Discussion Papers
Statistics Norway
Research department

No. 703
September 2012

Statistics Norway

Christian N. Brinch, Magne Mogstad and Matthew Wiswall

Beyond LATE with a discrete instrument

Heterogeneity in the quantity-quality interaction of children

Statistisk sentralbyrå

Discussion Papers No. 703, September 2012
Statistics Norway, Research Department

Christian N. Brinch, Magne Mogstad and Matthew Wiswall

Beyond LATE with a discrete instrument
Heterogeneity in the quantity-quality interaction of children

Abstract:
The interpretation of instrumental variables (IV) estimates as local average treatment effects (LATE) of instrument-induced shifts in treatment raises concerns about their external validity and policy relevance. We examine how to move beyond LATE in situations where the instrument is discrete, as it often is in applied research. Discrete instruments do not give sufficient support to identify the full range of marginal treatment effects (MTE) in the usual local instrumental variable approach. We show how an alternative estimation approach allows identification of richer specifications of the MTE when the instrument is discrete. One result is that the alternative approach identifies a linear MTE model even with a single binary instrument. Although restrictive, the linear MTE model nests the standard IV estimator: The model gives the exact same estimate of LATE while at the same time providing a simple test for its external validity and a linear extrapolation. Another result is that the alternative approach allows identification of a general MTE model under the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects. We apply these identification results to empirically assess the interaction between the quantity and quality of children. Motivated by the seminal quantity-quality model of fertility, a large and growing body of empirical research has used binary instruments to estimate LATEs of family size on child outcomes. We show that the effects of family size are both more varied and more extensive than what the LATEs suggest. Our MTE estimates reveal that the family size effects vary in magnitude and even sign, and that families act as if they possess some knowledge of the idiosyncratic effects in the fertility decision.

Keywords: Local average treatment effects, marginal treatment effects, discrete instrument, quantity-quality, fertility

JEL classification: C26, J13

Acknowledgements: We would like to thank Rolf Aaberge, Pedro Carneiro, Cathy Redmond and Terje Skjerpen for useful comments. While carrying out this research, Brinch and Mogstad have been associated with the centre of Equality, Social Organization, and Performance (ESOP) at the Department of Economics at the University of Oslo. ESOP is supported by the Research Council of Norway.

Address: Christian N. Brinch: Statistics Norway; Dept. of Economics, University of Oslo, Email: cnb@ssb.no,

Magne Mogstad: Dept. of Economics, University College London; Statistics Norway. Email: magne.mogstad@gmail.com

Matthew Wiswall: W.P. Carey School of Business, Arizona State University. Email: matt.wiswall@gmail.com

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:
Statistics Norway
Telephone: +47 62 88 55 00
E-mail: Salg-abonnement@ssb.no

ISSN 0809-733X
Print: Statistics Norway

Sammendrag

Instrumentvariabelestimater tolkes gjerne som lokale gjennomsnittlige behandlingseffekter (LATE) av endringer i behandlingsstatus indusert av det spesifikke instrumentet som har vært brukt i estimeringen. Denne tolkningen gir opphav til spørsmål knyttet til den eksterne validiteten og politikkrelevansen av estimatene. Vi undersøker her hvordan en kan komme lenger enn å estimere LATE i situasjoner der instrumentene er diskrete, slik de gjerne er i anvendt forskning. Diskrete instrumenter gir ikke tilstrekkelig dekning til å fullt ut identifisere marginale behandlingseffekter (MTE) med den lokale instrumentvariabelmetoden. Vi viser hvordan en alternativ estimeringsmetode lar oss identifisere rikere spesifikasjoner av MTE med diskrete instrumenter. Et resultat er at den alternative fremgangsmåten identifiserer en lineær MTE-modell selv med et enkelt binært instrument. Selv om modellen er restriktiv, inneholder estimatoren av den lineære MTE-modellen den vanlige IV estimatoren: Modellen gir opphav til eksakt samme estimat av LATE, samtidig som den gir en test av ekstern validitet og en lineær ekstrapolasjon. Et annet resultat er at den alternative metoden gir identifikasjon av en generell MTE-model under en ekstra antakelse om additiv separabilitet mellom effektene av observert og uobservert heterogenitet. Vi anvender disse resultatene til å undersøke interaksjonen mellom kvantitet og kvalitet i foreldres investeringer i barn. Motivert av den klassiske kvantitet-kvalitetsmodellen av fruktbarhet, har en stor og voksende gren av empirisk forskning brukt binære instrumenter til å estimere LATE av familiestørrelse på utfall hos barn. Vi viser at effektene av familiestørrelse er både mer varierende og større enn hva LATE-resultatene indikerer. Våre MTE-estimater viser at effekten av familiestørrelse varierer både i størrelsesorden og fortegn, slik at familiene oppfører seg som om de har noe kunnskap om effekten av flere barn på barnas utfall i sin egen familie, når de beslutter om de skal få flere barn.

1 Introduction

Many empirical papers use instrumental variables estimators (IV) to estimate a model of the following type

$$y = \mu + \beta D + X^{'}\delta + \epsilon, \tag{1}$$

where y is the dependent variable, X is a vector of covariates, D is the binary regressor of interest, and ϵ is the error term. The standard problem of selection bias (D correlated with ϵ conditional on X) is solved with a valid instrumental variable Z. Influential work by Imbens and Angrist (1994) has clarified the interpretation of IV estimates as local average treatment effects (LATE) when β is a random coefficient. With selection on gains (β is correlated with D), the LATE is only informative about the average causal effect of a specific instrument-induced shift in D. In general, agents induced to treatment by Z need not be the same agents induced to treatment by a given policy change, and the average β of the two groups can differ substantially. In addition, the LATE identified by a particular instrument will generally differ from conventional treatment parameters, such as the average treatment effect (ATE) and the average treatment effect on the treated (ATT).

To move beyond the LATE, Heckman and Vytlacil (1999, 2005, 2007) generalize the marginal treatment effects (MTE) introduced by Bjorklund and Moffitt (1987). The MTE has several useful features: (1) it plays the role of a functional that is invariant to the choice of instrument; (2) it has an attractive economic interpretation as a willingness to pay parameter for persons at a margin of indifference between participating in an activity or not; and (3) all conventional treatment parameters can be expressed as different weighted averages of MTE. Using the method of local instrumental variables (LIV), the MTE can be identified and estimated under the standard IV assumptions of conditional independence and monotonicity (see Vytlacil, 2002; Heckman, 2010). However, non-parametric identification of the full set of MTEs requires an instrument that generates continuous support on the probability of treatment $P(Z)$ from 0 to 1 for each value of X. In practice, however, instruments are often discrete, and many are binary. In such situations, auxiliary assumptions are needed to identify the MTE over the full unit interval, and to recover conventional treatment parameters.

This paper contributes by examining how to move beyond the LATE in situations with discrete instruments. We begin by showing that a polynomial MTE function of order $(N-1)$ can be identified under the standard IV assumptions when $P(Z)$ takes N different values for each value of X. One key implication is that a linear MTE model can be identified even with a single binary instrument. Although restrictive, the estimator based on the linear MTE model nests the standard IV estimator: The model gives the exact same estimate of LATE, while at the same time providing a simple test for its external validity and a linear extrapolation. Specifically, if the slope in the linear MTE model is non-zero so that the MTEs are non-constant, we reject the external validity of the LATE. In such cases, a given IV estimate is only informative about the instrument-induced effect of treatment.[1]

In some applications with discrete instruments, however, one may be reluctant to impose strong restrictions on the functional form of the MTE function. In such cases, an auxiliary assumption is required. We show that with a binary instrument and M different values of the covariates X, a polynomial MTE function of order M can be identified under the standard IV assumptions and the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects. Although restrictive, this auxiliary assumption is implied by additive separability between D and X, as imposed in equation (1), which is standard in applied work using IV.

Our identification results are based on an alternative estimation approach to the conventional LIV method. In the LIV approach, the MTE is identified by differentiating $E(Y \mid X = x; P(Z) = p)$ with respect to p, which can be computed over the empirical support of $P(Z)$ conditional on X. With a binary instrument, $P(Z)$ takes only two values for each value of X, and LIV cannot identify even a linear MTE function. The alternative approach, however, identifies the MTE from separately estimating $E(Y \mid X = x; P(Z) = p, D = 1)$ and $E(Y \mid X = x; P(Z) = p, D = 0)$. With a binary instrument, the advantage of the alternative estimation approach is that we have, for each value of X, two values of $P(Z)$ for the treated (always-takers vs. always-takers and compliers) and two values of $P(Z)$ for the untreated (never-takers vs. never-takers and compliers).[2] The additional information allows us to use a binary instrument to (i) estimate a linear MTE function under the standard IV assumptions, (ii) test the external validity of LATE, and (iii) estimate a general MTE function under the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects.[3]

We apply these identification results to empirically assess the interaction between

[1] Note that our test requires only a single binary instrument. In contrast, the approaches to test the external validity of LATE proposed by Angrist and Fernandez-Val (2010), Heckman, Schmierer, and Urzua (2010), and Heckman and Schmierer (2010) require either two (or more) instruments or one instrument that takes on multiple values. Our test is therefore a particularly useful complement in applications with a binary instrument.

[2] In the terminology of Angrist, Imbens, and Rubin (1996), the treated consist of compliers whose behavior is affected by the binary instrument at hand and always-takers who are treated irrespective of whether the instrument is switched off or on; the untreated are likewise composed of compliers and never-takers, where the latter group avoids treatment even when the instrument is switched on.

[3] See Heckman and Vytlacil (2007) and Carneiro and Lee (2009) for a discussion of the alternative estimation approach in situations with an instrument that generates continuous support on the probability of treatment $P(Z)$ from 0 to 1 for each value of X. With such instruments, Heckman and Vytlacil (2007) show that the alternative estimation approach can non-parametrically identify MTE over the full unit interval, while Carneiro and Lee (2009) use the approach to estimate the distribution of potential outcomes.

the quantity and quality of children. Motivated by the seminal quantity-quality (QQ) model of fertility by Becker and Lewis (1973), a large and growing body of empirical research has examined the effect of family size on child outcomes. Much of the early literature that tested the QQ model found that larger families reduced child quality, such as educational attainment (e.g. Rosenzweig and Wolpin (1980); Hanushek (1992)). However, recent studies from several developed countries have used binary instruments, such as twin births and same-sex sibship, to address the problem of selection bias in family size. The estimated LATEs suggest that family size little effect on children's outcomes.[4]

Although these recent studies represent a significant step forward, a concern is still that the effects of family size may be both more varied and more extensive than what the IV estimates suggest. To move beyond the LATE of family size, we apply our identification results to Norwegian administrative data, using same-sex siblings and twin births as instruments. We begin by using the same-sex instrument to estimate a linear MTE function, and find that the external validity of the LATE of family size can be rejected at conventional significance levels. We next impose the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects and estimate a general MTE function. We then find that the effects of family size vary in magnitude and even sign (i.e. β is random), and that families act as if they possess some knowledge of their idiosyncratic return in the fertility decision (β is correlated with D). We next use the twins instrument to validate the MTE estimates based on the same-sex instrument, exploiting that the MTE is a functional that is invariant to the choice of instrument. Lastly, we compare the MTE weights associated with the IV estimates to the MTE weights associated with ATE and ATT, and find that the latter treatment parameters assign much more weight to the positive part of the MTE distribution. This explains why the ATE and ATT of family size are sizeable and positive, while the LATEs are smaller and sometimes negative.

The remainder of the paper is organized as follows. Section 2 presents the generalized Roy model and uses it to define MTE. This section also reviews how LIV and the separate estimation approach identify and estimate MTE with a continuous instrument. Section 3 shows how to identify and estimate MTE with a discrete instruments. Section 4 presents our empirical analysis of the effects of family size on child outcomes. Section 5 concludes.

[4]Black, Devereux, and Salvanes (2005) conclude that "there is little if any family size effect on child education" (p. 697). Using data from the US and Isreal, Caceres-Delpiano (2006) and Angrist, Lavy, and Schlosser (2010) come to a similar conclusion. However, Mogstad and Wiswall (2011) re-examine the analysis by Black, Devereux, and Salvanes (2005), and find a significant but non-linear relationship between family size and child outcomes: While a second sibling increases the educational attainment of first born children, additional children have a negative effect.

2 Framework and estimation procedures

2.1 The Generalized Roy Model and MTE

The generalized Roy model is a basic choice-theoretic framework for empirical analysis. Let Y_1 be the potential outcome of an individual in the treated state ($D = 1$), and Y_0 denote his potential outcome in the untreated state ($D = 0$).[5] The observed outcome (Y) can be linked to the potential outcomes through the switching regression model:

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

We specify the potential outcomes as

$$Y_j = \mu_j(X) + U_j, \qquad j = 0, 1 \tag{2}$$

where $\mu_1()$ and $\mu_0()$ are unspecified functions, X a random vector of covariates and U_1 and U_0 are random variables for which we normalize $E(U_1|X = x) = E(U_0|X = x) = 0$ and assume that $E(U_j^2|X = x)$ exists for $j = 0, 1$, for all x in the support of X. We allow X to be stochastically dependent on (U_1,U_0).

The individual's net benefit of receiving treatment (I_D) depends on observed variables (Z) and an unobserved component (U_D):

$$I_D = \mu_D(Z) - U_D, \tag{3}$$

where $Z = (X, Z_-)$ is a vector Z_- represents the excluded instrument(s), $\mu_D()$ is an unspecified function, and U_D is a continuous random variable with a strictly increasing distribution function. An individual selects the treated state if the net benefit of treatment is positive: $D = 1\{I_D > 0\}$. Without loss of generality, the marginal distribution of U_D can be normalized to a uniform distribution on the unit interval (Carneiro, Heckman, and Vytlacil, 2011). The function $\mu_D(Z)$ is then interpretable as a propensity score: We therefore write $P(Z) = \mu_D(Z)$ so that $D = 1$ if $P(Z) > U_D$.

