heckman and Urzúa's structural model.

Finally, I turn to the issue of non-random attrition in randomized control trials. Rather than restricting analyses to the subset of individuals who contribute outcomes *ex post*, I propose using a pre-randomization stratification to isolate causal effects for compliers that would always provide survey outcomes. One promising choice of strata uses the common surveying practice of limited intensive follow-up. Since in practice second-round intensive surveying is often random, this stratification is likely uncorrelated with the distribution of complier treatment effects. Moreover to

[2]In another example of this framework, Feller et al. (2014) estimate differential effects of early childhood interventions across alternative care programs using a parametric assumption on the distribution of potential outcomes. Kline and Walters (2015) and Hull (2016) conduct related analysis using semi-parametric Roy selection models.

the extent further follow-up attempts are successful, average response rates will vary by surveying intensity, generating cross-stratum variation in complier shares. I use this logic to estimate the effects of Medicaid enrollment from the Oregon Health Insurance Experiment. Despite evidence of significant differential attrition, the results confirm robustness of the original Finkelstein et al. (2012) estimates for a variety of financial, health, and medical care outcomes.

2 Theoretical Framework

For each individual we observe a Bernoulli instrument Z, an outcome Y, a dummy covariate X, and a variable T which can equal either 1, a, or b. Here $T = 1$ indicates an individual in treatment, while someone with $T = a$ or $T = b$ is said to be in one of two possible untreated states, or "fallbacks."[3] Indicators for being in a fallback state are given by A and B, respectively; treatment is then indicated by $D = 1 - A - B$. As a stylized example, we may imagine Z indicates a quasi-experimental reduction in a student's GED passing standards, Y denotes her adult earnings, and $D = 1$ if the individual becomes GED-certified. Uncertified individuals may either be high school dropouts ($A = 1$) or have a traditional high school diploma ($B = 1$).

As in Rubin (1974), causal effects are defined in terms of potential outcomes. Potential treatment and fallback states when $Z = z \in \{0, 1\}$ are written D_z, A_z, and B_z, while Y_{zt} denotes potential realizations of Y when the instrument takes on the value z and the treatment status is $t \in \{1, a, b\}$. Potential outcomes and assignments are assumed to be independent across individuals, satisfying the usual stable unit treatment value assumption.

We start with the following three conditions on these latent variables:

Assumption 1 *Independence*: $((Y_{z1}, Y_{za}, Y_{zb}, A_z, B_z)_{z=0,1})$ is independent of Z, conditional on X

Assumption 2 *Exclusion*: $Pr(Y_{0t} = Y_{1t}|X) = 1$, for each $t \in \{1, a, b\}$

Assumption 3 *Monotonicity*: $Pr(A_1 \leq A_0|X) = Pr(B_1 \leq B_0|X) = 1$.

In the GED example, Assumption 1 states that the variation in passing standards captured by Z is as good as randomly assigned with respect to potential outcomes, within strata defined by X.

[3]It is straightforward to state the following assumptions and prove identification results in the general case of n untreated states and n distinct elements in the support of X. I work through the specific case where $n = 2$ for ease of notation and exposition.

Conditional independence is sufficient for identification of the reduced-form causal effects of Z on Y, A, and B. Interpretation of the earnings effect by way of schooling T requires an exclusion restriction (Assumption 2), which defines the single-indexed potential outcomes $Y_t = Y_{zt}$ for each t. Finally, we assume the effect of the instrument on treatment status is monotone, in the sense that no individual is induced to either untreated state by Z. Monotonicity is central to LATE identification and is naturally assumed in many contexts; in the stylized example it implies that no student is led to drop out or complete high school when it is easier to obtain a GED, which may be thought of as a revealed preference restriction.[4]

When Assumption 3 holds we may categorize individuals as one of four types by their potential treatment and fallback states:

	$D_1 = 0$	$D_1 = 1$
$D_0 = 0$	**1. Never-takers** ($A_1 = 1$, $A_0 = 1$ $B_1 = 0$, and $B_0 = 0$, or $A_1 = 0$, $A_0 = 0$, $B_1 = 1$, and $B_0 = 1$)	**2. a-compliers** ($A_1 = 0$, $A_0 = 1$, $B_1 = 0$, and $B_0 = 0$) **3. b-compliers** ($A_1 = 0$, $A_0 = 0$, $B_1 = 0$, and $B_0 = 1$)
$D_0 = 1$		**4. Always-takers** ($A_1 = 0$, $A_0 = 0$, $B_1 = 0$, and $B_0 = 0$)

In the GED example, never-takers are those who would either always drop out of high school or always obtain a traditional diploma, while always-takers are students that obtain a GED even when it is difficult to pass. Compliers are individuals who switch to a GED when the test becomes easier, and may either drop out of (a-compliers) or complete high school (b-compliers) when passing standards increase. Note that a-compliers are those with $A_1 < A_0$ while b-compliers have $B_1 < B_0$.

Although stated with expanded notation, it is straightforward to verify that Assumptions 1-3 are equivalent to those typically used to analyze causal effects with a single treatment channel. The payoff to the more elaborate setup is that we can now write the conditional LATE identified

[4]Students who would have otherwise completed high school may be led to *attempt* the GED when passing requirements are low but not actually pass and perhaps drop out instead. If passing forecast errors were systematically related to potential outcomes this may violate Assumption 3, though idiosyncratic monotonicity violations may be accommodated by extensions along the lines of de Chaisemartin (2015) in the single-treatment case.

by such analyses as an average of two fallback-specific LATEs for each of the now-differentiated a-complier and b-complier sub-populations. Specifically, we have the following:

Lemma 1 : Consider the IV regression of Y on D, instrumented by Z and conditional on X. Suppose $Pr(D_1 > D_0|X) \neq 0$. Then under Assumptions 1-3 the endogenous regressor coefficient identifies

$$E[Y_1 - Y_a|A_1 < A_0, X]\omega(X) + E[Y_1 - Y_b|B_1 < B_0, X](1 - \omega(X)), \tag{1}$$

where

$$\omega(X) = \frac{Pr(A_1 < A_0|X)}{Pr(A_1 < A_0|X) + Pr(B_1 < B_0|X)}. \tag{2}$$

The proof of Lemma 1, derived in the appendix along with all other propositions, uses the equivalence of Assumptions 1-3 and the assumptions of Imbens and Angrist (1994) in order to write the conditional IV estimand as a local average treatment effect. With two observable fallbacks to treatment, this LATE may in turn be written as the weighted average of average causal effects of the two complier groups, with weights equal to their population shares among all compliers. A typical IV regression with a single GED treatment channel weights together the LATE for students with a dropout counterfactual and the LATE for students who would have otherwise completed high school. Here we are interested in extracting these two causal effects from the overall average.

Note that while the model has been formulated in terms of multiple fallback states, we could equivalently write Assumption 3 with the inequalities reversed and prove Lemma 1 and subsequent results with A and B corresponding to distinct treatment states.[5] In this context it is worth noting that while Assumptions 1-3 ensure excludability of Z from Y given D (since Z is assumed excludable from Y given T, and Assumption 3 rules out the remaining possibility of Z shifting individuals across fallback states when $D = 0$), they do not rule out violations of exclusion when either A or B is considered in isolation. This is because an individual not observed in state a may still be induced into state b by the instrument if she is a b-complier, and similarly for b. Thus the theory derived here can also be thought of as addressing cases of known exclusion restriction violations, where the instrument affects untreated (from the perspective of one state) individuals

[5] Identification with multiple treatment states is quite similar to the literature on multiple "mediators." See Reardon and Raudenbush (2013) for related theory and Kling, Liebman, and Katz (2007) and Pinto (2015) for applications with the Moving to Opportunity experiment

by shifting them to another observable treatment state.[6]

The multiple-treatment version of Assumption 1-3 is closely related to the multiple "encouragement design" of Behaghel, Crepon, and Gurgand (2013), who consider two treatment states as well as an instrument that takes on three values, $\tilde{Z} \in \{1, a, b\}$. With $Z = \mathbf{1}[\tilde{Z} = 1]$ and the underlying realization of the multi-valued instrument attributed to individual heterogeneity, Assumptions 1-3 are implied by their framework.[7] Under similar assumptions, Kirkebøen, Leuven, and Mogstad (2016) discuss identification of fallback-specific LATEs when the econometrician is able to measure A_0 and B_0 for all individuals directly. The identification results derived here can be thought of as an alternative method for the more general situation in which only one valid instrument for treatment is available and where A_0 and B_0 (e.g. each student's potential schooling level when they not GED-certified) are not observed. Instead the current approach requires only a stratification that predicts some variation in the average unobserved counterfactual across the population. I next turn to the formal identification results.

