Beyond LATE with a Discrete Instrument

Author(s): Christian N. Brinch, Magne Mogstad and Matthew Wiswall

Source: *Journal of Political Economy*, August 2017, Vol. 125, No. 4 (August 2017), pp. 985-1039

Published by: The University of Chicago Press

Stable URL: https://www.jstor.org/stable/10.2307/26550434

Beyond LATE with a Discrete Instrument

Christian N. Brinch

BI Norwegian Business School

Magne Mogstad

University of Chicago, Statistics Norway, and National Bureau of Economic Research

Matthew Wiswall

Arizona State University, University of Wisconsin–Madison, and National Bureau of Economic Research

We show how a discrete instrument can be used to identify the marginal treatment effects under a functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment. Guided by this identification result, we perform a marginal treatment effect analysis of the interaction between the quantity and quality of children. Our 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.

I. Introduction

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

We thank three anonymous referees, the editor, and seminar participants at several universities and conferences for valuable feedback and suggestions. Supplementary material online describes how to apply for the data and contains replication files for the analyses.

Electronically published June 30, 2017
[*Journal of Political Economy*, 2017, vol. 125, no. 4]

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

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 the unobserved gain from treatment (β correlated with D), the LATE is informative only about the average causal effect of an 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. This raises concerns about the external validity and policy relevance of the LATE, unless the instrument-induced effect of treatment is the parameter of interest.

To move beyond the LATE, Heckman and Vytlacil (1999, 2005, 2007) generalize the marginal treatment effect (MTE) introduced by Björklund and Moffitt (1987). The MTE has several useful features: it plays the role of a functional that is invariant to the choice of instrument; 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 all conventional treatment parameters can be expressed as different weighted averages of the MTEs, such as the average treatment effect (ATE) and the average treatment effect on the treated (ATT). 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).

While the MTE has several useful features, full nonparametric identification is challenging because it requires instruments that generate continuous support on the probability of treatment $P(Z)$ from zero to one for each value of X. In practice, instruments are often discrete, and many are binary. In such situations, there are two alternative approaches to move beyond the LATE. One alternative is to abandon the goal of point identification and instead construct bounds on conventional treatment parameters. Unfortunately, bounds that do not involve additional assumptions tend to be very wide and therefore rarely informative.[1] Another alternative is to adopt some form of parametric or functional structure. For example, one could assume that treatment effects are the same for

[1] As discussed in detail later, Manski's (1989) no-assumption bounds on the ATE and the ATT are too wide to be informative in our context. Indeed, the bounds remain nearly uninformative even if we impose the IV assumptions.

everyone with the same observed characteristics or that treatment effects vary but individuals do not sort on the unobserved gain from treatment (see, e.g., Angrist and Fernandez-Val 2013). However, the constant effect assumption has been tested and rejected in a wide range of settings, and a number of studies find that individuals self-select on the basis of unobserved gain from treatment (see, e.g., Carneiro, Heckman, and Vytlacil 2011; Kirkebøen, Leuven, and Mogstad 2016).

In this paper, we show how a discrete instrument can be used to identify the MTE under functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment. We begin by showing that an MTE model with at most N parameters 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 approximation of a more general MTE model.[2]

In some applications with discrete instruments, however, one may be reluctant to impose strong parametric restrictions (e.g., linearity) on the MTE model. In such cases, an auxiliary assumption is required. We show that with a binary instrument and M different values of the covariates X, an MTE model with no more than $M + 1$ parameters 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 $E(Y|X = x; P(Z) = p)$ 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 model. The alternative approach, however, identifies the MTE from separately estimating $E(Y|X = x; P(Z) = p, D = 1)$ and $E(Y|X = x; P(Z) = p, D = 0)$. With a binary instrument, the advantage of the alternative estimation approach is that

[2] Note that our test requires only a single binary instrument. In contrast, the approaches to test the external validity of LATE proposed by Heckman and Schmierer (2010), Heckman, Schmierer, and Urzua (2010), and Angrist and Fernandez-Val (2013) 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 single binary instrument.