The generalized Roy model allows I_D to depend on Y_0 and Y_1, which leads to dependence between (U_1,U_0) and U_D. The key assumption about the random variables is

Assumption 1 CONDITIONAL INDEPENDENCE: *(U_0, U_1, U_D) is independent of Z, conditional on X.*

The traditional approach to estimating the model of equations (2) and (3) specifies a

[5] For simplicity, we consider only a binary treatment variable, as in most of the literature on MTE. Notable exceptions include Heckman and Vytlacil (2007), Heckman, Urzua, and Vytlacil (2006) and Heckman and Urzua (2010).

parametric joint distribution of the random variables (U_0, U_1, U_D) (see e.g. Bjorklund and Moffitt, 1987). Importantly, we will not make any assumption about the joint distribution of these variables. With Z stochastically independent of (U_0, U_1, U_D) given X, the model of equations (2) and (3) implies and is implied by the standard IV assumptions of conditional independence and monotonicity (see Vytlacil, 2002; Heckman, 2010).

To define MTE, we use the following notation for the conditional expectations of U_1 and U_0:

$$k_j(p, x) = E(U_j | Z = z, U_D = p), \qquad j = 0, 1,$$

and

$$k(p, x) = E(U_1 - U_0 | Z = z, U_D = p). \tag{4}$$

By Assumption 1, the expectations of U_j are functions of z only through x.

Definition 1 *The MTE is the expected treatment effect conditional on U_D and X:*

$$MTE(x, p) = E(Y_1 - Y_0 | X = x, U_D = p) = \mu_1(x) - \mu_0(x) + k(p, x).$$

Conditioning on $U_D = p$ is equivalent to conditioning on the intersection of $P(Z) = p$ and $I_D = 0$ (indifference to the choice of treatment). The MTE is the average treatment effect for individuals with characteristics $X = x$ and $U_D = p$.

The LATE is defined within the context of the generalized Roy model as integrals over MTE (Heckman and Vytlacil, 1999, 2005, 2007). In particular, with a binary instrument $(Z_{-} \in 0, 1)$ that shifts the propensity score from $P((x, 0)) = p_0(x)$ to $P((x, 1)) = p_1(x)$, the LATE can be written as

$$\begin{aligned} LATE(x) &= \frac{E(Y|Z = (x,1)) - E(Y|Z = (x,0))}{E(D|Z = (x,1)) - E(D|Z = (x,0))} \qquad (5) \\ &= \frac{1}{p_1(x) - p_0(x)} \int_{p_0}^{p_1} MTE(x, p) dp. \end{aligned}$$

2.2 Local Instrumental Variables

Heckman and Vytlacil (1999) show how MTE can be identified and estimated using LIV. This method is a two-stage procedure. In the first stage, the propensity score is estimated as a function of Z, denoted $\hat{P}(Z)$. In the second stage one estimates the nonparametric regression: $Y = L(\hat{P}(Z), X) + \epsilon$, with ϵ an error term. The MTE is given by the derivative of L with respect to $\hat{P}(Z)$.

Conditioning on the propensity score and inserting the model for potential outcomes

(2), we obtain

$$E(Y|P(Z)=p,X=x) = (1-p)(\mu_0(x)+E(U_0|U_D>p,X=x)) + p(\mu_1(x)+E(U_1|U_D\le p,X=x)). \quad (6)$$

Since $E(U_0|X=x)=0$, we have

$$(1-p)E(U_0|U_D>p,X=x) = -pE(U_0|U_D\le p,X=x),$$

giving

$$E(U_0|U_D>p,X=x) = -\frac{p}{1-p}E(U_0|U_D\le p,X=x). \quad (7)$$

Inserting (7) into (6) gives:

$$E(Y|P(Z)=p,X=x)=\mu_0(x)+p(\mu_1(x)-\mu_0(x))+K(p,x),$$

where

$$\begin{aligned} K(p,x) &= pE(U_1-U_0|U_D\le p,X=x) \\ &= \int_0^p E(U_1-U_0|U_D=u,X=x)du \end{aligned}$$

The MTE equals the following derivative:

$$\frac{\partial E(Y|P(Z)=p,X=x)}{\partial p} = \mu_1(x)-\mu_0(x)+k(p,x),$$

with k defined in equation (4). This means that $MTE(x,p)$ is identified under Assumption 1 over the support for the treated and the untreated of $P(Z)$ conditional on X.

2.3 A Separate Estimation Approach

As an alternative to LIV, Heckman and Vytlacil (2007) use a separate estimation approach to identify the MTE. The separate estimation approach is also a two-stage procedure. As in LIV, the first stage is to estimate the propensity score as a function of Z, denoted $\hat{P}(Z)$. Unlike LIV, the second stage consists of two separate nonparametric regressions: $Y_j = L_j(\hat{P}(Z),X)+\epsilon_j$ for $j=0,1$.

To be concrete, from (2) we obtain

$$E(Y_j|P(Z)=p,X=x,D=j)=\mu_j(x)+K_j(p,x),$$

for $j=0,1$, where

$$K_1(p,x)=E(U_1|U_D\le p,X=x)$$

and

$$K_0(p,x) = E(U_0|U_D > p, X = x).$$

By differentiating K_1 and K_0 with respect to p and rearranging, we get

$$k_1(p,x) = p\frac{\partial K_1(p,x)}{\partial p} + K_1(p,x)$$

and

$$k_0(p,x) = -(1-p)\frac{\partial K_0(p,x)}{\partial p} + K_0(p,x).$$

Since

$$k(p,x) = k_1(p,x) - k_0(p,x),$$

we can, under Assumption 1, use the separate estimation to recover the function $k(p,x)$ and identify $MTE(x,p)$ over the support for the treated and the untreated of $P(Z)$ conditional on X.

3 MTE with a Discrete Instrument

With an instrument that generates full support of $P(Z)$, both LIV and the separate estimation approach non-parametrically identify MTE over the full unit interval (Heckman and Vytlacil (2007)). We now show that with a discrete instrument, the separate estimation approach allows identification of richer specifications of the MTE function than LIV. We first show how the separate estimation approach allows us to identify and estimate a parametric MTE function under the standard IV assumptions. We next demonstrate that the separate estimation approach offers a simple test for the external validity of LATE. Lastly, we show how the separate estimation approach identifies and estimates a flexible MTE function under the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects.

3.1 Identification of MTE in a non-separable model

Throughout subsections 3.1 and 3.2, we assume only that Assumption 1 (Conditional Independence) holds. Without loss of generality, we keep the conditioning on X implicit and hence take $Z = Z_{-}$.

To fix ideas, we begin with an example showing how the separate estimation approach allows us to identify a linear MTE function with a single binary instrument.

Example 1 *The following equations specify a linear MTE function:*

$$k_0(p) = \alpha_0 p - \frac{1}{2}\alpha_0$$

and

$$k_1(p) = \alpha_1 p - \frac{1}{2}\alpha_1$$

where the constant terms ensure that the marginal expectations of U_1 *and* U_0 *are zero.*

From these expressions, we derive

$$K_1(p) = \frac{1}{p}\int_0^p E(U_1|U_D = u)du = \frac{1}{2}\alpha_1(p-1)$$

and

$$K_0(p) = \frac{1}{2}\alpha_0 p$$

and

$$K(p) = \frac{1}{2}(\alpha_1 - \alpha_0)p(p-1).$$

The MTE in this case is linear in p *and given by*

$$MTE = \mu_1 - \mu_0 + \frac{1}{2}(\alpha_1 - \alpha_0) - p(\alpha_1 - \alpha_0).$$

From the expressions above, we get

$$E(Y|P(Z) = p, D = 0) = \mu_0 + \frac{1}{2}\alpha_0 p, \tag{8}$$

$$E(Y|P(Z) = p, D = 1) = \mu_1 + \frac{1}{2}\alpha_1(p-1) \tag{9}$$

and

$$E(Y|P(Z) = p) = \mu_0 + p(\mu_1 - \mu_0) + \frac{1}{2}p(1-p)(\alpha_1 - \alpha_0). \tag{10}$$

Assume that $Z \in \{0, 1\}$*, such that* $P(Z = 1) = p_1$ *and* $P(Z = 0) = p_0$*, with* $p_1 \in (0, 1)$ *and* $p_0 \in (0, 1)$*.*

Recall that LIV is based on the integrated MTE in equation (10). Although the MTE function is linear in p*, equation (10) is quadratic in* p*. With a binary instrument, the empirical analog of* $E(Y|P(Z) = p)$ *is only observed for two different values of* p*. Thus, LIV does not identify a linear MTE function with a binary instrument.*

The separate estimation approach is based on equations (8) and (9). Both equations are linear in p*. With a binary instrument, the empirical analogs of* $E(Y|P(Z) = p, D = 1)$ *and* $E(Y|P(Z) = p, D = 0)$ *are observed for two different values of* p*. Thus, the separate estimation approach identifies a linear MTE function with a binary instrument.*

Geometry of linear MTE and LATE

Figure 1 illustrates the basic geometry of the linear MTE model and how it relates to LATE. The y-axis measures the outcome of interest, whereas the x-axis measures p. Recall that U_D has been normalized to be unit uniform, so that tracing MTE over the unit interval shows how the effect of treatment vary with different quantiles of the unobserved component of selection into treatment.

In this example, we consider a binary instrument with associate propensity score values of $p_1 = 0.8$ and $p_0 = 0.4$. In the data, we observe the average outcome for each combination of treatment state and value of the instrumental variable. Indicated by circles are the four conditional averages: $Y_1(0.8) = E(Y|D = 1, P(Z) = 0.8)$, $Y_1(0.4) = E(Y|D = 1, P(Z) = 0.4)$, $Y_0(0.8) = E(Y|D = 0, P(Z) = 0.8)$, and $Y_0(0.4) = E(Y|D = 0, P(Z) = 0.4)$. The dashed line that goes through the two conditional averages for the treated observations identifies the line $\mu_1 + K_1(p)$. The dashed line that goes through the two conditional averages for the untreated observations identifies the line $\mu_0 + K_0(p)$. The solid line $\mu_1 + k_1(p)$ has twice the slope as the dashed line $\mu_1 + K_1(p)$. The solid line $\mu_0 + k_0(p)$ has twice the slope as the dashed line $\mu_0 + K_0(p)$. Note that $k_0(1) = K_0(1)$ and $k_1(0) = K_1(0)$.

The MTE is given by the vertical difference between the solid lines at a given value $U_D = p$, i.e. $MTE(p) = \mu_1 - \mu_0 + k_1(p) - k_0(p)$. In this example, the MTE is negative for $U_D < 0.5$ and positive for $U_D > 0.5$. If the MTEs were constant (i.e. no heterogeneity in treatment effects), the solid lines would be parallel.

The LATE is given by the integrated MTE over the interval (p_0, p_1), which equals the vertical distance between the solid lines at the midpoint of the interval (p_0, p_1). If the MTEs were constant, the vertical distance between the solid lines would be the same at all points $U_D \in [0, 1]$. However, because the MTEs are non-constant, the different instruments will generally identify different LATEs.

Identifying MTE with a discrete instrument

Proposition 1 states the general identification result for a discrete instrument: the separate estimation approach allows identification of richer specifications of the MTE function than LIV. In terms of estimation, the MTE function can be recovered from the empirical analogs of $E(Y|P(Z) = p, D = 1)$, $E(Y|P(Z) = p, D = 0)$, and $P(Z)$ - all of which can be consistently estimated from sample data.

Proposition 1 *Suppose Assumption 1 holds. Assume that $P(Z)$ takes on N different values, $p_1, \ldots, p_N \in (0, 1)$.*

(i) Using LIV, the MTEs are identified provided k is specified as a polynomial of order no higher than $N - 2$.

(ii) Using the separate estimation approach, the MTEs are identified provided k_1 and k_0 are specified as polynomials of degree no higher than $N - 1$.

(The proof is given in appendix A.)

3.2 Extrapolating and testing the external validity of LATE

Assume that $Z \in \{0,1\}$, such that $P(1) = p_1$ and $P(0) = p_0$, with $p_1 \in (0,1)$ and $p_0 \in (0,1)$. The definition of LATE in equation (5) can be rewritten as

$$\begin{aligned} LATE &= \frac{p_1(\mu_1 + K_1(p_1)) + (1-p_1)(\mu_0 + K_0(p_1))}{p_1 - p_0} \\ &- \frac{(p_0(\mu_1 + K_1(p_0)) + (1-p_0)(\mu_0 + K_0(p_0)))}{p_1 - p_0} \end{aligned} \tag{11}$$

because

$$\int_{p_0}^{p_1} k_1(u)du = \int_{0}^{p_1} k_1(u)du - \int_{0}^{p_0} k_1(u)du = p_1 K_1(p_1) - p_0 K_1(p_0)$$

and

$$\int_{p_0}^{p_1} k_0(u)du = \int_{p_0}^{1} k_0(u)du - \int_{p_1}^{1} k_0(u)du = (1-p_0)K_0(p_0) - (1-p_1)K_0(p_1).$$

Equation (11) is useful because the linear MTE model is estimated by (i) computing the propensity scores as the sample proportions in treatment with the instrument switched on and off, and (ii) fitting the 4 parameters such that $\mu_0 + K_0(p_0)$, $\mu_0 + K_0(p_1)$, $\mu_1 + K_1(p_0)$, and $\mu_1 + K_1(p_1)$ are equal to their empirical counterparts. Hence, the estimator of LATE derived from the estimated linear MTE model can be expressed as

$$\hat{\gamma}^{LATE} = \frac{\left(\hat{p_1}\bar{Y}_1^c(\hat{p_1}) + (1-\hat{p_1})\bar{Y}_0^c(\hat{p_1})\right) - \left(\hat{p_0}\bar{Y}_1^c(\hat{p_0}) + (1-\hat{p_0})\bar{Y}_0^c(\hat{p_0})\right)}{\hat{p_1} - \hat{p_0}},$$

where $\hat{p}_z$ is the empirical analog of $P(Z = z)$ and $\bar{Y}_j^c(\hat{p_z})$ is the empirical analog of $E(Y|P(Z) = p_z, D = j)$, for $z = 0,1$ and $j = 0,1$. It then follows straightforwardly that $\hat{\gamma}^{LATE}$ is equal to the standard IV estimator:

$$\hat{\gamma}^{IV} = \frac{\bar{Y}^c(\hat{p_1}) - \bar{Y}^c(\hat{p_0})}{\hat{p_1} - \hat{p_0}}.$$

However, the separate estimation approach offers more than the standard IV estimator: A simple test for the external validity of the LATE and a linear extrapolation. Specifically, if the slope in the linear MTE function is non-zero so that the MTEs are non-constant, we reject the external validity of the LATE. In such cases, a given IV estimate is informative only insofar the instrument-induced effect of treatment is the question of interest.