2.1 IV Identification

Intuition for the main theoretical result can be seen in equations (1) and (2). Independence of the instrument ensures identification of the conditional first stage causal effect of Z on $1 - A$ and on $1 - B$, and the monotonicity assumption implies these are equal to $Pr(A_1 < A_0|X)$ and $Pr(B_1 < B_0|X)$, respectively. Thus the complier shares underlying the weighting scheme in Lemma 1 are identified. If the two counterfactual-specific LATEs are the same across the strata, equations (1) and (2) then constitute a system of two equations (one for each stratum) and two unknowns (the two counterfactual-specific LATEs), and IV identification is achieved as long as the two equations are not perfectly collinear. An IV regression with two endogenous variables instrumented by Z and the interaction of Z with X exactly implements this intuition.

[6]The multiple-treatment analogue of Assumptions 1-3 can also be placed in the ordered treatment model studied by Angrist and Imbens (1995). Namely, with $S = A + 2B$ and potentials defined accordingly, the ordered treatment setting may be modeled by Assumption 1 and 2 and a modified Assumption 3 that only requires $Pr(S_1 \geq S_0|X) = Pr(A_1+2B_1 \geq A_0+2B_0|X) = 1$ rather than the stronger condition that $Pr(A_1 \geq A_0|X) = 1$ and $Pr(B_1 \geq B_0|X) = 1$.

[7]The present approach to the multiple-treatment setting uses an IV regression with two instruments and two endogenous variables, as does Behaghel, Crepon, and Gurgand (2013). However the instruments proposed here will not satisfy their identifying assumptions in general, so that the approach is indeed distinct. In the special case where a-compliers and b-compliers are completely separated by the stratification the Behaghel, Crepon, and Gurgand (2013) assumptions *will* be satisfied, and Proposition 1 may be proved in a manner similar to their result. This case is quite far afield from the main motivation of this paper, however, as it would in fact involve a baseline stratification that perfectly distinguishes the two causal channels of interest, as in Kirkebøen, Leuven, and Mogstad (2016).

Formally, consider the just-identified, two-treatment IV system:

$$Y = \mu^y + \alpha(1 - A) + \beta(1 - B) + \gamma^y X + \epsilon^y \tag{3}$$

$$1 - A = \mu^a + \pi^a Z + \rho^a (Z \times X) + \gamma^a X + \epsilon^a \tag{4}$$

$$1 - B = \mu^b + \pi^b Z + \rho^b (Z \times X) + \gamma^b X + \epsilon^b. \tag{5}$$

where $E[(\epsilon^y, \epsilon^a, \epsilon^b)'|Z, X] = 0$. To economize on notation, define

$$\alpha(x) = E[Y_1 - Y_a | A_1 < A_0, X = x] \tag{6}$$

$$\beta(x) = E[Y_1 - Y_b | B_1 < B_0, X = x] \tag{7}$$

and

$$f_a(x) = Pr(A_1 < A_0 | X = x) \tag{8}$$

$$f_b(x) = Pr(B_1 < B_0 | X = x). \tag{9}$$

Here $\alpha(x)$ and $\beta(x)$ are stratum- and counterfactual-specific LATEs, while $f_a(x)$ and $f_b(x)$ denote the corresponding shares of a- and b-compliers. We then have the following result:

Proposition 1 : Suppose the matrix of complier shares,

$$\Pi = \begin{bmatrix} Pr(A_1 < A_0 | X = 0) & Pr(B_1 < B_0 | X = 0) \\ Pr(A_1 < A_0 | X = 1) & Pr(B_1 < B_0 | X = 1) \end{bmatrix}, \tag{10}$$

is nonsingular. Then under Assumptions 1-3 the endogenous regressor coefficients in equation (3) identify:

$$\alpha = \omega\alpha(0) + (1 - \omega)\alpha(1) + \delta_a(\beta(0) - \beta(1)) \tag{11}$$

$$\beta = (1 - \omega)\beta(0) + \omega\beta(1) + \delta_b(\alpha(0) - \alpha(1)), \tag{12}$$

where

$$\omega = \left(1 - \frac{f_a(1)}{f_b(1)} / \frac{f_a(0)}{f_b(0)}\right)^{-1} \tag{13}$$

$$\delta_a = \left(\frac{f_a(0)}{f_b(0)} - \frac{f_a(1)}{f_b(1)}\right)^{-1} \tag{14}$$

$$\delta_b = \left(\frac{f_b(0)}{f_a(0)} - \frac{f_b(1)}{f_a(1)}\right)^{-1}. \tag{15}$$

Proposition 1 extends the Imbens and Angrist (1994) interpretation of IV with heterogeneous effects to regressions on multiple endogenous variables. This is achieved by interacting a single instrument with a stratification that induces variation in the composition of a- and b-compliers (so that the first-stage matrix Π is invertible and the IV rank requirement is satisfied). Although similarly derived, equations (11)-(15) are, however, not as easily interpreted as in the single-treatment case. For example, the IV coefficient on $1 - A$ is a weighted average of average causal effects from a-compliers in the two strata plus a "bias" term reflecting the gap in average causal effects of b-compliers across strata. Since the coefficient on this term is identified by components of the first stage matrix Π, with a bounded outcome one could identify bounds on the average $\omega\alpha(0)+(1-\omega)\alpha(1)$, with more narrow intervals given by stratifications where the ratio $f_a(X)/f_b(X)$ is more heterogeneous. Note, however, that since these ratios are always positive given Assumption 3, the parameter ω will be contained in $(-\infty, 0) \cup (1, \infty)$ so that the weighting scheme is never convex.

As can be seen in equations (11)-(12), a special case in which one may wish to estimate a regression of the form of equations (3)-(5) is when treatment effects are constant. More generally, α and β are easily interpreted causal parameters when the chosen stratification is mean-independent of the average treatment effect for the two groups of compliers. That is, suppose in addition to Assumptions 1-3 we have:

Assumption 4 *LATE homogeneity*: $E[Y_1 - Y_a|A_1 < A_0, X]$ and $E[Y_1 - Y_b|B_1 < B_0, X]$ do not depend on X

The form of equations (11)-(12) then makes the following result immediate:

Corollary to Proposition 1 : Suppose Π is of full rank. Then under Assumptions 1-4 the endogenous regressor coefficients in equation (3) identify $\alpha = E[Y_1 - Y_a|A_1 < A_0]$ and $\beta = E[Y_1 - Y_b|B_1 < B_0]$.

With LATE homogeneity, therefore, the multiple endogenous variable IV regression correctly deconvolutes the weighted average of fallback-specific LATEs given by Lemma 1.[8] In the stylized

[8]It is straightforward to extend Proposition 1 to consider multi-valued X and the over-identified IV regression instrumented by multiple stratum interactions. When Assumption 4 holds across all values in the support of X the two LATEs will be identified by any such regression, just as in the constant effects case. A test of overidentifying restrictions would thus be valid for jointly testing Assumptions 1-4.

example, α and β identify the average returns to GED certification for individuals induced to the GED from a dropout and high school completion counterfactual, respectively, provided these do not vary systematically with the covariate X.[9]

It is instructive to consider what kinds of data-generating processes may accommodate Assumptions 1-4. Consider a model of an individual deciding between the treatment and alternative states to maximize her state-specific latent utility ν_{ti}:

$$A_i = \mathbf{1}[\nu_{ai} \geq \nu_{bi}, \nu_{ai} \geq \nu_{1i}] \tag{16}$$

$$B_i = \mathbf{1}[\nu_{bi} \geq \nu_{ai}, \nu_{bi} \geq \nu_{1i}] \tag{17}$$

$$D_i = \mathbf{1}[\nu_{1i} \geq \nu_{ai}, \nu_{1i} \geq \nu_{bi}]. \tag{18}$$

To satisfy Assumption 3 it is sufficient to have, for unidimensional η_i,

$$\nu_{1i} = h(X_i, Z_i, \eta_i), \tag{19}$$

such that $h(x, 1, \eta_i) \geq h(x, 0, \eta_i)$ almost-surely for $x = 0, 1$. Exclusion and LATE homogeneity then hold if potential outcomes may be written

$$Y_{ti} = \mu_t + \gamma X_i + \epsilon_{ti}, \tag{20}$$

such that

$$E[\epsilon_{1i} - \epsilon_{ai} | A_{1i} < A_{0i}, X_i] = E[\epsilon_{ai} - \epsilon_{ai} | A_{1i} < A_{0i}] \tag{21}$$

$$E[\epsilon_{1i} - \epsilon_{bi} | B_{1i} < B_{0i}, X_i] = E[\epsilon_{ai} - \epsilon_{bi} | B_{1i} < B_{0i}], \tag{22}$$

while Assumption 1 holds if the vector of structural disturbances $(\eta_i, \nu_{ai}, \nu_{b1}, \epsilon_{1i}, \epsilon_{ai}, \epsilon_{bi})'$ is independent of Z_i, conditional on X_i. Note that in writing equations (20)-(22) we are neither assuming that the stratum indicator X_i is excludable from the structural outcome equation (20) nor that it is independent of its error ϵ_{ti} (in which case X_i may itself be thought of as an instrument); rather, LATE homogeneity asserts that X_i enters the outcome equation in an additively-separable way, and that *differences* in the residual determinants of Y_{ti} are mean-independent of X_i in the compliant

[9] Proposition 1 also provides a way to indirectly validate Assumption 4 given an exogenous control G thought to be correlated with individual treatment effects. For example, estimating equations (3)-(5) by setting $Y = G \times A$ yields a second-stage coefficient on $1 - B$ of $\delta_b(E[G|A_1 < A_0, X = 1] - E[G|A_1 < A_0, X = 0])$ under Assumptions 1-3, since then $\beta(0) = \beta(1) = 0$. One could therefore test whether the control G systematically varies for a-compliers across the X-stratification (and likewise for b-compliers). It is straightforward to generalize this test along the lines of the following section.

sub-populations. Section 3 and appendix section A.4 give a parametric example of such a model.

2.2 Relaxing Independence and Homogeneity

That cross-stratum comparisons are informative for a common pair of treatment effects is essential for their identification by IV. For any given application, which stratification is most likely to maintain an independent instrument and homogeneous LATEs while still producing first-stage variation depends on the specific context. Helpfully, as with intent-to-treat effect and conventional LATE identification (Hirano, Imbens, and Ridder, 2003; Abadie, 2003), this approach may be extended to settings where Assumptions 1-4 only hold conditional on a rich set of predetermined covariates. The strategy is again intuitive: one could imagine running conditional versions of the IV regression (3)-(5) at each point in the support of a discretely-valued control W. When conditional cross-stratum comparisons identify conditional fallback-specific LATEs, averaging the resulting coefficients over the marginal complier distribution of W will recover population LATEs. Such a procedure is conceptually possible yet likely infeasible when W is continuous or takes on many discrete values.[10] I next outline an alternative, more flexible implementation of this basic idea for a generic vector of controls W.

We start by considering the conditional analogues of the key identifying assumptions:

Assumption 1′ $((Y_{z1}, Y_{za}, Y_{zb}, A_z, B_z)_{z=0,1})$ is independent of Z, conditional on W and X

Assumption 2′ $Pr(Y_{0t} = Y_{1t}|W, X) = 1$, for each $t \in \{1, a, b\}$

Assumption 3′ $Pr(A_1 \leq A_0|W, X) = Pr(B_1 \leq B_0|W, X) = 1$

Assumption 4′ $E[Y_1 - Y_a|A_1 < A_0, W, X]$ and $E[Y_1 - Y_b|B_1 < B_0, W, X]$ do not depend on X

Here Assumption 1′ only requires the instrument Z to be as good as randomly assigned once potential confounders in W and X are held fixed, while Assumption 4′ allows for arbitrary cross-stratum heterogeneity in average complier treatment effects that is captured non-parametrically by W. Assumptions 2′ and 3′ further relax the exclusion and monotonicity restrictions to hold only conditional on W and X. We then have the following result:

[10] As Hirano, Imbens, and Ridder (2003) note, a related issue is whether standard asymptotic theory adequately approximates the sampling distributions of such manually-reweighted estimators. See Robins and Ritov (1997) and Angrist and Hahn (2004) for a discussion of this problem.

Proposition 2 : Suppose $Pr(Z = 1|W, X)$ and $Pr(X = 1|W)$ are bounded away from zero and one and that the matrix of conditional complier shares

$$\Pi(W) = \begin{bmatrix} Pr(A_1 < A_0|W, X = 0) & Pr(B_1 < B_0|W, X = 0) \\ Pr(A_1 < A_0|W, X = 1) & Pr(B_1 < B_0|W, X = 1) \end{bmatrix} \tag{23}$$

is nonsingular with probability one. Define

$$\lambda = \frac{E[Z|W, X] - Z}{E[Z|W, X](1 - E[Z|W, X])}, \tag{24}$$

and

$$\mu_a = \frac{E[\lambda AX|W] - E[\lambda A|W]X}{E[X|W](1 - E[X|W])} \tag{25}$$

$$\mu_b = \frac{E[\lambda BX|W] - E[\lambda B|W]X}{E[X|W](1 - E[X|W])}. \tag{26}$$

Then, under Assumptions 1′-4′,

$$E[Y_1 - Y_a|A_1 < A_0] = E\left[\frac{E[\lambda A|W]}{E[\lambda A]} \frac{\lambda\mu_b}{E[\lambda\mu_b A|W]} Y\right] \tag{27}$$

and

$$E[Y_1 - Y_b|B_1 < B_0] = E\left[\frac{E[\lambda B|W]}{E[\lambda B]} \frac{\lambda\mu_a}{E[\lambda\mu_a B|W]} Y\right]. \tag{28}$$

Furthermore these weighting schemes are non-parametrically identified by the conditional expectations $E[X|W]$, $E[Z|W, X]$, $E[A|W, X, Z]$, and $E[B|W, X, Z]$.

The proof of Proposition 2, given in the appendix, shows that conditional-on-W versions of, for example, the coefficient on $1 - A$ in equation (3) can be written as the ratio of $E[\lambda\mu_b Y|W]$ to $E[\lambda\mu_b A|W]$. Averaging this ratio over the marginal distribution of W for a-compliers (using $E[\lambda A|W]/E[\lambda A]$ weights) thus identifies $E[Y_1 - Y_a|A_1 < A_0]$. For these results the conditional IV estimand must be well-defined along the support of W, so that both $Z|W, X$ and $X|W$ must be almost-surely stochastic and the conditional first-stage matrix $\Pi(W)$ must be always-surely invertible. Thus while W should be general enough to make Z ignorable and stratum-specific LATEs homogeneous, it must still allow for the kind of cross-stratum variation in complier shares underlying the basic IV approach.

As in the unconditional case, we can write a model consistent with Assumptions 1′-4′ by adding covariates to the structural equations for an individual's treatment utility and her potential out-

comes:

$$\nu_{1i} = h(W_i, X_i, Z_i, \eta_i) \tag{29}$$

$$Y_{ti} = f(W_i, X_i) + g_t(W_i) + \epsilon_{ti}, \tag{30}$$

where again we assume differences in ϵ_{ti} are mean-independent of X_i, the vector of structural errors is independent of Z given X and W, and the function $h(w, x, z, \eta_i)$ is almost-surely monotone in z given W and X. In such a model, the conditional LATEs that are weighted together by Proposition 2 may be written

$$E[Y_1 - Y_a | A_1 < A_0, W] = g_1(W) - g_a(W) + E[\epsilon_1 - \epsilon_a | A_1 < A_0, W] \tag{31}$$

$$E[Y_1 - Y_b | B_1 < B_0, W] = g_1(W) - g_b(W) + E[\epsilon_1 - \epsilon_b | B_1 < B_0, W]. \tag{32}$$

Proposition 2 suggests a non-parametric estimation procedure for recovering the unconditional LATEs when Assumptions 1′-4′ hold. Namely, a researcher may in a first step flexibly approximate four conditional expectation functions: $E[X|W]$, $E[Z|W, X]$, $E[A|W, X, Z]$, and $E[B|W, X, Z]$. The appendix shows how these can then be used to form sample analogues of λ, μ_a, μ_b, $E[\lambda A|W]$, $E[\lambda B|W]$, $E[\lambda\mu_b A|W]$, and $E[\lambda\mu_a B|W]$, and thus of the weighting schemes in equations (27) and (28). Unlike with the IV procedure in Proposition 1, inference for this multi-step estimator will in general be non-standard. Under appropriate regularity conditions, finite-sample approximations to the asymptotic distribution of the estimator may be based on either a bootstrap procedure or on analytic expressions derived by the approaches of Andrews (1991) and Newey (1994a, 1994b).

3 Applications

3.1 The Returns to GED Certification

In 1997 the GED Testing Service required all U.S. states to meet new passing score requirements. Prior to this reform five states – Louisiana, Mississippi, Nebraska, New Mexico, and Texas – awarded GEDs to students that obtained either a minimum score of 40 (out of a possible 80) on each of five standardized sub-tests *or* an average score of 45 across all sub-tests, while starting January 1st, 1997 both criteria were required nationwide. In a difference-in-differences design, Heckman et al. (2012) show that this increase in test difficulty, plausibly exogenous from the perspective

of current students, significantly increased the share of high school graduates in affected states. The authors further show that the effect was concentrated among students who were older at the time of the policy change and were thus likely less constrained in their ability to drop out of high school when facing a harder GED exam. To the extent an individual's age at the time of reform was not directly priced in the relative labor market returns she faced in subsequent decades, a stratification that filters the differential reduced-form effect of the policy across birth cohorts through differential rates of high school completion may be used to separate the overall causal effect into counterfactual-specific effects along the lines of Proposition 1.

I first illustrate this approach with a stylized model of degree choice and subsequent earnings. Heckman and Urzúa (2010) use such a model to demonstrate identification of multiple GED effects under large support or parametric conditions; I extend their simulated data-generating process to accommodate a stratification scheme consistent with Assumptions 1-4. Here Y denotes an individual's log hourly earnings in adulthood, D indicates GED certification, while A and B indicate the two GED fallbacks of dropping out and completing high school, respectively. The stratification X indicates an individual's age (either 16 or 17) when a quasi-experimental reduction in GED passing standards, Z, is announced. An appendix section contains a full description of how these variables are generated from draws of latent correlated factors in the Heckman and Urzúa (2010) parameterization.