The test for the external validity of LATE is simple to implement and does not require estimation of the linear MTE model. Testing the null hypothesis of constant MTE (i.e., U_1-U_0 is mean independent of U_D) versus the alternative hypothesis of linear but non-constant MTE (i.e. U_1-U_0 is a linear function of U_D) is equivalent to testing whether

$$\Delta_1 = \Delta_0, \tag{12}$$

where

$$\Delta_j = E(Y|D=j, Z=1) - E(Y|D=j, Z=0) \text{ for } j = \{0, 1\}.$$

This is a standard statistical test with known properties. It is easily seen from Figure 1 that constant MTE in the linear MTE model corresponds to equation (12). If there are covariates in the model, the test statistic can be computed conditional on X, and it is straightforward to test jointly if all of the MTEs are constant.[6]

Note that our test requires only a single binary instrument. In contrast, the approaches to test the exernal validity of LATE proposed by Angrist and Fernandez-Val (2010), Heckman, Schmierer, and Urzua (2010), and Heckman and Schmierer (2010) require either two (or more) instruments or one instrument that takes multiple values. Our test is therefore a particularly useful complement in applications with a binary instrument.

3.3 Identification of MTE with separability

Without stronger assumptions than Assumption 1, we can only identify a fairly restrictive parametric MTE function, where the degree of the flexibility depends on the support of the discrete instrument. This subsection shows how an auxiliary assumption allows us to identify a more general MTE function in the separate estimation approach.

The auxiliary assumption is:

Assumption 2 *[Separability of marginal treatment effects]*

$$E(Y_j|U_D, X=x) = \mu_j(x) + E(U_j|U_D), \qquad j = 0, 1.$$

Assumption 2 implies that the conditional expectation function of $U_1 - U_0$ as a function of U_D does not depend on X, so that MTE is additively separable in X and U_D:

$$MTE(x, p) = \mu_1(x) - \mu_0(x) + E(U_1 - U_0|U_D = p).$$

Although restrictive, Assumption 2 is implied by additive separability between D and X, as imposed in equation (1), which is a standard auxiliary assumption in applied work

[6] In comparison, testing for no selection bias is equivalent to testing whether $\Delta_1 = \Delta_0 = 0$, which implies that (U_1,U_0) is mean independent of U_D.

using IV.[7]

Proposition 2 states the usefulness of the auxiliary assumption.

Proposition 2 *Suppose Assumptions 1 and 2 hold. Assume that X takes on M different values and Z takes on N different values for each X, giving MN values of $P(Z)$, labeled $(p_1, \ldots, p_{MN}) \in \mathcal{P} = (0,1)^{MN}$.*

(i) Using LIV, the MTEs are identified with $(p_1, \ldots, p_{MN})$ a.e. in $\mathcal{P}$ provided k is specified as a polynomial of order no higher than $(N-2)M$.

(ii) Using the separate estimation approach, the MTEs are identified with $(p_1, \ldots, p_{MN})$ a.e. in $\mathcal{P}$ provided k_1 and k_2 are specified as polynomials of order no higher than $(N-1)M$.

(The proof is in Appendix A.)

The almost everywhere (a.e.) condition in Proposition 2 is necessary because, even if we require all the p's to be different, there exist particular combinations of the p's such that the parameters will not be uniquely identified. An illustration is given below, in Example 2. We conjecture that this possibility of non-identification has little empirical relevance.

An important implication of Proposition 2 is that with a binary instrument and M different values of the covariates X, the separate estimation approach can identify a polynomial MTE function of order M under Assumptions 1 and 2. In contrast, LIV cannot even identify a linear MTE function under the same assumptions. Example 2 illustrates the differences across the estimation procedures in the simple case of a single binary X.

Example 2 *Consider first the case without any covariates. The following equations specify a quadratic MTE function:*

$$k_0(u) = \alpha_{01}u + \alpha_{02}u^2 - \frac{1}{2}\alpha_{01} - \frac{1}{3}\alpha_{02}$$

and

$$k_1(u) = \alpha_{11}u + \alpha_{12}u^2 - \frac{1}{2}\alpha_{11} - \frac{1}{3}\alpha_{12}$$

where the constant terms ensure that the marginal expectations of U_1 and U_0 are zero.

From these expressions, we derive

$$K_0(p) = \frac{1}{2}\alpha_{01}p + \frac{1}{3}\alpha_{02}p(p+1),$$

$$K_1(p) = \frac{1}{2}\alpha_{11}(p-1) + \frac{1}{3}\alpha_{12}(p^2-1)$$

[7] In fact, Assumption 2 is weaker, as it allows the treatment effects to vary by X and U_D, though not by the interaction of the two terms. In contrast, additive separability between D and X assumes that the treatment effects are the same for all values of X.

and

$$K(p) = \frac{1}{2}(\alpha_{11} - \alpha_{01})p(p-1) + \frac{1}{3}(\alpha_{12} - \alpha_{02})p(p^2-1).$$

As shown in Proposition 1, with only a binary instrument, neither LIV nor the separate estimation approach will identify the quadratic MTE function.

Suppose we introduce a single binary covariate to the model. With a binary instrument, Assumptions 1 and 2 now give us four different values of p for the treated and the untreated. At the same time, we have additional parameters that we need to estimate since the model allows the $\mu_1(X)$ and $\mu_0(X)$ terms to vary with X.

The LIV approach is based on the equation

$$E(Y|X=x, P(Z)=p) = \mu_{00} + \mu_{01}x + p(\mu_{10} - \mu_{00}) + px(\mu_{11} - \mu_{01}) + K(p)$$

where under Assumption 2, the $K()$ function does not depend on X. From this equation, the four values of p are insufficient for identification of the model under Assumptions 1 and 2. In fact, the inclusion of X does not allow for identification of even a linear MTE function.

The separate estimation approach is based on the equations

$$E(Y|X=x, P(Z)=p, D=0) = \mu_{00} + \mu_{01}x + \frac{1}{2}\alpha_{01}p + \frac{1}{3}\alpha_{02}p(p+1) \tag{13}$$

and

$$E(Y|X=x, P(Z)=p, D=1) = \mu_{10} + \mu_{11}x + \frac{1}{2}\alpha_{11}(p-1) + \frac{1}{3}\alpha_{12}(p^2-1). \tag{14}$$

In each equation, we have four parameters and data that allow us to evaluate the expectation for four values of p. This shows that under Assumptions 1 and 2, the separate estimation approach identifies a quadratic MTE function with a binary Z_{-} and a binary X.

There is one exception to the conclusion in the above paragraph - which illustrates the reason for the a.e. condition in Proposition 2. Explicit specification of the linear equation system necessary to solve for the parameters in (13) and (14) shows that the parameters are uniquely identified if

$$\begin{vmatrix} 1 & 1 & p_1 & p_1^2 \\ 1 & 1 & p_2 & p_2^2 \\ 1 & 0 & p_3 & p_3^2 \\ 1 & 0 & p_4 & p_4^2 \end{vmatrix} = (p_2 - p_1)(p_4 - p_3)(p_4 + p_3) - (p_2 - p_1)(p_4 - p_3)(p_2 + p_1) \neq 0,$$

where p_1 and p_2 are the two propensity scores associated with $X = 1$, and p_3 and p_4 are the two propensity scores associated with $X = 0$. Proposition 2 assumes that the propensity *scores differ for each value of X, so that $p_1 \neq p_2$ and $p_3 \neq p_4$. The system will then have a unique solution, except if $p_1 + p_2 = p_3 + p_4$ which is the reason for the a.e. condition in Proposition 2. Although this may happen by chance, we conjecture that there will be a unique solution in most empirical applications. The exception is if Z_- is generated from a randomized controlled trial with perfect compliance.*

3.4 Weights on MTE for conventional treatment parameters

Heckman and Vytlacil (2005, 2007) show that conventional treatment parameters, such as LATE, ATE, ATT, and the average treatment effect on the untreated (ATUT), can be expressed as different weighted averages of the MTE. Specifically they show that treatment parameter j for a given X, denoted $\Delta_j(x)$, can be written as:

$$\Delta_j(x) = \int_0^1 MTE(x,u)h_j(x,u)du,$$

where the weights can be consistently estimated from sample data. The population treatment parameter, Δ_j, is simply the weighted sum of covariate-specific treatment parameters, $\Delta_j(x)$.

The formulas for weights derived by Heckman and Vytlacil assume that the MTEs are estimated separately for each value of X. In practice, however, researchers rarely estimate covariate-specific treatment parameters. Brinch, Mogstad, and Wiswall (2012) show how instrumental variables estimators can be expressed as different weighted averages of the MTE in situations with parametric specifications in X. In the part of the empirical analysis where we will be making parametric specifications in X, we use these weight expressions. As before, the weights can be consistently estimated from sample data.

4 Empirical analysis

4.1 Data and descriptive statistics

As in Black, Devereux, and Salvanes (2005), our data are based on administrative registers from Statistics Norway covering the entire resident population of Norway who were between 16 and 74 of age at some point during the period 1986-2000. The family and demographic files are merged by unique individual identifiers with detailed information about educational attainment reported annually by Norwegian educational establishments. The data also contains identifiers that allow us to match parents to their children. As we observe each child's date of birth, we are able to construct birth order indicators for every child in each family. We refer to Black, Devereux, and Salvanes (2005) for a more detailed description of the data as well as of relevant institutional details for Norway.

We follow the sample selection used in Black, Devereux, and Salvanes (2005). We begin by restricting the sample to children who were aged at least 25 in 2000 to make it likely that most individuals in our sample have completed their education. Twins are excluded from the estimation sample because of the difficulty of assigning birth order to these children. To increase the chances of measuring completed family size, we drop families with children aged less than 16 in 2000. We exclude a small number of children with more than 5 siblings as well as a handful of families where the mother had a birth before she was aged 16 or after she was 49. In addition, we exclude a small number of children where their own or their mother's education is missing. Rather than dropping the larger number of observations where information on fathers is missing, we include a separate category of missing for father's education and father's age.

Regressors and instruments

As in Black, Devereux, and Salvanes (2005), our measure of family size is the number of children born to each mother. Throughout the empirical analysis, we follow much of the previous literature in focusing on the treatment effect on a first born child from being in a family with 2 or more siblings rather than 1 sibling. The outcome of interest is the child's years of schooling, which is often used as a proxy for child quality. The child's education is collected from year 2000, and the education of the parents is measured at age 16 of the child.

In line with much of the previous literature on family size and child outcomes, we use the following two instruments: twin birth and same-sex sibship. The twins instrument is a dummy for a multiple second birth (2nd and 3rd born children are twins). This instrument rests on the assumptions that the occurrence of a multiple birth is as good as random, and that a multiple birth affects child development solely by increasing fertility. The same-sex instrument is a dummy variable equal to one if the two first children in a family have the same sex. This instrument is motivated by the fact that parents with two children are more likely to have a third child if the first two are of the same sex than if sex-composition is mixed. The validity of the same-sex instrument rests on the assumptions that sibling sex composition is essentially random and that it affects child development solely by increasing fertility. It should be emphasized that our focus is not on the validity of these instruments: Our aim is to move beyond the LATE of family size, applying commonly used instruments.[8]

[8] See e.g. Black, Devereux, and Salvanes (2005) and Angrist, Lavy, and Schlosser (2010) for an assessment of the validity of the instruments.

Summary statistics and fertility decision model

Our sample consists of 514,049 children. Table 1 displays the basic descriptive statistics. In 50 percent of the sample, the two first children are of the same sex, whereas a twin birth at second parity occurs in about 1 percent of the families. As expected, fathers are, on average, slightly older and more educated than mothers. In 50 percent of the sample, there are at least three children in the family, and the average family size is 2.7 children.

Table 2 presents estimates of the average marginal effects from a logit model for the choice of having 3 or more children (instead of 2 children). In terms of the choice model defined by (3), I_D represents the net benefit from having more than 2 children, which is assumed to depend on an unobserved component, the covariates and the instrument(s) listed in Table 1. Recall that we do not assume that the covaraites are exogenous; all we assume is that the instruments are independent of the unobservables conditional on the covariates.

We see that the instruments are (individually and jointly) strong predictors of family size. The average effect of twin birth is about 0.52. This means that 48 percent of mothers with two or more children would have had a third birth anyway. We also see that parents of same-sex sibship are, on average, about 5.7 percentage points more likely to have a third birth than parents of mixed-sex sibship. It is also evident that families with three or more children were decreasing over the period we study, which is reflected in the sizable marginal effect of child's age in the year 2000. Mothers age at first birth is also predictive of family size: The propensity score decreases by as much as 1.6 percentage points if the mother is one year older at the first birth.

4.2 IV estimates with treatment heterogeneity

We specify the following outcome equation:

$$Y = \mu + \beta D + X'\delta + \epsilon, \tag{15}$$

where Y denotes child's years of schooling, X is a vector of controls for (pre-determined) child and parental characteristics, and ϵ is the error term. In line with much of the previous literature, we will throughout the empirical analysis focus on the treatment effect on a first born child from being in a family with 2 or more siblings ($D = 1$) rather than 1 sibling ($D = 0$).

Table 3 shows how IV estimates of the effects of family size vary in magnitude and even sign with the choice of instrument. The IV estimates reported in Column 1 are based on the first stage

$$D = \gamma + Z'\_\theta + X'\xi + v. \tag{16}$$

While the effect of family size induced by twins is only 0.051, the effect based on the same-sex instrument is as large as 0.165. The fact that the IV estimates vary with the choice of instrument indicates non-constant MTEs. When including both instruments in the first stage, we estimate that being in a family with 2 or more siblings rather than 1 sibling raises educational attainment by 0.076 years.

In Column 2, we follow Carneiro, Heckman, and Vytlacil (2011) in specifying the first stage as

$$D = \gamma + \delta P(Z) + X^{'}\xi + v, \tag{17}$$

where $P(Z) \equiv Pr(D = 1 \mid Z)$ is used as the instrument for family size. We construct $P(Z)$ using the parameter estimates from the logit model, for which average effects are reported in Table 2. We report IV estimates based on (17) for each instrument separately and when using both instruments.

Both (16) and (17) provide consistent estimates of the LATE from instrument-induced shifts in family size under the same assumptions (Carneiro, Heckman, and Vytlacil, 2011). However, as $P(Z)$ incorporates interactions between the controls and the instrument in the fertility choice, the LATE of a $P(Z)$-shift in D does not need to be same as the LATE of a Z_{-}-shift in D. Indeed, the IV estimates differ substantially across Columns 1 and 2: While the estimated LATEs based on Z_{-} are positive for every instrumental variable, the estimated LATEs based on $P(Z)$ are negative for every instrumental variable. This suggests that the MTEs vary in sign and that the IV estimates based on $P(Z)$ assign more weight to negative MTEs as compared to the IV estimates based on Z_{-}.

MTE weights of treatment parameters

As a first step towards understanding why the IV estimates vary so much with the choice of instrument, we estimate the distribution of instrument-specific weights across the support of the MTE distribution. Figure 3 displays the distribution of weights for the IV estimates, and compares them to the distribution of weights of the ATE, the ATT, and the ATUT. The y-axis measures the density of the distribution of weights, whereas the x-axis measures the unobserved component U_D of parents' net gain from having 3 or more children ($D = 1$) rather than 2 children ($D = 0$). Recall that a high value of U_D means that a family is less likely to have 3 or more children.

There are clear patterns in the distribution of weights. First, the IV estimates based on the twins instrument assign more weight to individuals with high values of U_D as compared to the same-sex instrument. This pattern is quite intuitive: With twin births there are no never-takers, so that even families very unlikely to have another child are induced to increase family size; with same-sex sibship, the complier group consists of parents whose preferences for mixed-sex sibship induce them to have a third child.[9]

Second, the distributions of weights are more skewed to the right for IV estimates using Z_{-} as the instrument as compared to those using $P(Z)$ as the instrument. The difference is particularly stark for the same-sex instrument, in which case the IV estimate based on $P(Z)$ assigns the vast majority of weight to MTEs in the interval $(0.3, 0.6)$. The large disparity in the distribution of weights for the same-sex instrument resonates well with the substantial difference in estimated LATEs based on the same-sex instrument

Third, both ATT and ATE assign much more weight to families who are likely to have 2 or more siblings as compared to the IV estimates. In contrast, ATUT and the IV estimates based on the twins instrument assign most of the weight to families unlikely to have another child. This pattern is also quite intuitive. With twins there are no never-takers, so the untreated consist only of compliers with the twins instrument switched off. If the occurrence of a twin birth is as good as random (conditional on covariates), the LATE representing the average effect for the twin birth compliers is equal to the average effect for all compliers given by ATUT. This implies that the distributions of weights with the twins instruments should mirror the distribution of weights for the ATUT.

Heterogeneity in the generalized Roy model

To fully understand what the LATEs of family size identify and why the IV estimates vary so much with the choice of instrument, we need to know the underlying distribution of MTEs. But before turning attention to the actual estimation of MTE, it can be useful to get a better sense of the pattern of heterogeneity in the relationship between the quantity and quality of children that is consistent with the generalized Roy model.

Consider first the traditional approach to estimating the model of equations (2) and (3), which assumes that (U_0, U_1, U_D) are joint normal distributed and independent of Z (see e.g. Bjorklund and Moffitt, 1987). Although this normal selection model restricts the shape of the MTE function, it is consistent with IV estimates of different magnitude and sign depending on the choice of instrument: the MTE is either constant, monotonically declining (i.e. positive selection on gains) or monotonically increasing (i.e. negative selection on gains) in U_D; the MTE tends towards $\pm\infty$ as U_D tends towards 0 or 1 (unless the MTE is constant); the distribution of MTE is symmetric in U_D, so that the slope of the MTE takes the same absolute value for $U_D = u$ and $U_D = 1 - u$.

Although the joint normality assumption is convenient, it can mask essential heterogeneity in the effects of family size if the population is segmented in preferences and/or constraints. For example, preference for mixed-sex sibships is unlikely to be manifested with equal force by all groups in the population. Mixture distributions arise naturally

[9] Angrist and Fernandez-Val (2010) characterize the complier groups and find that twins and same-sex compliers are clearly different. For example, twins compliers are likely to be college graduates, while same-sex compliers are unlikely to be college graduates.

when the population contains two or more distinct sub-populations.[10] In Figure 2, we present a simple example of MTE in a mixture model with two subpopulations of equal size. Specifically, let the unobserved component U_D of parents' net gain from having 3 or more children be generated from as a mixture of two random variables U_{D1} and U_{D2} with equal probability. We assume that U_{D1} is standard normal, while U_{D2} is normal with mean zero and variance 2. Individuals in the first subgroup have constant MTE of 1, while individuals in the second subgroup have constant MTE of -1. Figure 2 shows that the MTE derived from this mixed model has a U-shape. Individuals with high MTE are overrepresented in the tails, whereas individuals with low MTE tend to be in the middle ranges of U_D. The reason is that the first subgroup has a relatively high variance of U_D: This could, for example, be due to weaker preferences for mixed-sex sibship such that the unobserved component explains more of the variation in the choice of family size.

Lastly, we note that several sources can generate MTE of different magnitude and sign, including heterogeneity in preferences over child quality and quantity, differences in the technologies available to produce child quality, and variability in the economic resources available to families. For example, the QQ model of fertility by Becker and Lewis (1973) is consistent with both positive and negative effects of family size depending on whether quantity and quality are complements or substitutes (Rosenzweig and Wolpin, 1980). Also other theories, outside the Becker and Lewis model, suggest essential heterogeneity in the effects of family size on child outcome. In particular, for some families additional siblings may benefit existing children if they stabilize parental relationship (see e.g. Becker, 1998), make maternal employment less likely (see e.g. Ruhm, 2008), or if there are positive spillover effects among siblings (see e.g. Bandura, 1977).

4.3 MTE estimates with the same-sex instrument

This subsection shows how the separate estimation approach and our identification results can be used to move beyond the LATE of family size. We begin by estimating a linear MTE function and use it to test the external validity of LATE. We next impose the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects (Assumption 2) and estimate a general MTE function.

Linear MTE model and external validity of LATE

Consider first a linear MTE model without covariates. For now, we only use the same-sex instrument, but we will later provide estimates using both instruments. Table 4 displays the results: Panel (a) shows estimates of the intercept and the slope of the linear MTE

[10] Morduch and Stern (1997) show how a mixture model provides an empirical framework which is consistent with theoretically and empirically based concerns about population heterogeneity with regards to gender bias in fertility and child investment.

model as well as its underlying components; Panel (b) reports the LATE derived from the linear MTE model and compares it to the LATE estimated by standard IV.

We find that 53.1 percent of parents with same-sex siblings have a third child, whereas only 47.3 percent of parents with mixed-sex sibship have 3 or more children. It is also evident that first born children with same-sex siblings have slightly lower educational attainment (12.281 years of schooling) as compared to first born children with mixed-sex sibship (12.284 years of schooling). The estimated LATE of the same-sex-induced increase in family size is given by

$$\hat{\gamma}^{IV} = \frac{\bar{Y}^c(\hat{p}_1) - \bar{Y}^c(\hat{p}_0)}{\hat{p}_1 - \hat{p}_0} = \frac{12.281 - 12.284}{0.531 - 0.473} = -0.065$$

Table 4 shows that our separate estimation approach provides the exact same estimate of the LATE. To be specific, we estimate that

$$\hat{\mu_1} + \hat{K}_1(p) = 11.720 + 0.773p$$

$$\hat{\mu_0} + \hat{K}_0(p) = 12.780 - 0.216p$$

and

$$\begin{aligned} \hat{\mu_1} + \hat{k}_1(p) &= 0.773p + 11.720 + 0.773p = 11.720 + 1.546p \\ &= 11.720 + 1.546p \end{aligned}$$

$$\begin{aligned} \hat{\mu_0} + \hat{k}_0(p) &= -0.216(1-p) - (12.780 - 0.216p) \\ &= 12.780 - 0.432p. \end{aligned}$$

The last step in the separate estimation approach to derive the LATE is:

$$\hat{\mu_1} - \hat{\mu_0} + \int_{0.471}^{0.531} \hat{k}_1(u) - \hat{k}_0(u)du = -0.065.$$

This illustrates that in situations with a binary instrument, the separate estimation approach of the linear MTE model gives the exact same estimate of LATE as standard IV estimation.

However, the separate estimation approach offers more: A simple test for the external validity of the LATE. Table 4 shows that the slope of the linear MTE model is different from zero at conventional significance levels. We therefore reject the external validity of LATE, which suggests that it is only informative about the same-sex-induced effect of family size.

Recall that our test for the external validity of LATE can actually be performed without estimating the linear MTE model. Specifically, testing the null hypothesis of constant MTEs versus the alternative hypothesis of linear but non-constant MTEs is equivalent to testing whether

$$\begin{aligned} E(Y|D=1,Z=1) \quad &- \quad E(Y|D=1,Z=0) \\ &= \\ E(Y|D=0,Z=1) \quad &- \quad E(Y|D=0,Z=0), \end{aligned}$$

which is a standard statistical test with known properties. In this application, we reject the null hypothesis of a constant MTE at the 1 percent significance level (p-value 0.0001).

There is one important caveat to the rejection of the external validity of the LATE: The same-sex instrument may be correlated with other variables than family size. If these variables affect children's education then Z depends on (U_1, U_2, U_D), implying that the results reported in Table 4 are biased. We address this concern by controlling for the set of covariates listed in Table 1. Specifically, we partition our sample into 64 groups based on these covariates and estimate the linear MTE model separately for each group. Tables D-1 and D-2 reported in the Appendix display the results. Although most of the LATEs are too imprecisely estimated to draw any conclusions about the covariate-specific effects of family size, we find that the slopes of the linear MTE models are jointly different from zero at the 10 percent significance level (p-value 0.064). This suggests that the rejection of the external validity of the LATE is unlikely to be driven by differences in observables across families with same-sex and mixed-sex sibship.[11]

A flexible MTE function in a separable model

If all we are willing to assume is that (U_1, U_2, U_D) is independent of Z given X (Assumption 1), then a binary instrument identifies a linear MTE function only. This means that unless one is willing to use the linear MTE function to extrapolate, it is not possible to recover the MTE over a wide range of U_D. As an alternative to such a linear extrapolation, we proceed by invoking the auxilary assumption of additive separability between observed and unobserved heterogeneity in treatment effects (Assumption 2).

Figure 4 shows the empirical support of $P(Z) \equiv Pr(D = 1 \mid Z)$ under Assumptions 1 and 2, using same-sex as the instrument for family size. The common support is defined as the intersection of the support of $P(Z)$ given $D = 1$ and the support of $P(Z)$ given $D = 0$.

[11] The rejection of the external validity of the LATE is robust to how we partition the sample and what covariates we include. The robustness results are available upon request.

As for the IV estimates reported in Table 3, we construct $P(Z)$ using the parameter estimates from the logit model whose average derivatives are reported in Table 2. We see that the auxiliary assumption yields substantial support in the interval $(0.20, 0.75)$. We do not, however, obtain support in the tails, which implies that we cannot identify MTE as U_D approaches 0 or 1.