Population first-stage and reduced-form coefficients from an IV regression of Y on D are reported in Panel A of Table 1. An exogenous decrease in GED passing standards increases the share of GED-certified students by 5.5 percentage points and decreases hourly earnings by an average of 1.2 percent, figures quite consistent with Heckman and Urzúa's original model. By monotonicity the former represents the total share of a- and b-compliers in the population; from Lemma 1 the ratio of reduced-form to first-stage effects, -0.209, is the overall average complier GED effect. Both effects may be decomposed into strata-specific first-stage and reduced-form moments, displayed in Panel B. Compliers in the dropout-constrained ($X = 0$) subsample are more likely to obtain a high school diploma when untreated, while older compliers (with $X = 1$) are more likely to drop out in response to stricter GED passing standards. The model parameterizes adult wages such that students tend to see gains when shifted to the GED from a dropout counterfactual and losses when the GED replaces a high school diploma. Consequently, the reduced form effect of an easier GED exam is

higher when $X = 1$ than when $X = 0$. Population coefficients in the two-treatment IV regression, $\alpha = 0.277$ and $\beta = -0.396$, are obtained by inverting the first stage matrix in columns 2 and 3 and multiplying by the reduced form vector in column 4. Since the model satisfies LATE homogeneity (an individual's cohort X is allowed to affect the level of her wages but not the returns-to-schooling frontier), by Proposition 1 these represent counterfactual-specific local average treatment effects of GED certification on adult earnings.

I next compare the performance of IV estimators in the simulated model with real-world estimates of the returns to GED certification using quasi-experimental variation in GED passing standards from the 1997 reform. I construct a sample of 22,923 individuals born in the U.S. in 1978-1979 and in 1981-1982 who report positive earnings and hours worked in the 2013 American Community Survey and that completed at least two years of high school.[11] These individuals were of age 16 and 17 in either the year 1995 (one year prior to the mandated GED score change) or 1998 (one year after the change) and were likely to face differential GED costs while in high school. As in Heckman et al. (2012) I exclude the actual year of the policy change, which occurred in the middle of the school year.[12] Trends in the rate of GED attainment and in log hourly earnings across birth cohorts and by birth states are plotted in Figure 1A. The proportion of GED-certified individuals born in the five affected states declines sharply for the later birth cohorts, from 19.5 to 14.5 percentage points, while rates from other states show only a modest decrease. Importantly, similar comparisons between earlier birth cohorts not affected by the policy show almost no difference in certification trends, supporting the claim of Heckman et al. (2012) that such difference-in-differences comparisons may be causal. As in the calibrated model, an increase in passing standards led to a decline in GED completion by around 5 percentage points, with only a negligible overall increase in subsequent labor market earnings; these estimates are plotted in Figure 1B.

Under assumptions analogous to those of Imbens and Angrist (1994), the ratio of difference-

[11] Self-employed individuals and college-educated individuals are also excluded for ease of interpretation. Hourly wages are constructed by dividing annual wage and salary income by the product of the usual number of weeks worked within a year and the number of hours worked per week. The latter is imputed from categories reported in the 2013 ACS using the average number of hours reported within the same categories in the 2007 ACS, the last year in which underlying hours were reported. All reported results are robust to these sample construction choices.

[12] In their application, Heckman et al. (2012) drop states that were otherwise affected by the policy change but that had already required candidates to meet both a minimum and mean score requirement, while showing robustness to the choice of control group. To increase power I use all states other than the five affected by the "and/or" scoring change as controls, though the results are similar without the already-required states.

in-differences effects of the policy change on earnings to effects on GED completion identifies the average return to the GED for policy compliers (Hudson, Hull, and Liebersohn, 2015). Assumptions 1-3 may similarly be extended to accommodate this instrumented difference-in-differences framework, in which case the LATE explicitly corresponds to a weighted average of compliers with different non-GED schooling counterfactuals. Difference-in-differences estimates of the effect of the policy on high school dropout and completion rates are 2.3 and 2.7 percentage points, respectively. By monotonicity this suggests that among the 5 percent complier population, 46% would have dropped out under the stricter GED testing regime, while 54% would have completed high school instead. Consistent with Heckman et al. (2012) the dropout counterfactual appears concentrated in the older ($X = 1$) stratum, with 17 year-olds seeing a 4.3 percentage point increase in the probability of dropping out, compared with only 0.1 percentage points among 16 year-olds.

IV estimates of α and β using the 1997 reform are reported in column 1 of Table 2.[13] The overall effect of GED certification on log hourly wages for all compliers is estimated at -0.12, but as in the calibrated model an IV regression with two endogenous variables suggests a compelling underlying story. Compliers who are induced to the GED from a dropout counterfactual appear to see an average increase in hourly wages of 15 percent, while those who are drawn from a high school diploma are estimated to take a massive average 35 percent cut in their hourly earnings. While this application and its estimates are intended as illustrative (indeed, inference based on birth-state clusters with a treatment group of only five states yields standard errors that fail to reject a large range of possible estimates), it is quite striking how closely they resemble the corresponding moments of the model parameterized to match the Heckman and Urzúa (2010) priors, reproduced in column 2. Monte carlo replications of similarly-powered IV regressions reported in column 3 also closely track the real-world LATE estimates.

3.2 Differential Attrition in an RCT

Distinguishing between multiple treatment alternatives can be of first-order importance in a randomized control trial with imperfect follow-up. Suppose program offers Z are randomly assigned to an initial population who may then choose whether or not to comply with the treatment *and*

[13] Due to the small and likely weak sources of identifying variation in this example, I report bias-adjusted (Fuller) 2SLS estimates of these parameters, though unadjusted estimates are essentially the same. All regressions control for state of birth and residency.

whether or not to report subsequent outcomes, Y. Individuals can then be said to select between three possible states: being treated and reporting outcomes (D), not being treated and reporting outcomes (A), and not reporting outcomes (B). Since outcomes are only measured in states D and A (suppose the researcher arbitrarily sets $Y = 0$ for anyone with $B = 1$), the estimable local average treatment effect in the entire sample,

$$E[Y_1 - Y_0|D_1 > D_0] = E[Y_1 - Y_a|A_1 < A_0]\omega + E[Y_1|B_1 < B_0](1 - \omega) \tag{33}$$

$$\text{for} \quad \omega = \frac{Pr(A_1 < A_0)}{Pr(A_1 < A_0) + Pr(B_1 < B_0)}, \tag{34}$$

is not a weighted average of causal treatment effects on the latent, potentially unreported outcome whenever there are any compliers with B as a fallback (that is, when $Pr(B_1 < B_0) \neq 0$). Facing such endogenous attrition, researchers often choose to conduct their analyses on a restricted sub-sample of individuals that report outcomes in the hope of identifying causal effects. Such a procedure, however, is also unlikely to be easily interpreted when $Pr(B_1 < B_0) \neq 0$, as conditioning on *ex post* outcomes ($B_Z = 0$) will then introduce imbalance in the distribution of the instrument Z.[14]

As the form of equations (33) and (34) suggests, the differential attrition problem can be mapped to the multiple-counterfactual setting of Assumptions 1-3. For a given baseline stratification X, independence of Z from potential outcomes (Assumption 1) is ensured by virtue of the randomized design, and in many settings a program offer is likely to have no direct effect on latent outcomes and to not deter program participation. Assumptions 2 and 3 would then be satisfied provided that (1) Z has no direct effect on attrition behavior given treatment status and (2) the effect of assignment on attrition through treatment is monotone. These primitive assumptions that place the differential attrition problem within the general setting considered here are the same as those used to estimate non-parametric bounds on causal parameters by the methods of Lee (2009) and Behaghel et al. (2009).[15]

To solve the differential attrition problem with Proposition 1, we require an appropriate pre-randomization stratification that induces variation in response behavior while maintaining LATE homogeneity. One plausible candidate exploits the practice of randomized intensive follow-up, a

[14]Common approaches to the differential attrition problem include parametric sample selection modeling (Gronau, 1974; Heckman, 1976) and partial non-parametric identification of causal effects (Lee, 2009; Behaghel et al., 2009; Engberg et al., 2014). Methods involving Bayesian inference (Little and Rubin, 1987) and covariate re-weighting (Frölich and Huber, 2014) have also been proposed under different assumptions than those considered here.

[15]Note that monotonicity of response behavior with respect to the instrument is central to the latent index framework most commonly used to study and assess general selection bias (Angrist, 1997).

common surveying technique that is often recommended when attrition rates are large (e.g. Duflo, Glennerster, and Kremer, 2008). Suppose, upon initially measuring outcomes, a researcher selects among the attritors a random fraction p for additional follow-up attempts. Denote this set and another random fraction p of initial responders by $X = 1$ and let $X = 0$ for all other individuals.[16] Since X is highly correlated with an individual's probability of facing more intensive follow-up, the $X = 1$ strata is likely to contain a relatively larger proportion of untreated compliers with an observed outcome. Moreover, since X is randomly assigned in the population, LATEs for both types of compliers will be the same across strata to the extent the additional follow-up attempt draws second-round responses from individuals representative of the pool of initial non-responders.[17]

Assumption 4 is, however, not guaranteed by randomized intensive follow-up *per se*, and researchers hoping to use Propositions 1 or 2 to resolve differential attrition concerns should carefully design the intensive surveying scheme generating such stratifications. As a simplistic but instructive example, suppose a researcher randomly assigns offers for a job-training program and initially conducts phone interviews on employment outcomes throughout the day. As treated individuals may be more likely to be employed, the offer may have an effect (through treatment) on the probability an individual will be home to answer the survey: these people will have $B_1 < B_0$. However, suppose the exact timing of follow-up interviews is as good as random with respect to working hours (perhaps due to alphabetical or other quasi-random queuing of survey attempts), and that the random second round of interviews occurs in similar fashion on a subsequent day. In this case individuals in the intensive stratum of X will face a higher probability of being home when surveyed (on either day one or day two), but those successfully interviewed on the second day will not vary systematically from those interviewed in the initial round. Assumption 4 would then hold, and Proposition 1 may be used to solve the differential attrition problem.

I follow this approach to estimate the effects of Medicaid on survey outcomes in the first year following a lottery of roughly 90,000 low-income adults in Oregon. Finkelstein et al. (2012) discuss the setting for the Oregon Health Insurance Experiment, which selected roughly 35,000 individuals over eight lottery drawings from March through September 2008. Selected individuals became

[16] It will, by Rao-Blackwell logic, in fact be more efficient to let $X = p$ for all initial responders, rather than employing a randomized estimator. I follow this approach in the following empirical application.

[17] In an unpublished manuscript, DiNardo, McCrary, and Sanbonmatsu (2006) discuss parametric and semi-parametric methods of using randomized intensive follow-up to overcome differential attrition in estimating intent-to-treat effects. The current approach is distinct from their approach, both in the interest in local average treatment effects and the specific way the intensive follow-up scheme is used for identification.

eligible for enrollment in OHP Standard, a comprehensive Medicaid program, and roughly 30% of lottery winners successfully enrolled. In addition to administrative hospital discharge data, Finkelstein et al. (2012) collected outcomes by a mail survey, distributed one year later in the summer of 2009, and found evidence that Medicaid increased health care utilization, decreased out-of-pocket expenditure and debt, and improved overall health among survey responders. The relatively low rate of response (at 50%) and moderate imbalance (at around 2 percentage points) in the probability of response by eligibility status, however, suggest caution in interpreting these restricted IV estimates.[18]

I use the Finkelstein et al. (2012) public-use database to replicate the authors' main survey analysis sample. For simplicity I restrict attention to the largest experimental stratum, consisting of 9,770 members of single-person households in the seventh survey wave. Attrition appears to be a more serious issue in this sub-sample, with an overall response rate of only 42% and with eligible individuals roughly 4 percentage points less likely to respond to any survey question.[19] As in the main sample, 30% of initial non-respondents were selected for additional follow-up attempts by mail and phone. The average yield on such intensive surveying was around 22%, suggesting a strong contrast across the stratification scheme described above. I let the stratum indicator $X = 1$ for those designated for intensive follow-up and for a proportionate random sample of initial respondents. The endogenous variables A, B, and D are constructed from survey response and treatment indicators as outlined above.

IV estimates of the effect of Medicaid enrollment on a variety of health, financial, and medical care outcomes are reported in Table 3. As in Finkelstein et al. (2012), columns 1 reports "restricted" IV estimates from specifications with a single treatment variable D, estimated over the subsample of individuals with successfully recorded outcomes.[20] Column 2 instead reports estimates of the coefficient on $1 - A$ in IV regressions of the form of equations (3)-(5) from the full experimental

[18] Finkelstein et al. (2012) address attrition concerns by showing balance in eligibility status across baseline covariates in the survey respondent sub-sample. The authors also construct Lee (2009) bounds for intent-to-treat effects, finding generally robust results for health care use and financial strain outcomes while not able to reject the null of no effect on self-reported health.

[19] Although equations (33)-(34) describe a model in which Z makes attrition less likely, Proposition 1 may also be used to recover causal effects when the instrument increases (through treatment) the probability of non-response. In either case using $1 - A = 1 - (1 - C)R$ and $1 - B = R$ as the two endogenous variables in the two-instrument IV regression (where C indicates treatment receipt and R denotes survey response) will identify the average causal effect of treatment among compliers who always respond by the coefficient on $1 - A$.

[20] I follow Finkelstein et al. (2012) in weighting all restricted IV estimates by the inverse probability of being included in the intensive follow-up group. In practice this has little effect on the restricted IV point estimates.

sample. By Proposition 1, these represent local average treatment effects for compliers who would always provide survey outcomes. Interestingly, the two-treatment IV specification yields point estimates quite close to those obtained by the restricted single-treatment model across virtually every outcome. Although the former is generally less precisely-estimated, the two are highly correlated so that estimated differences (reported in column 3 of Table 3) are tightly distributed around zero. This suggests that, despite apparent endogenous attrition, the estimates reported in Finkelstein et al. (2012) serve as reliable measures of true causal effects of Medicaid enrollment.

4 Conclusions

Although originally formulated within the context of additive, constant-effects models, the method of instrumental variables is occasionally robust to deviations from such parametric frameworks. Indeed, IV estimation of treatment effects has often clarified the minimal assumptions needed for causal interpretation in a fully heterogeneous world. This paper adds to this tradition by extending the theoretical framework of Imbens and Angrist (1994) to settings where more than one causal channel is needed to answer an economic or causal question but only one quasi-experiment is available. The ease by which Proposition 1 may be applied, using an estimator with statistical properties familiar to most applied researchers, is readily apparent in the above empirical applications. More involved, though still tractable estimation may be used to relax the key identifying assumptions given sufficiently rich controls. As the discussion in Section 3.2 illustrates, in some cases one may be able to increase the plausibility of the key LATE homogeneity assumption by a carefully-constructed surveying design. This suggests new tools for overcoming the fundamental issue of differential attrition in randomized program evaluation.

Figure 1A: Trends in GED attainment and log hourly earnings by the 1997 "and/or" GED scoring change