We proceed by semi-parametric estimation of the MTE under Assumptions 1 and 2. Our estimation procedure follows closely the approach used in Heckman, Urzua, and Vytlacil (2006) and Carneiro, Heckman, and Vytlacil (2011). The first step is the construction of the estimated $P(Z)$, and the second step is the estimation of $\mu_1(X)$ and $\mu_2(X)$ using the estimated $P(Z)$. The first step is carried out as for Figure 4. Our specification is quite flexible, and alternative functional form specifications for the choice model (e.g. probit or linear probability model) produce results similar to the ones reported here. The second step uses the method proposed by Robinson (1988) for estimating partially linear models, as extended in Heckman, Ichimura, and Todd (1997). Lastly, the functions K_1 and K_0 are estimated using local quadratic regression of $Y_1 - \hat{\mu}_1(X)$ and $Y_0 - \hat{\mu}_0(X)$ on the estimated $P(Z)$, where $\hat{\mu}_1$ and $\hat{\mu}_0$ are the estimates from the second step.[12]

Figure 5 displays how the MTE depends on U_D, with 95 percent confidence intervals computed from a non-parametric bootstrap.[13] The MTE estimates are evaluated at mean values of X. Our estimates show that the effects of family size vary in magnitude (i.e. β is random) and even sign, and that families act as if they possess some knowledge of their idiosyncratic return (β is correlated with D). As in the study of the marginal return to education by Carneiro, Heckman, and Vytlacil (2005), the MTE is clearly U-shaped and the magnitude of heterogeneity is substantial. As discussed above, this pattern for the MTE could not be uncovered with the normal selection model, but it is consistent with a mixture model where the population is segmented according to preferences and/or constraints. Specifically, our estimates show that an increase in family size raises the average educational attainment of first born children in families with U_D less than 0.40. This means that first born in families that are likely to have another child (in terms of their unobservables) would gain from an increase in family size. The family size effects are negative for values of U_D in the interval (0.40,0.62), indicating a quantity-quality tradeoff in families where preferences for mixed sibling sex composition plays a more important role in the decision to have another child. For values of U_D above 0.62, the estimated MTE is positive. This means that the educational attainment of first born in families

[12] We use rectangular kernels and choose the bandwidth that minimizes the square prediction error when the current observation is left out of the analysis. This gives us an estimate of the optimal bandwidth of 0.055.

[13] Heckman, Ichimura, and Todd (1997) show that the bootstrap provides a better approximation to the true standard errors than asymptotic standard errors for the estimation of the parameters in a model similar to the one we present here. We use 100 bootstrap replications. Throughout the paper, in each iteration of the bootstrap we re-estimate $P(Z)$ so all standard errors account for the fact that $P(Z)$ is itself an estimated object.

unlikely to have a third child will benefit from an increase in family size; However, the parents still decide not to have an additional child because the unobserved (psychic or financial) costs are too high.

We have already discussed the rejection of the hypothesis that MTE is constant in U_D, based on the estimates of the linear MTE model with and without covariates. But we can also directly test whether the semi-parametric MTE is constant in U_D or not. We evaluate the MTE in five intervals equally spaced between 0.2 and 0.75. As in Carneiro, Heckman, and Vytlacil (2011) we construct pairs of adjacent intervals and compare the mean of the MTEs for each pair. Table 5 reports the outcome of these comparisons. For example, the first column reports

$$E(Y_1 - Y_0|X = \bar{X}, 0.3 \leq U_D \leq 0.35) \quad - \quad E(Y_1 - Y_0|X = \bar{X}, 0.25 \leq U_D \leq 0.2) = 0.955.$$

The p-value of the test of the hypothesis that this difference is equal to zero is reported below and is equal to 0.051. Table 5 shows that most of the adjacent LATEs are different at conventional levels of significance. A joint test that the difference across all adjacent LATEs is different from zero has a p-value close to zero. This is further evidence that families select into family size based on heterogeneous returns.

4.4 Model validation using the twins instrument

So far, we have only used the sex instrument in the estimation of the MTE. We now use the twins instrument to validate the MTE estimates based on the same-sex instrument, exploiting that the MTE is a functional that is invariant to the choice of instrument. If the MTE estimates vary significantly with the choice of instrument, it would raise serious concerns about the validity of the instruments (or Assumption 2).

Figure 6 compares estimates of MTE based on the same-sex instrument to estimates of MTE using both the same-sex and the twins instrument. In both cases, we use the semi-parametric method described above. It is reassuring to find that the two MTE estimates display the same U-shaped pattern. Indeed, the point estimates are similar in magnitude and never statistically different. This finding suggests that the differences in the IV estimates by the choice of instrument is because of different weighting of the MTE, rather than invalidity of the instruments.

A concern with the validation exercise presented in Figure 6 is that the same-sex instrument is driving both MTE estimates. To address this concern, it would be useful to estimate the MTE separately for each instrument. However, with twins there are no never-takers, so the function k_0 and thus the MTE cannot be identified under Assumptions 1 and 2 from the twins instrument only. Nevertheless, we can use the twins instrument to estimate the function k_1 (i.e. the expected outcome as treated), since there are both always-takers and compliers. Figure 7 shows how the semi-parametric estimates of the function k_1 vary with the choice of instrument. For each instrument, we use the semi-parametric method described above. The similarity in the estimates of k_1 gives credibility to the semi-parametric MTE estimates reported in Figure 5.

4.5 Summary Measures of Treatment Effects

As shown by Heckman and Vytlacil (1999, 2005, 2007), all conventional treatment parameters can be expressed as different weighted averages of the MTE. Recovering these treatment parameters from estimates of MTE, however, requires full support of $P(Z)$ on the unit interval. Since we do not have full support of $P(Z)$, we follow Carneiro, Heckman, and Vytlacil (2011) in constructing bounds and in rescaling the weights so that they integrate to one over the region of common support.

We use the semi-parametric MTE estimates based on the same-sex instrument, reported in Figure 5, to construct rescaled estimates and lower bounds on the ATE, the ATET and the ATUT. While there are regions of $P(Z)$ with negative MTEs, the MTEs are positive and sizeable in the tails of the common support. We therefore construct the lower bounds assuming that the MTE in the region outside the common support ($U_D \in [0.20, 0.75]$) are non-negative.

Table 6 displays the lower bounds and rescaled support estimates for the ATE, ATT, and ATUT parameters. The lower bound estimates are 0.194 for the ATUT, 0.232 for the ATE, and 0.313 for the ATT. The rescaled support estimates are even larger, reflecting that no weight is given to the MTE outside the region of support. This evidence stands in stark contrast to the IV estimates reported in Table 3, which range between 0.174 and -0.208. As shown in Figure 3, the reason is that the IV estimates assign much more weight to the regions with negative MTE as compared to the ATE and ATT. This illustrates the need to be cautious in going from the mean impact of family size on compliers to the average effects on the entire population or the subpopulation of (non)treated.

5 Conclusions

The interpretation of IV estimates as effects of instrument-induced shifts in treatment raises concerns about their external validity. This paper examines how to move beyond the LATE in situations with a discrete instrument with finite support. Discrete instruments do not give sufficient support to identify the full range of marginal treatment effects (MTE) in Heckman and Vytlacil's (1999, 2005, 2007) local instrumental variable (LIV) approach. We show how an alternative estimation approach allows identification of richer specifications of the MTE with discrete instruments.

One key result is that we can identify a linear MTE model even with a single binary instrument. Although restrictive, the linear MTE model nests the standard IV estimator:

The model gives the exact same estimate of LATE while at the same time providing a simple test for its external validity and a linear extrapolation. Another key result is that the alternative estimation approach allows identification of a general MTE model under the auxiliary assumption of additive separability between observed and unobserved heterogeneity in treatment effects.

We apply these identification results to empirically assess the interaction between the quantity and quality of children. Motivated by the seminal quantity-quality model of fertility, a large and growing body of empirical research has used IV to examine the effect of family size on child outcomes. We find that the effects of family size vary in magnitude and even sign, and that families act as if they possess some knowledge of the idiosyncratic effects. We also reject the external validity of the LATEs of family size at conventional significance levels. When comparing the MTE weights associated with the IV estimates to the MTE weights associated with ATE and ATT, we found that the latter treatment parameters assign much more weight to positive MTEs. This explains why the ATE and ATT of family size are sizeable and positive, while the LATEs are smaller and sometimes negative.

References

Angrist, J., and I. Fernandez-Val (2010): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," *NBER Working Papers.*

Angrist, J., G. W. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434), 444–472.

Angrist, J., V. Lavy, and A. Schlosser (2010): "Multiple experiments for the causal link between the quantity and quality of children," *Journal of Labor Economics*, 28(4), 773–824.

Bandura, A. (1977): *Social Learning Theory.* Prentice Hall, Englewood Cliffs, NJ.

Becker, G. S. (1998): *A Treatise on the Family. Enlarged Version.* Harvard University Press, Cambridge, MA.

Becker, G. S., and H. G. Lewis (1973): "On the Interaction between the Quantity and Quality of Children," *Journal of Political Economy*, 81(2), 279–288.

Bjorklund, A., and R. Moffitt (1987): "The estimation of wage gains and welfare gains in self-selection models," *The Review of Economics and Statistics*, pp. 42–49.

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

Brinch, C. N., M. Mogstad, and M. Wiswall (2012): "Understanding instrumental variables in models with covariates," manuscript.

Caceres-Delpiano, J. (2006): "The Impacts of Family Size On Investment in Child Quality," *Journal of Human Resources*, 41(4), 738–754.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2005): "Understanding What Instrumental Variables Estimate: Estimating the Average and Marginal Return to Schooling," *Working Paper, University of Chicago.*

——— (2011): "Estimating Marginal Returns to Education," *American Economic Review*, 101(6), 2754–2781.

Carneiro, P., and S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149(2), 191–208.

Hanushek, E. A. (1992): "The Trade-off between Child Quantity and Quality," *Journal of Political Economy*, 100(1), 84–117.

Heckman, J. J. (2010): "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy," *Journal of Economic Literature*, 48(2), 356–398.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *The Review of Economic Studies*, pp. 605–654.

Heckman, J. J., and D. Schmierer (2010): "Tests of hypotheses arising in the correlated random coefficient model," *Economic Modelling*, 27(6), 1355–1367.

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158(2), 177–203.

Heckman, J. J., and S. Urzua (2010): "Comparing IV with structural models: What simple IV can and cannot identify," *Journal of Econometrics*, 156(1), 27–37.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

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

Heckman, J. J., and E. J. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Heckman, J. J., and E. J. Vytlacil (2007): "Econometric evaluation of social programs, part II," *Handbook of Econometrics*, 6, 4875–5143.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

Mogstad, M., and M. Wiswall (2011): "Testing the Quantity-Quality Model of Fertility: Estimation using Unrestricted Family Size Models," *Working Paper*.

Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," *Econometrica*, 56(4), 931–954.

Rosenzweig, M. R., and K. I. Wolpin (1980): "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment," *Econometrica*, 48(1), 227–240.

Ruhm, C. J. (2008): "Maternal Employment and Adolescent Development," *Labour Economics*, 15(5), 958–983.

Vytlacil, E. J. (2002): "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, 70(1), 331–341.

6 Tables

Table 1: Descriptive Statistics

	Mean	Std. Dev.
Outcome:		
Years of schooling	12.3	2.7
Instruments:		
Same sex, 1st and 2nd child	0.501	0.5
Twins at second birth	0.0096	0.097
Endogenous regressor:		
At least three children	0.5021	0.5
Covariates:		
Female	0.47	0.50
Age in 2000	39.5	9.2
Mother's age at first birth	24.0	4.2
Father's age at first birth	26.8	4.5
Mother's years of schooling	10.0	1.4
Father's years of schooling	10.1	2.6

Note: Descriptive statistics are for 514,049 children. All children are first born with at least one sibling. Twins at first birth are excluded from the sample. All children, parents and siblings are aged between 16 and 74 years at some point between 1986 and 2000.

Table 2: Fertility decision model - Average Derivatives

	Average effect (std. err.)
Covariates:	
Age in 2000	0.0163 (0.0011)
Mother's age at first birth	-0.0161 (0.0013)
Father's age at first birth	0.0007 (0.0008)
Mother's years of schooling	0.0030 (0.0016)
Father's years of schooling	-0.0038 (0.0019)
Female	-0.0016 (0.0018)
Instruments:	
Same sex, first and second	0.0567 (0.0012)
Twins at 2nd parity	0.5179 (0.0007)

Note: This table reports the average partial effect (average treatment effect for binary variables) from a logit model for the probability of being in a family with 2 or more siblings rather than 1 sibling. The emodel is specified in the following way: We use a third order polynomial in "Age in 2000", "Mother's age at first birth", "Father's age at first birth birth", "Mother's years of schooling" and "Father's years of schooling"; We include interactions between the first order terms of all covariates; "Same sex, first and second" enters the model without interaction terms; "Twins at 2nd parity" is interacted with all covariates (including higher order terms and interactions) to ensure that the model is consistent with the fact that there are no never takers with twins. Standard errors in parantheses are computed by nonparametric bootstrap with 100 bootstrap replications.

Table 3: OLS and IV estimates

	Z_{-} as instrument	$P(Z)$ as instrument
IV:		
Same-sex instrument	0.174	-0.208
	(0.115)	(0.104)
Twins instrument	0.050	-0.060
	(0.063)	(0.063)
Both instruments	0.076	-0.015
	(0.055)	(0.054)
OLS	-0.052	
	(0.007)	

Note: This table reports OLS and IV estimates of the effect of family size on the educational attainment of first born children. The first column (Z_{-} as instrument) uses the first stage equation (16). The second column (P(Z) as instrument) uses the first stage equation (17). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. The second stage is given by equation (15). We use the same specification for the covariates as reported in Table 2. The first row uses the "Same sex, first and second" instrument, the second row uses the "Twins at 2nd parity" instrument, and the third row uses both instruments. The OLS estimates is reported in the fourth row. Standard errors in parantheses are heteroskedasticity-robust.

Table 4: LATE and linear MTE estimates with same-sex instrument, no covariates

(a) Estimates of linear MTE model and its components

	p=0.473	p=0.531	intercept	slope
Linear MTE model:				
$\mu_1 + K_1(P) = E(Y_1 \mid U_D < p)$	12.086	12.131	11.720	+ 0.773 p
	(0.008)	(0.007)	(0.073)	(0.180)
$\mu_0 + K_0(P) = E(Y_0 \mid U_D > p)$	12.462	12.450	12.564	-0.216 p
	(0.007)	(0.008)	(0.090)	(0.184)
$\mu_1 + k_1(p) = E(Y_1 \mid U_D = p)$	12.453	12.576	11.720	+1.550 p
	(0.082)	(0.103)	(0.073)	(0.360)
$\mu_0 + k_0(p) = E(Y_0 \mid U_D = p)$	12.576	12.551	12.780	-0.432 p
	(0.100)	(0.790)	(0.270)	(0.0368)
$MTE(p) = E(Y_1 - Y_0 \mid U_D = p)$	-0.123	-0.008	-1.006	+1.981 p
	(0.129)	(0.130)	(0.285)	(0.514)

(b) LATE from IV and linear MTE model

Instrumental variables:	
$\left(E(Y \mid Pr(D) = 0.531) - E(Y \mid Pr(D) = 0.473)\right) / (0.531 - 0.473)$	-0.065
	(0.129)
LATE from linear MTE model:	
$\int_{0.471}^{0.531} MTE(p) = MTE((0.531 + 0.471)/2)$	-0.065
	(0.128)

Note: This table displays LATE and linear MTE estimates of family size on the educational attainment of first born children. Panel (a) reports estimates of the linear MTE-model with "Same sex, first and second" as instrument and no covariates. Panel (b) reports estimates of LATE from the IV estimator and the linear MTE model, with "Same sex, first and second" as instrument and no covariates. Standard errors in parantheses are computed by nonparametric bootstrap with 100 bootstrap replications.

Table 5: Tests of constant MTE: Comparing LATEs at different propensity score ranges

LATE over interval - LATE over interval	(0.20,0.25) (0.30,0.35)	(0.30,0.35) (0.40,0.45)	(0.40,0.45) (0.50,0.55)	(0.50,0.55) (0.60,0.65)	(0.60,0.65) (0.70,0.75)
point est.	1.109	1.285	0.053	-0.752	-1.239
std. err.	0.441	0.371	0.294	0.285	0.390
p-value	0.012	0.001	0.857	0.008	0.002
joint p-value	0.000				

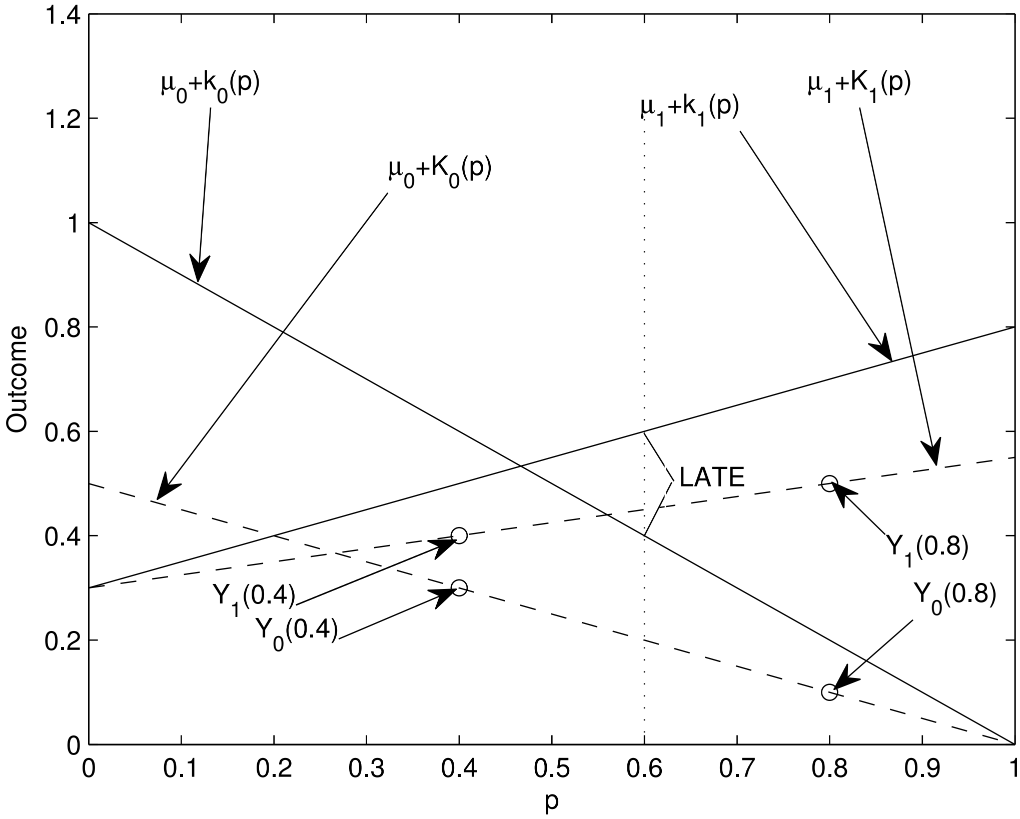
Note: This table reports tests of constant MTE of family size on the educational attainment of first born children. The MTE estimates are from the semiparametric generalized Roy model based on Assumptions 1 and 2, with "Same sex, first and second" as instrument (see Figure 5). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. We use the same specification for the covariates as reported in Table 2. The MTE estimates are based on double residual regression separately for the treated and non-treated, using local quadratic regression with rectangular kernel and bandwidth of 0.055. The LATEs are derived from the MTE estimates by integrating over the indicated intervals. Standard errors are based on nonparametric bootstrap (of both estimation stages) with 100 bootstrap replications.

Table 6: Treatment effect parameters using same-sex instrument

model	ATE	ATT	ATUT
lower bound	0.232	0.313	0.194
	(0.060)	(0.086)	(0.061)
rescaled support	0.423	0.756	0.553
	(0.110)	(0.171)	(0.150)

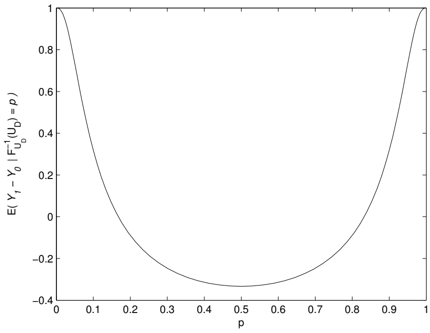
Note: This table reports ATE, ATET, and ATUT of family size on the educational attainment of first born children. *Lower bound:* We use estimates of MTE in the region (0.20,0.75). In the regions (0,0.20) and (0.75,1) the MTE is set equal to 0. *Rescaled support:* We use estimates of MTE in the region (0.20,0.75), and rescale the weights to integrate to one over this region. In both cases, the MTE estimates are from the semiparametric generalized Roy model based on Assumptions 1 and 2, with "Same sex, first and second" as instrument (see Figure 5). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. . We use the same specification for the covariates as reported in Table 2. The MTE estimates are based on double residual regression separately for the treated and non-treated, using local quadratic regression with rectangular kernel and bandwidth of 0.055. Standard errors are based on nonparametric bootstrap (of both estimation stages), with 100 bootstrap replications.

Figure 1: The geometry of the linear MTE model and LATE



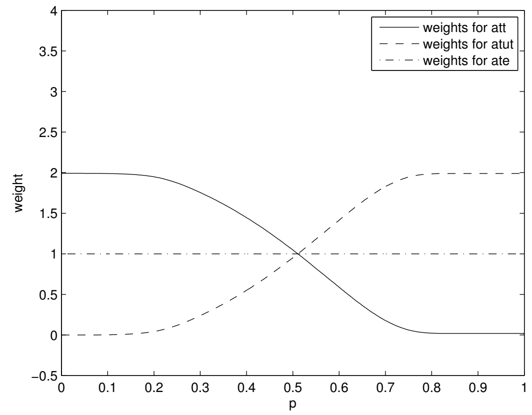
Note: This figure shows the geometry of the linear MTE-models and LATE. We consider a binary instrument with associated propensity scores $p_0 = 0.4$ and $p_1 = 0.8$. The four circles indicate the average outcome for each combination of treatment state and instrument value. The dashed line that goes through the two conditional averages for the treated observations identifies the line $\mu_1 + K_1(p)$. The dashed line that goes through the two conditional averages for the untreated observations identifies the line $\mu_0 + K_0(p)$. The solid line $\mu_1 + k_1(p)$ has twice the slope as the dashed line $\mu_1 + K_1(p)$. The solid line $\mu_0 + k_0(p)$ has twice the slope as the dashed line $\mu_0 + K_0(p)$. Note that $k_0(1) = K_0(1)$ and $k_1(0) = K_1(0)$. We identify MTE from the vertical difference between the solid lines at a given value $U_D = p$, i.e. $MTE(p) = \mu_1 - \mu_0 + k_1(p) - k_0(p)$. The LATE is given by the integrated MTE over the interval (p_0, p_1), which equals the vertical distance between the solid lines at the midpoint of the interval (p_0, p_1) (indicated by the vertical dotted line).

Figure 2: Example of MTE generated from a mixture model

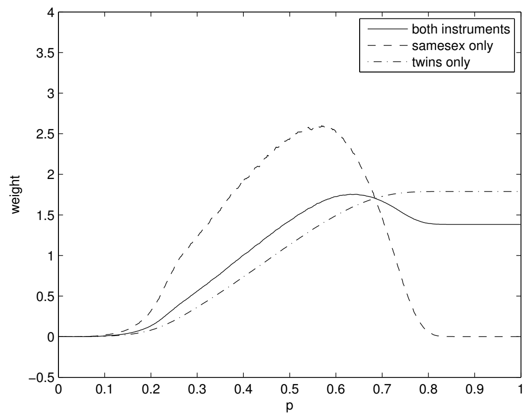


Note: This figure displays the distribution of MTE that is generated from a mixture of two normal selection models. The population consists of two equally sized subgroups: One with constant marginal treatment effects equal to 1; the other with constant marginal treatment effects equal to -1. In the selection equation both groups enter treatment if a random variable exceeds a threshold of 0. The group with negative marginal treatment effects has a random variables that is standard normal, while the group with positive marginal treatment effects has random variables that is normal with mean zero and variance 2. The y-axis measures the value of the MTE, whereas the x-axis represents the unobserved component of parents' net gain from treatment. A high value of p means that treatment is less likely.

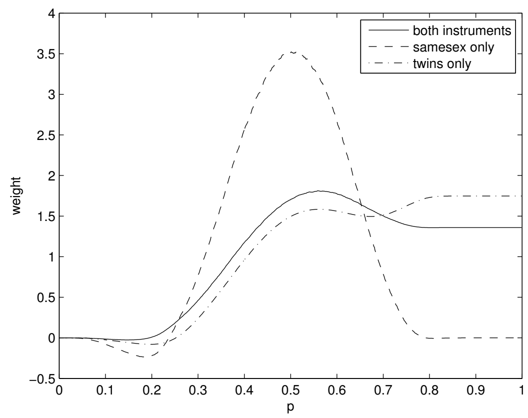
Figure 3: Weight of MTE for treatment effects parameters and instruments