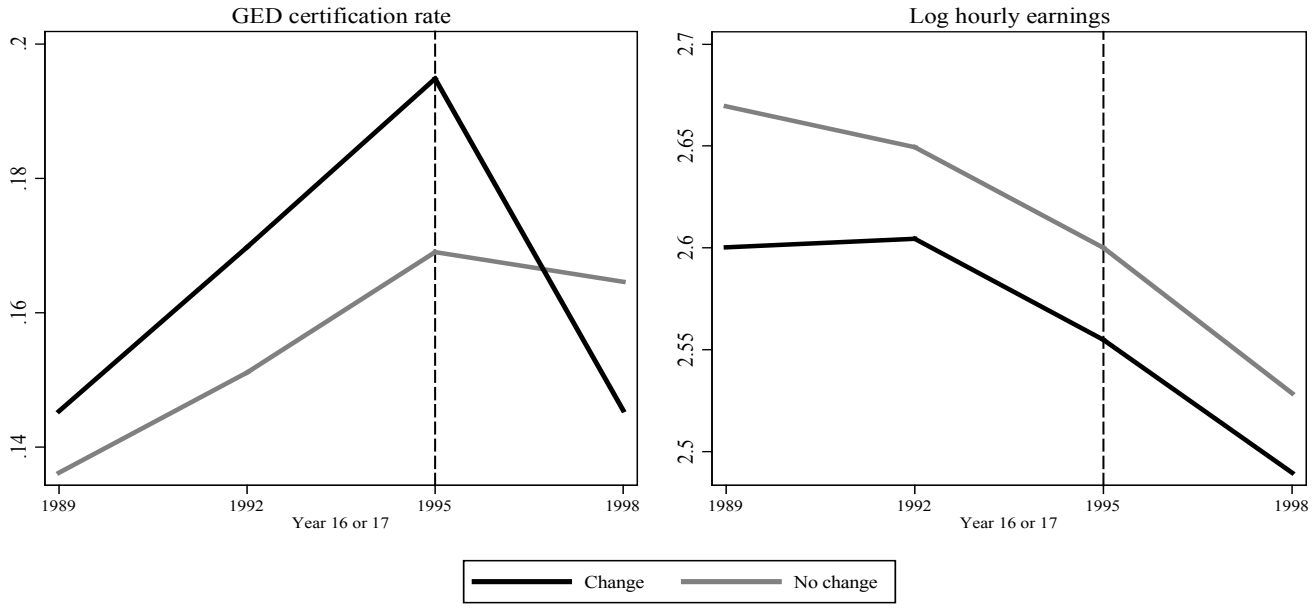
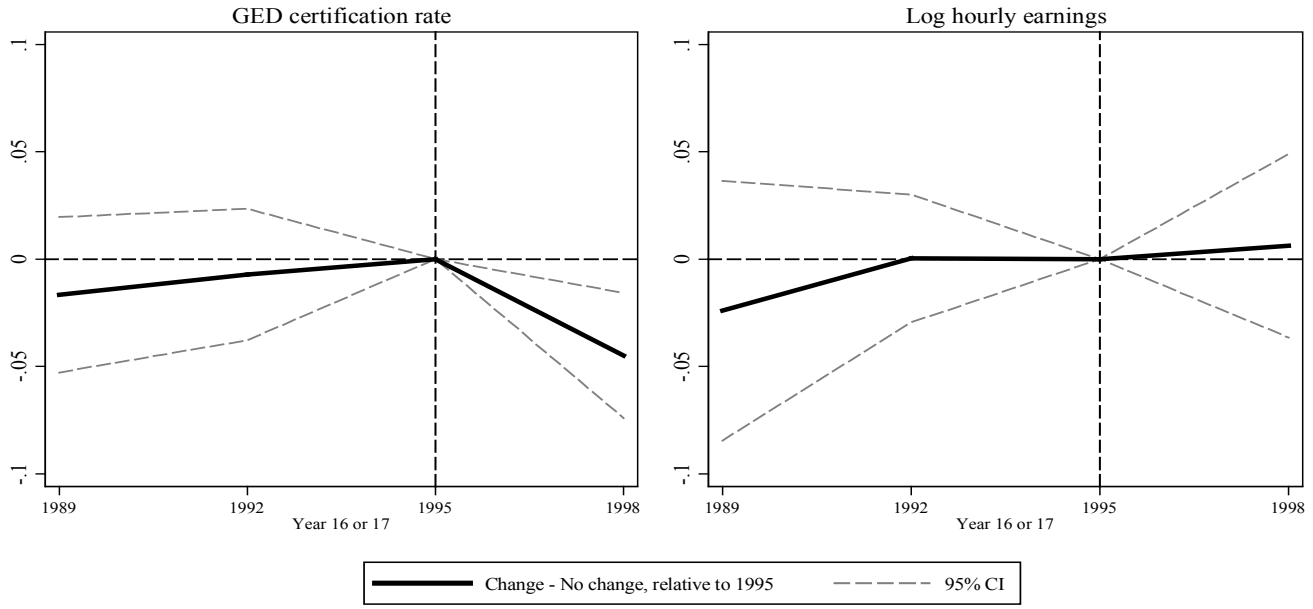


Figure 1B: Difference-in-differences estimates of the effect of the 1997 "and/or" scoring change



Notes: Figure 1A plots average GED certification rates and 2013 log hourly earnings, by birth year, for a sample of employed individuals born in states that were and were not required to eliminate the "and/or" scoring option on the GED in 1997. Figure 1B plots growth in these variables relative to the cohort that was age 16 or 17 in 1995. 95% confidence intervals are based on robust standard errors that cluster by birth state.

Table 1: Simulated first-stage and reduced-form effects of a GED scoring change

	First stage (complier shares)			Reduced form
	D (all compliers)	1-A (dropout counterfactual)	1-B (high school diploma counterfactual)	Y (log hourly earnings)
	(1)	(2)	(3)	(4)
		A. Single-treatment IV		
Full sample	0.055			-0.012
		B. Two-treatment IV		
Dropout-constrained stratum (X=0)		0.005	0.050	-0.019
Unconstrained stratum (X=1)		0.073	0.029	0.009

Notes: This table reports moments from the simulated model of GED effects inspired by Heckman and Urzúa (2010) and described in the text. Column 4 reports reduced-form effects of the instrument, a decrease in GED passing standards by 0.75 standard deviations, on log hourly earnings in the full sample (Panel A) and in two subsamples differentiated by the average difficulty of dropping out of high school (Panel B). Column 1 reports first-stage effects of the instrument on an indicator for completing the GED in the full sample, while columns 2 and 3 report first-stage effects on 1-A and 1-B in each subsample, where A indicates a student dropping out of high school and B indicates high school completion. The single-treatment IV coefficient is -0.209, the ratio of reduced-form to first-stage effects in Panel A. The two-treatment (isoLATE) IV coefficients are 0.277 and -0.396, the inverse of the first-stage matrix in Panel B post-multiplied by the reduced-form vector.

Table 2: Estimated and simulated returns to GED certification

	ACS data	Data calibrated to the Heckman and Urzúa (2010) model	
	isoLATE estimates	Population LATEs	isoLATE monte carlo
	(1)	(2)	(3)
		A. Single-treatment IV	
All compliers	-0.121	-0.209	-0.207
	(0.268)		(0.109)
		B. Two-treatment IV	
Dropout counterfactual compliers	0.150	0.277	0.257
	(0.506)		(0.418)
High school diploma counterfactual compliers	-0.347	-0.396	-0.392
	(0.208)		(0.203)

Notes: Column 1 of this table reports estimates of local average treatment effects in a sample of 22,923 employed individuals who were either 16 or 17 in either 1995 or 1998. The outcome is 2013 log hourly earnings. A cohort indicator and state of birth and residency indicators are included as controls, with the interaction of cohort and an indicator for being born in a state subject to a "and/or" score change in 1997 as the excluded instrument. The isoLATE stratification is by those born in the earlier vs. later year of their cohort. Column 2 reports corresponding moments of the model parameterized according to Heckman and Urzua (2010) and described in the text, while column 3 reports average IV estimates of these moments from 500 monte carlo replications of the two-treatment IV specification (N=100,000). Robust standard errors, clustered by birth state, are reported in parentheses in column 1; estimate standard deviations are reported in parentheses in column 3.

Table 3: Estimated Medicaid effects from the Oregon Health Insurance Experiment

	Estimation		Difference
	Restricted IV	isoLATE IV	
	(1)	(2)	(3)
A. Healthcare access			
Have usual place of clinic-based care	0.335	0.397	-0.062
	(0.073)	(0.140)	(0.096)
Have personal doctor	0.264	0.184	0.080
	(0.069)	(0.147)	(0.113)
Got all needed medical care, last six	0.266	0.215	0.051
	(0.061)	(0.120)	(0.088)
Got all needed drugs, last six months	0.242	0.199	0.043
	(0.054)	(0.096)	(0.069)
Didn't use ER for nonemergency, last six months	-0.037	-0.080	0.043
	(0.040)	(0.086)	(0.064)
B. Healthcare utilization			
Using prescription drugs currently	-0.040	-0.107	0.067
	(0.078)	(0.185)	(0.145)
Outpatient visits, last six months	0.199	0.159	0.040
	(0.066)	(0.120)	(0.075)
ER visits, last six months	0.038	0.043	-0.005
	(0.063)	(0.104)	(0.054)
Inpatient hospital admissions, last six months	0.046	0.085	-0.039
	(0.041)	(0.088)	(0.065)
C. Financial strain			
Any out of pocket medical expensis,	-0.204	-0.223	0.019
	(0.069)	(0.119)	(0.068)
Owe money for medical expenses	-0.257	-0.253	-0.005
	(0.068)	(0.117)	(0.067)
Borrowed/skipped bills to pay medical bills, last six months	-0.196	-0.155	-0.041
	(0.066)	(0.118)	(0.075)
Refused treatment because of medical debt, last six months	-0.015	-0.005	-0.010
	(0.039)	(0.056)	(0.027)
D. Health outcomes			
Heath good/very good/excellent	0.225	0.192	0.033
	(0.071)	(0.129)	(0.079)
Health not poor	0.113	0.148	-0.035
	(0.046)	(0.092)	(0.063)
Health same or better, last six months	0.225	0.225	0.000
	(0.063)	(0.107)	(0.059)

Notes: This table reports 2SLS estimates of the effects of Medicaid using randomized Medicaid offers from the Oregon Health Insurance Experiment as instruments. Columns 1 and 4 use a single treatment variable, restrict estimation to those individuals with valid survey responses for each outcome, and weight by the inverse probability of intensive follow-up, as in Finkelstein et al.(2012). Columns 1 and 4 are estimated with two endogenous variables as described in the text using the full sample of 9,770 single-person households in the 7th survey wave. Robust standard errors are reported in parentheses.

References

Abadie, A. (2003): "Semiparametric Instrumental Variables Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231–263.

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2016): "Charters Without Lotteries: Testing Takeovers in New Orleans and Boston," *American Economic Review*, 106(7), 1878–1920.

Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307–345.

Angrist, J. D. (1997): "Conditional Independence in Sample Selection Models," *Economics Letters*, 54(2), 103–112.

Angrist, J. D., and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," Advances in Economics and Econometrics. Cambridge University Press.

Angrist, J. D., and J. Hahn (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects," *Review of Economics and Statistics*, 86(1), 1–15.

Angrist, J. D., and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 430(90), 431–442.

Angrist, J. D., and M. Rokkanen (2016): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, forthcoming.