(a) ATT, ATUT, and ATE



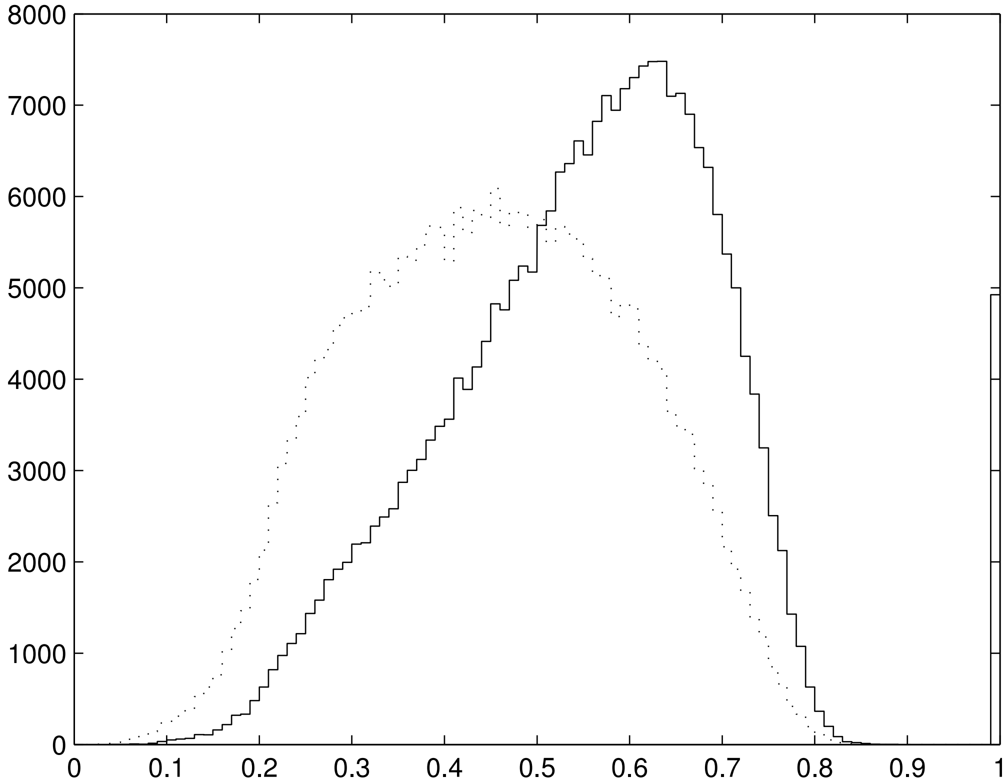
(b) IV with Z_{-} as instrument



(c) IV with P(Z) as instrument

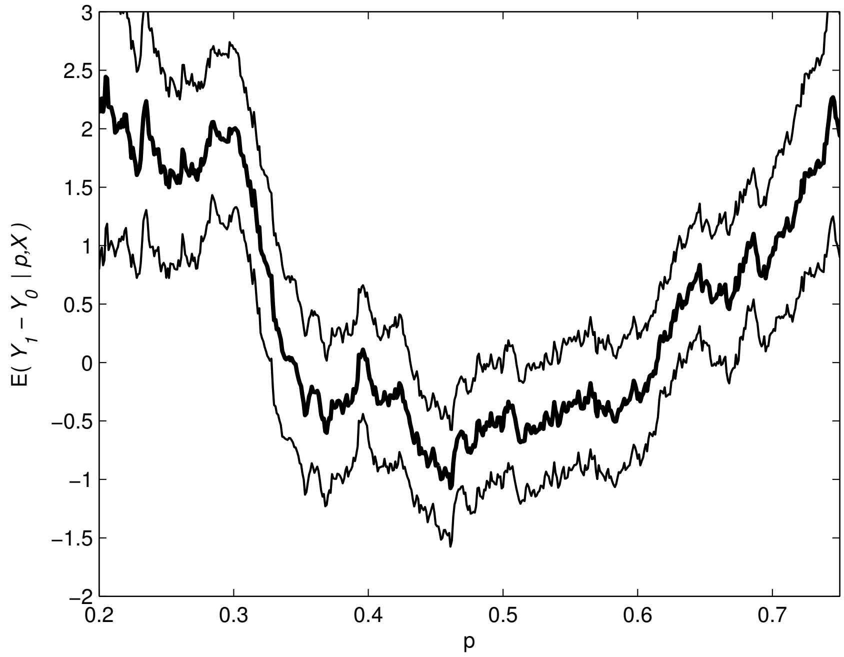
Note: The upper panel graphs MTE weights associated with the average treatment effect on the treated (ATT), the average treatment effect (ATE), and the average treatment effect on the untreated (ATUT). The middle panel (Z_{-} as instrument) and lower panel ($P(Z)$ as instrument) graph MTE weights associated with the IV estimates presented in Table 3. To compute the weights, we use the weight formulas described in the Appendix. The y-axis measures the density of the distribution of weights, whereas the x-axis represents the unobserved component of parents' net gain from having 3 or more children rather than 2 children. A high value of p means that a family is less likely to have 3 or more children.

Figure 4: Histogram of propensity scores with same-sex instrument, for the treated (solid) and the untreated (dotted)



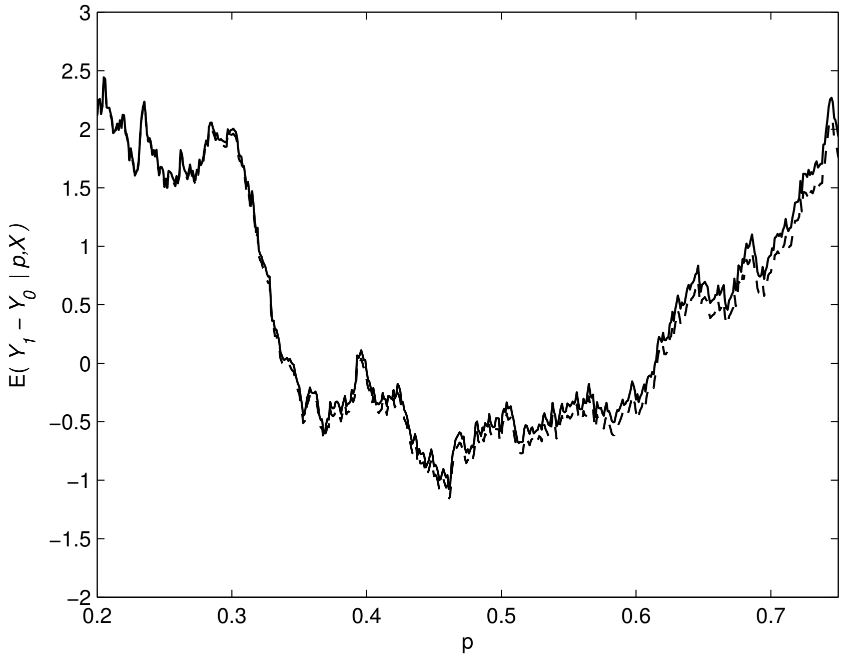
Note: This figure shows the empirical support of $P(Z) \equiv Pr(D = 1 \mid X, Z)$ under Assumptions 1 and 2, with "Same sex, first and second" as instrument. The common support is defined as the intersection of the support of $P(Z)$ given $D = 1$ (solid) and the support of $P(Z)$ given $D = 0$ (dotted). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. We use the same specification for the covariates as reported in Table 2.

Figure 5: MTE estimates with same-sex instrument



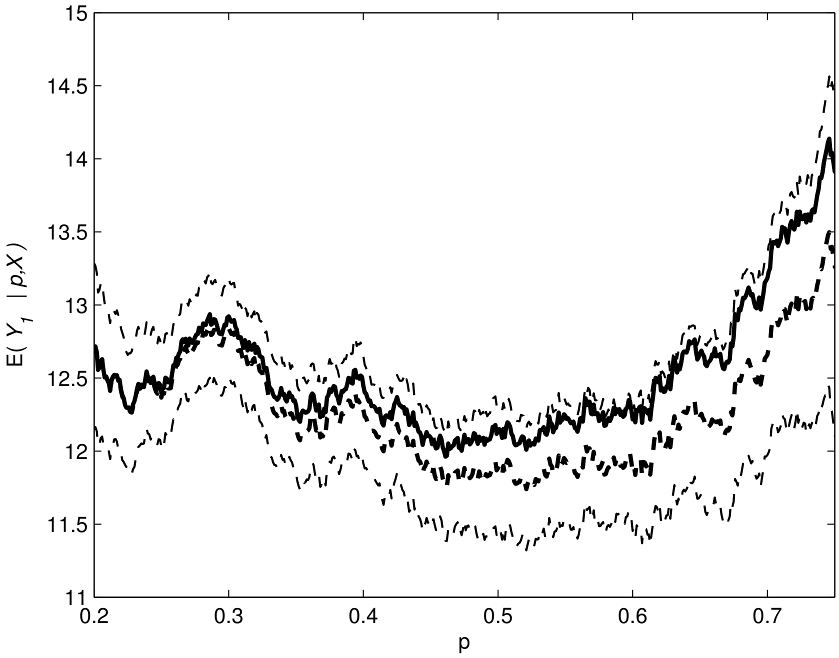
Note: This figure displays the MTE estimates from the semiparametric generalized Roy model based on Assumptions 1 and 2, with "Same sex, first and second" as instrument. We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. We use the same specification for the covariates as reported in Table 2. The MTE estimates are based on double residual regression separately for the treated and non-treated, using local quadratic regression with rectangular kernel and bandwidth of 0.055. The 95 percent confidence interval is computed from a non-parametric bootstrap with 100 bootstrap replications. The y-axis measures the value of the MTE in years of schooling, whereas the x-axis represents the unobserved component of parents' net gain from having 3 or more children rather than 2 children. A high value of p means that a family is less likely to have 3 or more children.

Figure 6: MTE estimates with same-sex instrument only and with both same-sex and twins instruments



Note: This figure displays the MTE estimates from the semiparametric generalized Roy model based on Assumptions 1 and 2. We show estimates with "Same sex, first and second" as the only instrument (solid line) and when both the "Same sex, first and second" instrument and the "Twins at 2nd parity" instrument are included (dashed line). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. We use the same specification for the covariates as reported in Table 2. The MTE estimates are based on double residual regression separately for the treated and non-treated, using local quadratic regression with rectangular kernel and bandwidth of 0.055. The 95 percent confidence interval is computed from a non-parametric bootstrap with 100 bootstrap replications. The y-axis measures the value of the MTE in years of schooling, whereas the x-axis represents the unobserved component of parents' net gain from having 3 or more children rather than 2 children. A high value of p means that a family is less likely to have 3 or more children.

Figure 7: Expected outcome as treated for each instrument



Note: This figure displays estimates of expected years of schooling with 2 or more siblings ($\mu_1(X)+k_1(p)$) from the semiparametric generalized Roy model based on Assumptions 1 and 2. We show estimates with "Same sex, first and second" as the only instrument (solid line) and with "Twins at 2nd parity" as the only instrument (dashed line). We construct P(Z) using the parameter estimates from the logit model with average derivatives reported in Table 2. We use the same specification for the covariates as reported in Table 2. The MTE estimates are based on double residual regression separately for the treated and non-treated, using local quadratic regression with rectangular kernel and bandwidth of 0.055. The 95 percent confidence interval (dotted lines) pertains to the MTE estimates based on the "Twins at 2nd parity" instrument, and is computed from a non-parametric bootstrap with 100 bootstrap replications. The y-axis measures the outcome in years of schooling, whereas the x-axis represents the unobserved component of parents' net gain from having 3 or more children rather than 2 children. A high value of p means that a family is less likely to have 3 or more children.

A Proofs

Proposition 1

Proof. Suppose Assumption 1 holds. Assume that $P(Z)$ takes on N different values, $p_1, \dots, p_N \in (0,1)$. Without loss of generality, we keep the conditioning on X implicit and take $Z = Z_-$.

The expected outcome as a function of the propensity score is given by

$$E(Y|P(Z) = p) = \mu_0 + p(\mu_1 - \mu_0) + K(p). \tag{18}$$

Specify $k(p) = \sum_{k=0}^{L} \alpha_k p^k$. Inserting for $k(p)$ in $K(p)$ gives

$$K(p) = \int_0^p k(u)du = \sum_{k=0}^{L} \frac{\alpha_k}{k+1} p^{k+1},$$

and it follows that $E(Y|P(Z) = p)$ is a polynomial in p of order $L+1$

Let $\bar{Y^c}(p_i)$ denote the conditional average of Y given $P(Z) = p_i$. If $L = N-2$, there is exactly one combination of parameters

$$\theta = (\mu_0, \mu_1 - \mu_0 + \alpha_0, \alpha_1, \dots, \alpha_L)$$

that fits the expectations in equation (18) to the observed conditional averages

$$\{(p_1, \bar{Y^c}(p_1)), \dots, (p_N, \bar{Y^c}(p_N))\}$$

according to the unisolvence theorem. Because

$$E(U_1 - U_0) = \int_0^1 k(u)du = 0$$

implies

$$\alpha_0 = -\sum_{k=1}^{L} \alpha_k/(1+k)$$

then μ_1 is also identified. In contrast, if $L > N-2$, there are several combinations of the parameters θ that fit the expectations in equation (18) to the observed conditional averages. Thus, using LIV the MTEs are identified provided k is specified as a polynomial of order no higher than $N-2$.

The expected outcome as function of propensity scores and treatment status is given by

$$E(Y|P(Z) = p, D = j) = \mu_j + K_j(p),\ j = 0,1 \tag{19}$$

Specify $k_1(p) = \sum_{k=0}^{L} \alpha_k^1 p^k$. Inserting for $k_1(p)$ in $K_1(p)$ gives,

$$K_1(p) = \frac{1}{p} \int_0^p k_1(u)du = \sum_{k=0}^{L} \frac{\alpha_k^1}{k+1} p^k,$$

and it follows that $E(Y|P(Z) = p, D = 1)$ is a polynomial in p of order L.

Let $\bar{Y_1^c}(p_i)$ denote the conditional average of Y given $P(Z) = p_i$ and $D = 1$. If $L = N - 1$, there is exactly one combination of parameters

$$\xi^1 = (\mu_1 + \alpha_0^1, \alpha_1^1, \ldots, \alpha_L^1)$$

that fits the expectations in equation (19) to the observed conditional averages

$$\{(p_1, \bar{Y_1^c}(p_1)), \ldots, (p_N, \bar{Y_1^c}(p_N))\}$$

according to the unisolvence theorem. Because

$$E(U_1) = \int_0^1 k_1(u)du = 0$$

implies

$$\alpha_0^1 = -\sum_{k=1}^{L} \alpha_k^1/(1+k)$$

then μ_1 is also identified. In contrast, if $L > N - 1$, there are several combinations of the parameters ξ^1 that fit the expectations in equation (18) to the observed conditional averages.

The proof for identification of μ_0 and the parameters in $K_0(p)$ follows the above procedure. Thus, using the separate estimation approach the MTEs are identified provided k_1 and k_0 are specified as polynomials of degree no higher than $N - 1$.

■

Proposition 2

Proof. Suppose Assumptions 1 and 2 hold. Assume that X takes on M different values and Z takes on N different values for each X, giving MN values of $P(Z)$, labeled $(p_1, \ldots, p_{MN}) \in \mathcal{P} = (0,1)^{MN}$.