Behaghel, L., B. Crepon, and M. Gurgand (2013): "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial," IZA Discussion Paper 7447.

Behaghel, L., B. Crepon, M. Gurgand, and T. L. Barbanchon (2009): "Sample Attrition Bias in Randomized Experiments: A Tale of Two Surveys," IZA Discussion Paper 4162.

Cohodes, S. R. (2015): "The Long-Run Impacts of Tracking High-Achieving Students: Evidence from Boston's Advanced Work Class," Working Paper.

de Chaisemartin, C. (2015): "Tolerating Defiance? Identification of Treatment Effects without Monotonicity," Working Paper.

DiNardo, J., J. McCrary, and L. Sanbonmatsu (2006): "Constructive Proposals for Dealing with Attrition: An Empirical Example," Working Paper.

Duflo, E., R. Glennerster, and M. Kremer (2008): "Using Randomization in Development Economics Research: A Toolkit," Handbook of Development Economics 4, pp. 3895–3962. Elsevier.

Engberg, J., D. Epple, J. Imbrogno, H. Sieg, and R. Zimmer (2014): "Education Programs That Have Lotteried Admission and Selective Attrition," *Journal of Labor Economics*, 32(1), 27–63.

Feller, A., T. Grindal, L. Miratrix, and L. Page (2014): "Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Settings," Working Paper.

Frangakis, C. E., and D. B. Rubin (2002): "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Frölich, M., and M. Huber (2014): "Treatment Evaluation with Multiple Outcome Periods under Endogeneity and Attrition," IZA Discussion Paper 7972.

Gronau, R. (1974): "Wage Comparisons – A Selectivity Bias," *Journal of Political Economy*, 82(6), 1119–1143.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," vol. 5 of *Annals of Economic and Social Measurement*, pp. 475–492. NBER.

Heckman, J. J., J. E. Humphries, P. A. LaFontaine, and P. L. Rodríguez (2012): "Taking the Easy Way Out: How the GED Testing Program Induces Students to Drop Out," *Journal of Labor Economics*, 30(3), 495–520.

Heckman, J. J., and S. Urzúa (2010): "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify," *Journal of Econometrics*, 156, 27–37.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Hudson, S., P. Hull, and C. J. Liebersohn (2015): "Interpreting Instrumented Difference-in-Differences," Working Paper (available upon request).

Hull, P. (2016): "Estimating Hospital Quality with Quasi-experimental Data," Working Paper.

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

Kirkebøen, L., E. Leuven, and M. Mogstad (2016): "Field of Study, Earnings, and Self-Selection," *Quarterly Journal of Economics*, 101(3), 1057–1111.

Kline, P., and C. Walters (2015): "Evaluating Public Programs with Close Substitutes: The Case of Head Start," UC Berkeley Working Paper.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Expeirmental Analysis of Neighborhood Effects," *Econometrica*, pp. 83–119.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76(3), 1071–1102.

Little, R. J. A., and D. B. Rubin (1987): *Statistical Analysis with Missing Data*. Wiley.

Newey, W. K. (1994a): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1994b): "Series Estimation of Regression Functionals," *Econometric Theory*, 10, 1–28.

Pinto, R. (2015): "Selection Bias in a Controlled Experiment: The Case of Moving to Opportunity," Working Paper.

Reardon, S. F., and S. W. Raudenbush (2013): "Under What Assumptions Do Site-by-Treatment Instruments Identify Average Causal Effects?," *Sociological Methods Research*, 42.

Robins, J. M., and Y. Ritov (1997): "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine*, 16, 285–319.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Appendix

A.1 Proof of Lemma 1

Consider the reduced-form regression of Y on Z, conditional on X. By the excludability of Z given treatment and fallback status (Assumption 2), we can write

$$Y = Y_1 + (Y_a - Y_1)A + (Y_b - Y_1)B,$$

so that the regression coefficient on Z identifies:

$$\begin{aligned}
E[Y|Z=1,X] - E[Y|Z=0,X] =& E[Y_1 + (Y_a - Y_1)A + (Y_b - Y_1)B|Z=1,X] \\
&- E[Y_1 + (Y_a - Y_1)A + (Y_b - Y_1)B|Z=0,X] \\
=& E[Y_1|Z=1,X] - E[Y_1|Z=0,X] \\
&+ E[(Y_a - Y_1)A_1|Z=1,X] - E[(Y_a - Y_1)A_0|Z=0,X] \\
&+ E[(Y_b - Y_1)B_1|Z=1,X] - E[(Y_b - Y_1)B_0|Z=0,X] \\
=& E[(Y_a - Y_1)(A_1 - A_0)|X] + E[(Y_b - Y_1)(B_1 - B_0)|X] \\
=& E[Y_1 - Y_a|A_1 < A_0, X]Pr(A_1 < A_0|X) \\
&+ E[Y_1 - Y_b|B_1 < B_0, X]Pr(B_1 < B_0|X),
\end{aligned}$$

where the third equality follows by independence of Z given X (Assumption 1) and the fourth by monotonicity (Assumption 3). Furthermore, the conditional first-stage regression of D on Z is

$$\begin{aligned}
E[D|Z=1,X] - E[D|Z=0,X] &= E[D_1 - D_0|X] \\
&= E[(1 - A_1 - B_1) - (1 - A_0 - B_0)|X] \\
&= E[A_0 - A_1|X] + E[B_0 - B_1|X] \\
&= Pr(A_1 < A_0|X) + Pr(B_1 < B_0|X).
\end{aligned}$$

This again follows by Assumptions 1 and 3. The conditional IV coefficient on D is the ratio of reduced-form to first-stage expressions, completing the proof □

A.2 Proof of Proposition 1

The proof to Lemma 1 shows that under Assumptions 1-3 the conditional reduced form identifies

$$E[Y|Z=1,X] - E[Y|Z=0,X] = \alpha(X)f_a(X) + \beta(X)f_b(X).$$

Furthermore, the conditional first-stage regressions for $1-A$ and $1-B$ are

$$\begin{aligned} E[1-A|Z=1,X] - E[1-A|Z=0,X] &= -E[A_1 - A_0|X] \\ &= Pr(A_1 < A_0|X) \\ &= f_a(X) \end{aligned}$$

and

$$E[1-B|Z=1,X] - E[1-B|Z=0,X] = f_b(X).$$

As with Lemma 1, these follow from Assumptions 1 and 3.

Consider the multiple-endogenous variable IV regression of equations (3)-(5). Let $\mathbf{Y}$ denote a vector of observations of Y, $\mathbf{X}$ a matrix of observations of $1-A$ and $1-B$, and $\mathbf{Z}$ a matrix of observations of Z and ZX. The endogenous regressor coefficients then satisfy:

$$\begin{aligned} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} &= p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) \\ &= p\lim\left(((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}})^{-1}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right), \end{aligned}$$

where $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{X}}$ are matricies of residuals from regressing $\mathbf{Z}$ and $\mathbf{X}$ on X and a constant. By above,

$$\begin{aligned} p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\widetilde{\mathbf{X}}\right) &= \begin{bmatrix} f_a(0) & f_b(0) \\ f_a(1) & f_b(1) \end{bmatrix} = \Pi \\ p\lim\left((\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'\mathbf{Y}\right) &= \begin{bmatrix} \alpha(0)f_a(0) + \beta(0)f_b(0) \\ \alpha(1)f_a(1) + \beta(1)f_b(1) \end{bmatrix}. \end{aligned}$$

When Π is invertible, the continuous mapping theorem and Slutsky's theorem imply:

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} f_a(0) & f_b(0) \\ f_a(1) & f_b(1) \end{bmatrix}^{-1} \begin{bmatrix} \alpha(0)f_a(0) + \beta(0)f_b(0) \\ \alpha(1)f_a(1) + \beta(1)f_b(1) \end{bmatrix}.$$

Proposition 1 follows from simplifying this expression. □

A.3 Proof of Proposition 2

Define for each $V \in \{A, B, Y\}$

$$\begin{aligned}\delta^V_{W,X} &= E[V|Z=0,W,X] - E[V|Z=1,W,X] = \frac{E[(1-Z)V|W,X]}{1-E[Z|W,X]} - \frac{E[ZV|W,X]}{E[Z|W,X]} \\ &= E\left[\frac{E[Z|W,X]-Z}{E[Z|W,X](1-E[Z|W,X])}V|W,X\right] \\ &= E[\lambda V|W,X],\end{aligned}$$

and note that by Assumptions $1'$ and $2'$ we have

$$\begin{aligned}\delta^A_{W,X} &= E[A_0 - A_1|W,X] \\ \delta^B_{W,X} &= E[B_0 - B_1|W,X] \\ \delta^Y_{W,X} &= E[(Y_a - Y_1)(A_0 - A_1) + (Y_b - Y_1)(B_0 - B_1)|W,X].\end{aligned}$$

Next define the random vector

$$\begin{bmatrix}\alpha_W \\ \beta_W\end{bmatrix} = \begin{bmatrix}\delta^A_{W,0} & \delta^B_{W,0} \\ \delta^A_{W,1} & \delta^B_{W,1}\end{bmatrix}^{-1}\begin{bmatrix}\delta^Y_{W,0} \\ \delta^Y_{W,1}\end{bmatrix}.$$

Here α_W and β_W are conditional analogues of the multiple endogenous variable IV specification used in Proposition 1. Focusing on the first, we may write

$$\alpha_W = \frac{\delta^Y_{W,0}\delta^B_{W,1} - \delta^Y_{W,1}\delta^B_{W,0}}{\delta^A_{W,0}\delta^B_{W,1} - \delta^A_{W,1}\delta^B_{W,0}}.$$

Furthermore,

$$\begin{aligned}\delta^Y_{W,0}\delta^B_{W,1} - \delta^Y_{W,1}\delta^B_{W,0} =& E[(Y_a - Y_1)(A_0 - A_1)|W,X=0]E[B_0 - B_1|W,X=1] \\ &+ E[(Y_b - Y_1)(B_0 - B_1)|W,X=0]E[B_0 - B_1|W,X=1] \\ &- E[(Y_a - Y_1)(A_0 - A_1)|W,X=1]E[B_0 - B_1|W,X=0] \\ &- E[(Y_b - Y_1)(B_0 - B_1)|W,X=1]E[B_0 - B_1|W,X=0] \\ =& E[Y_1 - Y_a|A_1 < A_0, W, X=0]Pr(A_1 < A_0|W,X=0)Pr(B_1 < B_0|W,X=1) \\ &+ E[Y_1 - Y_b|B_1 < B_0, W, X=0]Pr(B_1 < B_0|W,X=0)Pr(B_1 < B_0|W,X=1) \\ &- E[Y_1 - Y_a|A_1 < A_0, W, X=1]Pr(A_1 < A_0|W,X=1)Pr(B_1 < B_0|W,X=0) \\ &- E[Y_1 - Y_b|B_1 < B_0, W, X=1]Pr(B_1 < B_0|W,X=1)Pr(B_1 < B_0|W,X=0) \\ =& E[Y_1 - Y_a|A_1 < A_0, W](\delta^A_{W,0}\delta^B_{W,1} - \delta^A_{W,1}\delta^B_{W,0}),\end{aligned}$$

where the second equality follows by Assumptions $3'$ and the third by Assumption $4'$. Thus

$$E[Y_1 - Y_a | A_1 < A_0, W] = \alpha_W,$$

so that, by the Law of Iterated Expectations,

$$\begin{aligned} E[Y_1 - Y_a | A_1 < A_0] &= E\left[\frac{E[(Y_1 - Y_a)(A_1 - A_0)|W]}{Pr(A_1 < A_0)}\right] \\ &= E\left[\frac{Pr(A_1 < A_0|W)}{Pr(A_1 < A_0)}\alpha_W\right]. \end{aligned}$$

Finally, note that we can write

$$\begin{aligned} \delta^V_{W,0}\delta^B_{W,1} - \delta^V_{W,1}\delta^B_{W,0} &= E[\lambda V|W, X=0]E[\lambda B|W, X=1] - E[\lambda V|W, X=1]E[\lambda B|W, X=0] \\ &= \frac{E[\lambda V(1-X)|W]E[\lambda BX|W]}{(1-E[X|W])E[X|W]} - \frac{E[\lambda VX|W]E[\lambda B(1-X)|W]}{E[X|W](1-E[X|W])} \\ &= E\left[\lambda \frac{E[\lambda BX|W] - E[\lambda B|W]X}{E[X|W](1-E[X|W])} V|W\right] \\ &= E\left[\lambda \mu_b V|W\right], \end{aligned}$$

and

$$\begin{aligned} Pr(A_1 < A_0|W) &= E[E[A_0 - A_1|W, X]|W] \\ &= E[E[\lambda \cdot A|W, X]|W] \\ &= E[\lambda A|W]. \end{aligned}$$

Thus, once again applying the Law of Iterated Expectations,

$$E[Y_1 - Y_a | A_1 < A_0] = E\left[\frac{E[\lambda A|W]}{E[\lambda A]} \frac{\lambda \mu_b Y}{E\left[\lambda \mu_b A|W\right]}\right]$$

The same steps show the result for $E[Y_1 - Y_b | B_1 < B_0]$.

Note that the function $\lambda(w, x, z)$ generating $\lambda = \lambda(W, X, Z)$ is identified by the conditional expectation function $E[Z|W, X]$ and that

$$\begin{aligned} E[\lambda A|W = w] = \sum_{x=0,1} \sum_{z=0,1} &\lambda(w, x, z) E[A|W = w, X = x, Z = z] \\ &\times Pr(Z = z|W = w, X = x) Pr(X = x|W = w) \end{aligned}$$

and similarly for $E[\lambda B|W]$. Moreover,

$$E[\lambda BX|W=w] = \sum_{z=0,1} \lambda(w,1,z)E[B|W=w,X=1,Z=z]Pr(Z=z|W=w,X=1)E[X|W=w].$$

Thus both the weights $E[\lambda A|W=w]/E[\lambda A]$ and the function $\mu_b(w,x)$ generating $\mu_b = \mu_b(W,X)$ are identified by the conditional expectation functions $E[X|W]$, $E[Z|W,X]$, $E[A|W,X,Z]$, and $E[B|W,X,Z]$. Finally, note that

$$\begin{aligned} E[\lambda\mu_b A|W=w] = \sum_{x=0,1}\sum_{z=0,1} &\lambda(w,x,z)\mu_a(w,z)E[A|W=w,X=x,Z=z] \\ &\times Pr(Z=z|X=x,W=w)Pr(X=x|W=w) \end{aligned}$$

We can thus form sample analogues of the weighting schemes identifying $E[Y_1 - Y_a|A_1 < A_0]$ from non-parametric estimates of these conditional expectation functions. The same result follows for $E[Y_1 - Y_b|B_1 < B_0]$. □

A.4 GED Selection Model

Section 3.1 simulates data on educational attainment and labor market returns using a model inspired by Heckman and Urzúa (2010). Potential log hourly earnings are given by

$$Y_{ti} = \mu_t + \gamma X_i + \epsilon_{it}, \tag{35}$$

where $X_i = 1$ is a cohort indicator and $t \in \{1, a, b\}$ indexes the individual's educational status: GED-certified, high school dropout, or traditional high school graduate. Individuals observe the schooling environment and chooses the alternative t that maximizes ν_{ti}, where

$$\nu_{1i} = \Phi(\pi_1 \tilde{Z}_{1i} - \eta_{1i}) \tag{36}$$

$$\nu_{ai} = \Phi(\pi_a \tilde{Z}_{ai} - \eta_{ai}) \mathbf{1}[X_i \geq \xi_i] \tag{37}$$

$$\nu_{bi} = \Phi(\pi_b \tilde{Z}_{bi} - \eta_{bi}), \tag{38}$$

and where $\Phi(\cdot)$ denotes the normal CDF. That is, individuals choose the schooling level that gives them the highest latent utility, subject to the constraint that some may not be allowed to drop out of high school by virtue of being too young ($\mathbf{1}[X_i \leq \xi_i]$). To simulate the model, I let $(\tilde{Z}_{1i}, \tilde{Z}_{ai}, \tilde{Z}_{bi}) \sim N(\mu_Z, \Sigma_Z)$ and $(\epsilon_{1i}, \epsilon_{ai}, \epsilon_{bi}, \eta_{1i}, \eta_{ai}, \eta_{bi}) \sim N(0, \Sigma_{\epsilon\nu})$ where

$$\Sigma_Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_{\epsilon\nu} = \begin{bmatrix} 0.64 & 0.16 & 0.16 & 0.024 & -0.32 & 0.016 \\ 0.16 & 1 & 0.2 & 0.02 & -0.3 & 0.01 \\ 0.16 & 0.2 & 1 & 0.02 & -0.4 & 0.04 \\ 0.024 & 0.02 & 0.02 & 1 & 0.6 & 0.1 \\ -0.32 & -0.3 & -0.4 & 0.6 & 1 & 0.2 \\ 0.016 & 0.01 & 0.04 & 0.1 & 0.2 & 1 \end{bmatrix}, \tag{39}$$

and where $(\mu_1, \mu_a, \mu_b) = (0.3, 0.1, 0.7)$ and $(\pi_1, \pi_a, \pi_b) = (0.2, 0.3, 0.1)$. With $X_i = \xi_i = 0$, this model is the same as the one in Heckman and Urzúa (2010). To generate cross-strata first stage variation I let $\xi \sim N(0.5, 0.025)$ and draw X uniformly with probability 0.5. Setting $\gamma = 0.2$ allows an individual's cohort to affect the level of her adulthood wages, but not her relative returns to schooling. As in Heckman and Urzúa (2010), I apply Proposition 1 with an instrument Z_i that represents an exogenous increase in $\tilde{Z}_{1i}$ by 0.75 standard deviations.