The expected outcome as a function of the propensity score is given by

$$E(Y|P(Z) = p, X = x) = \mu_0(x) + p(\mu_1(x) - \mu_0(x)) + K(p).$$

Specify $k(p) = \sum_{k=0}^{L} \alpha_k p^k$. Inserting for $k(p)$ in $K(p)$ gives

$$K(p) = \int_0^p k(u)du = \sum_{k=0}^{L} \frac{\alpha_k}{k+1} p^{k+1}.$$

Let $\bar{Y^c}(p_i, x)$ denote the conditional average of Y given $P(Z) = p_i$ and $X = x$. There are MN different values of $\bar{Y^c}(p_i, x)$. The vector of parameters

$$\begin{aligned}\theta \ = \ & (\alpha_1/2, \ldots, \alpha_L/(L+1), \\ & \mu_0(1), \mu_1(1) - \mu_0(1) + \alpha_0, \ldots, \mu_0(N), \mu_1(N) - \mu_0(N) + \alpha_0)'\end{aligned}$$

are identified if no more than one solution exist for the equation system

$$E(Y|P(Z) = p_i, X = x) = \bar{Y^c}(p_i, x)$$

for all p_i and X. The equation system is linear in parameters, with $A\theta = \bar{Y^c}$ where

$$A = \begin{pmatrix} p_1^2 & \cdots & p_1^{L+1} & 1 & p_1 & 0 & 0 & \cdots & 0 & 0 \\ p_2^2 & \cdots & p_2^{L+1} & 1 & p_2 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & & & & \vdots & \vdots \\ p_N^2 & \cdots & p_N^{L+1} & 1 & p_N & 0 & 0 & \cdots & 0 & 0 \\ p_{N+1}^2 & \cdots & p_{N+1}^{L+1} & 0 & 0 & 1 & p_{N+1} & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ p_{MN}^2 & \cdots & p_{MN}^{L+1} & 0 & 0 & 0 & 0 & \cdots & 1 & p_{MN} \end{pmatrix}$$

and $\bar{Y^c}$ is an appropriately sorted (column) vector of $\bar{Y^c}(p_i, x)$.

If $L > (N-2)M$, there are several combinations of the parameters θ that solves the equation system. If $L = (N-2)M$, the equation system has a unique solution if and only if the determinant $D(p_1, \ldots, p_{MN}) = |A| \neq 0$. Note first that $D(p_1, \ldots, p_{MN})$ is not 0 for all $(p_1, \ldots, p_{MN}) \in \mathcal{P}$ (this can easily be verified numerically for any choice of $M > 1$ and $N > 2$). Further, D is analytic in $(p_1, \ldots, p_{MN})$. Since D is not zero for all $(p_1, \ldots, p_{MN}) \in \mathcal{P}$ and analytic, it is not zero on any open subset of $\mathcal{P}$. Hence, the equation system has a unique solution a.e. in $\mathcal{P}$. Because

$$E(U_1 - U_0) = \int_0^1 k(u)du = 0$$

implies

$$\alpha_0 = -\sum_{k=1}^{L} \alpha_k/(1+k)$$

then $\mu_1(1), ..., \mu_1(N)$ are also identified.

The expected outcome as a function of the propensity score and treatment status is given by

$$E(Y|P(Z) = p, X = x, D = j) = \mu_j(x) + K_j(p),\ j = 0, 1.$$

Inserting for $k_1(p)$ in $K_1(p)$ gives

$$K_1(p) = \frac{1}{p}\int_0^p k_1(u)du = \sum_{k=0}^{L} \frac{\alpha_k^1}{k+1} p^k.$$

Let $\bar{Y_1^c}(p_i, x)$ denote the conditional average of Y given $P(Z) = p_i$, $D = 1$, and $X = x$. There are MN different values of $\bar{Y_1^c}(p_i, x)$. The vector of parameters

$$\xi^1 \quad = \quad (\alpha_1^1/2, \ldots, \alpha_L^1/(L+1), \mu_1(1) + \alpha_0^1, .., \mu_1(N) + \alpha_0^1)'$$

are identified if no more than one solution exist for the equation system

$$E(Y|P(Z) = p_i, D = 1, X = x) = \bar{Y_1^c}(p_i, x)$$

for all p_i and X. The equation system is linear in parameters, with $A_1\xi^1 = \bar{Y_1^c}$ where

$$A_1 = \begin{pmatrix} p_1^1 & \cdots & p_1^L & 1 & 0 & \cdots & 0 \\ p_2^1 & \cdots & p_2^L & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ p_N^1 & \cdots & p_N^L & 1 & 0 & \cdots & 0 \\ p_{N+1}^1 & \cdots & p_{N+1}^L & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ p_{MN}^1 & \cdots & p_{MN}^L & 0 & 0 & \cdots & 1 \end{pmatrix}$$

and $\bar{Y_1^c}$ is an appropriately sorted (column) vector of $\bar{Y_1^c}(p_i, x)$.

If $L > (N-1)M$, there are several combinations of the parameters ξ^1 that solves the equation system. If $L = (N-1)M$, the equation system has a unique solution if and only if the determinant $D_1(p_1, \ldots, p_{MN}) = |A_1| \neq 0$. Note first that $D_1(p_1, \ldots, p_{MN})$ is not 0 for all $(p_1, \ldots, p_{MN}) \in \mathcal{P}$ (this can easily be verified numerically for any choice of $M > 1$ and $N > 1$). Further, D_1 is analytic in $(p_1, \ldots, p_{MN})$. Since D_1 is not zero for all $(p_1, \ldots, p_{MN}) \in \mathcal{P}$ and analytic, it is not zero on any open subset of $\mathcal{P}$. Hence, the equation system has a unique solution a.e. in $\mathcal{P}$. Because

$$E(U_1) = \int_0^1 k(u)du = 0$$

implies

$$\alpha_0^1 = -\sum_{k=1}^{L} \alpha_k^1/(1+k)$$

then $\mu_1(1), ..., \mu_1(N)$ are also identified.

The proof for identification of the parameters $\mu_0(x)$ and those in $K_0(p)$ follows the above procedure.

■

B Supplementary tables

Table B-1: LATE and Linear MTE slope coefficients for 64 subgroups

Group						Size	p_0	p_1	LATE	(se)	SLOPE	(se)
X_1	X_2	X_3	X_4	X_5	X_6							
0	0	0	0	0	0	12136	0.519	0.574	-0.0856	(0.618)	-1.77	(2.48)
0	0	0	0	0	1	12855	0.687	0.717	0.226	(1.38)	10.4	(6.03)
0	0	0	0	1	0	1453	0.43	0.458	-5.04	(6.03)	19.4	(15.9)
0	0	0	0	1	1	4396	0.548	0.605	0.145	(1.29)	0.765	(5.23)
0	0	0	1	0	0	5030	0.551	0.596	-0.896	(1.24)	1.76	(4.96)
0	0	0	1	0	1	13150	0.658	0.693	-2.56	(1.37)	-2.94	(5.3)
0	0	0	1	1	0	4240	0.367	0.391	1.64	(3.01)	9.73	(11.6)
0	0	0	1	1	1	23920	0.439	0.498	-1.34*	(0.591)	-1.14	(2.32)
0	0	1	0	0	0	9070	0.432	0.496	0.584	(0.663)	-4.57	(2.62)
0	0	1	0	0	1	3349	0.651	0.731	-0.0908	(1.03)	2.38	(4.53)
0	0	1	0	1	0	1643	0.396	0.444	-3.35	(2.75)	-6.47	(8.8)
0	0	1	0	1	1	1348	0.576	0.628	0.654	(2.67)	-17.5	(10.8)
0	0	1	1	0	0	4212	0.473	0.518	1.5	(1.6)	5.51	(6.01)
0	0	1	1	0	1	4114	0.632	0.691	0.139	(1.42)	-4.14	(6.07)
0	0	1	1	1	0	5242	0.344	0.387	-0.778	(1.52)	-1.28	(6.25)
0	0	1	1	1	1	12853	0.431	0.493	0.885	(0.843)	4.13	(3.33)
0	1	0	0	0	0	14081	0.461	0.516	-0.415	(0.637)	6.58*	(2.55)
0	1	0	0	0	1	7835	0.662	0.699	-0.619	(1.51)	4.3	(6.5)
0	1	0	0	1	0	2286	0.356	0.416	0.236	(1.55)	2.42	(6.45)
0	1	0	0	1	1	2756	0.535	0.592	-1.77	(1.85)	-10.5	(7.09)
0	1	0	1	0	0	2553	0.469	0.539	-0.631	(1.26)	0.532	(5.02)
0	1	0	1	0	1	3136	0.66	0.693	-0.162	(2.75)	10.4	(11.6)
0	1	0	1	1	0	4131	0.295	0.363	0.241	(1.05)	-3.86	(4.47)
0	1	0	1	1	1	8416	0.417	0.465	4.00*	(1.54)	8.72	(4.88)
0	1	1	0	0	0	29784	0.395	0.471	0.779	(0.362)	-1.38	(1.45)
0	1	1	0	0	1	7093	0.655	0.715	2.29*	(1.13)	0.727	(4.5)
0	1	1	0	1	0	11955	0.379	0.457	-0.661	(0.585)	-2.16	(2.36)
0	1	1	0	1	1	4436	0.581	0.648	-0.205	(1.29)	13.1*	(5.3)
0	1	1	1	0	0	6510	0.443	0.51	-0.308	(0.899)	2.02	(3.6)
0	1	1	1	0	1	3880	0.65	0.705	0.637	(1.68)	-2.19	(7.14)
0	1	1	1	1	0	22664	0.318	0.385	0.889	(0.495)	-0.813	(2.08)
0	1	1	1	1	1	20248	0.453	0.514	0.224	(0.664)	-3.42	(2.66)

Note: X_1 - female, X_2 - Father's years of schooling > 9, X_3 - Mothers years of schooling > 9, X_4 - Father's age at first birth > 26, X_5 - Mother's age at first birth > 23, X_6 - Age in 2000 > 39. Chi-square test of H_0 : "all slope coefficients are equal to zero" has 64 degrees of freedom and gives test statistic 82.02 (p-value: 0.0642).

Table B-1: continued

Group						Size	p_0	p_1	LATE	(se)	SLOPE	(se)
X_1	X_2	X_3	X_4	X_5	X_6							
1	0	0	0	0	0	11787	0.518	0.572	0.394	(0.699)	-0.813	(2.78)
1	0	0	0	0	1	10862	0.687	0.731	1.51	(0.98)	-0.0123	(4.1)
1	0	0	0	1	0	1344	0.429	0.472	1.66	(3.01)	-1.29	(11.6)
1	0	0	0	1	1	3356	0.552	0.595	-1.74	(1.89)	1.57	(7.28)
1	0	0	1	0	0	4823	0.541	0.623	-0.432	(0.752)	-0.457	(3.06)
1	0	0	1	0	1	10416	0.668	0.717	0.907	(0.911)	2.63	(3.9)
1	0	0	1	1	0	4135	0.358	0.39	-2.5	(2.47)	-2.42	(9.24)
1	0	0	1	1	1	18253	0.446	0.473	2.23	(1.49)	7.12	(5.37)
1	0	1	0	0	0	8727	0.434	0.492	-0.137	(0.813)	2.7	(3.27)
1	0	1	0	0	1	3016	0.687	0.721	0.32	(2.59)	-3.28	(11.3)
1	0	1	0	1	0	1522	0.381	0.446	3.78	(2.34)	-0.629	(7.35)
1	0	1	0	1	1	1195	0.589	0.594	-44	(248)	241	(120)
1	0	1	1	0	0	4081	0.488	0.552	-0.306	(1.15)	5.1	(4.61)
1	0	1	1	0	1	3501	0.659	0.682	-1.57	(3.95)	-23	(16.3)
1	0	1	1	1	0	5089	0.338	0.399	-1.08	(1.13)	-1.43	(4.64)
1	0	1	1	1	1	10545	0.431	0.492	0.0253	(0.867)	-4.93	(3.48)
1	1	0	0	0	0	13453	0.45	0.517	-0.327	(0.573)	3.86	(2.3)
1	1	0	0	0	1	6959	0.656	0.725	-0.965	(0.835)	2.29	(3.58)
1	1	0	0	1	0	2212	0.363	0.427	0.497	(1.54)	10.8	(6.32)
1	1	0	0	1	1	2404	0.557	0.603	-2.72	(2.56)	-7.7	(9.15)
1	1	0	1	0	0	2552	0.475	0.551	-1.2	(1.22)	-1.7	(4.77)
1	1	0	1	0	1	2591	0.625	0.708	0.691	(1.18)	6.8	(4.96)
1	1	0	1	1	0	4040	0.32	0.341	5.23	(5.25)	35.4*	(15.1)
1	1	0	1	1	1	7003	0.407	0.464	-0.848	(1.1)	-5.35	(4.4)
1	1	1	0	0	0	28294	0.403	0.47	0.042	(0.42)	1.97	(1.7)
1	1	1	0	0	1	6445	0.65	0.721	0.781	(0.937)	7.91	(4.02)
1	1	1	0	1	0	11341	0.373	0.444	0.442	(0.62)	2.67	(2.53)
1	1	1	0	1	1	3976	0.612	0.674	1.79	(1.43)	5.1	(5.78)
1	1	1	1	0	0	6300	0.448	0.514	-1.65	(0.97)	-4.94	(3.68)
1	1	1	1	0	1	3403	0.654	0.712	2.03	(1.65)	13.8*	(6.8)
1	1	1	1	1	0	21990	0.317	0.382	0.146	(0.48)	-1.62	(2.02)
1	1	1	1	1	1	17614	0.456	0.514	0.749	(0.688)	0.463	(2.74)

Note: X_1 - female, X_2 - Father's years of schooling > 9, X_3 - Mothers years of schooling > 9, X_4 - Father's age at first birth > 26, X_5 - Mother's age at first birth > 23, X_6 - Age in 2000 > 39. Chi-square test of H_0 : "all slope coefficients are equal to zero" has 64 degrees of freedom and gives test statistic 82.02 (p-value: 0.0642).

B

Returadresse:
Statistisk sentralbyrå
NO-2225 Kongsvinger

Statistics Norway

Oslo:
PO Box 8131 Dept
NO-0033 Oslo
Telephone: + 47 21 09 00 00
Telefax: + 47 21 09 00 40

Kongsvinger:
NO-2225 Kongsvinger
Telephone: + 47 62 88 50 00
Telefax: + 47 62 88 50 30

E-mail: ssb@ssb.no
Internet: www.ssb.no

ISSN 0809-733X

Statistisk sentralbyrå
Statistics Norway