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).[3] The additional information allows us to use a binary instrument to (i) estimate a linear (approximation of the) MTE model under the standard IV assumptions, (ii) test the external validity of LATE, and (iii) estimate a flexible MTE model under the additional assumption of additive separability between observed and unobserved heterogeneity in treatment effects.[4]

Guided by these identification results, we empirically assess the interaction between 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 effects of family size on child outcomes. Much of the early literature that tested the QQ model found that larger families reduced observable measures of 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 has little effect on children's outcomes.[5]

Although these studies represent a significant step forward, a concern is 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 the separate estimation approach to Norwegian administrative data. We begin by using the same-sex instrument to estimate a linear MTE model. Our test suggests 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

[3] 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.

[4] 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 zero to one for each value of X. With such instruments, Heckman and Vytlacil (2007) show that the alternative estimation approach can nonparametrically identify MTE over the full unit interval, while Carneiro and Lee (2009) use the approach to estimate the distribution of potential outcomes. None of these studies consider situations with discrete instruments.

[5] Black, Devereux, and Salvanes (2005) conclude that "there is little if any family size effect on child education" (697). Using data from the United States and Israel, Caceres-Delpiano (2006) and Angrist, Lavy, and Schlosser (2010) come to a similar conclusion. However, Mogstad and Wiswall (2016) reexamine the analysis by Black et al. (2005) and find a significant but nonlinear relationship between family size and child outcomes: While a second sibling increases the educational attainment of firstborn children, additional children have a negative effect.

and unobserved heterogeneity in treatment effects and estimate a flexible specification of the MTE model. We then find that the effects of family size vary in magnitude and even sign (i.e., β is heterogeneous) and that families act as if they possess some knowledge of their idiosyncratic return in the fertility decision (i.e., β is correlated with D). We show that these results are robust to a number of specification checks, and we use the twins instrument to validate the MTE estimates based on the same-sex instrument. Finally, we compare the MTE weights associated with the IV estimates to the MTE weights associated with the ATE and the ATT, and we find that the latter treatment parameters assign much more weight to the positive part of the MTE distribution. This explains why the ATE and the ATT of family size are sizable and positive, while the LATEs based on twin births or same-sex sibship are smaller and sometimes negative.

The remainder of the paper is organized as follows. Section II 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 MTEs with a continuous instrument. Section III shows how to identify and estimate MTEs with discrete instruments. Sections IV and V present our empirical analysis of the effects of family size on child outcomes. Section VI presents conclusions.

II. Framework and Estimation Procedures

A. *The Generalized Roy Model*

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 ($D = 0$).[6] 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

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

where $\mu_1(\cdot)$ and $\mu_0(\cdot)$ are unspecified functions, X is a random vector of covariates, while 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 do not assume that X and (U_0, U_1) are independent.

[6] For simplicity, we consider only a binary treatment variable. See Heckman, Urzua, and Vytlacil (2006), Heckman and Vytlacil (2007), Heckman and Urzua (2010), and Kirkebøen et al. (2016) for discussions of IV and MTE estimation with multiple treatment variables.

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 $\mu_D(\cdot)$ is an unspecified function, U_D is a continuous random variable with a strictly increasing distribution function, and $Z = (X, Z_-)$ is a vector in which Z_- represents the excluded instrument(s), that is, a variable that enters the selection equation (3) but is excluded from the potential outcome equations (2). An individual selects the treated state if the net benefit of treatment is positive: $D = 1\{I_D > 0\}$. The marginal distribution of U_D can be normalized to a uniform distribution on the unit interval. The function $\mu_D(Z)$ is then interpretable as a propensity score: We therefore write $P(Z) \equiv \Pr(D = 1|Z) = \mu_D(Z)$ 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 as follows.

Assumption 1. (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 parametric joint distribution of the random variables (U_0, U_1, U_D) (see, e.g., Björklund and Moffitt 1987). By contrast, we will not make assumptions 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 the MTE, we use the following notation for the conditional expectations of U_1 and U_0:

$$k_j(p, x) = E(U_j|X = x, U_D = p), \quad j = 0, 1,$$

and

$$k(p, x) = k_1(p, x) - k_0(p, x) = E(U_1 - U_0|X = x, U_D = p). \tag{4}$$

The MTE can then be expressed as follows.

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

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

Note that conditioning on $U_D = p$ is equivalent to conditioning on the intersection of $P(Z) = p$ and $I_D = 0$. As a result, the MTE can be interpreted as the average effect of treatment for persons on a margin of indifference between participation in treatment and nonparticipation.

In the generalized Roy model, a LATE can be defined as an integral over the MTEs (Heckman and Vytlacil 1999, 2005, 2007). In particular, with a binary instrument ($Z_- \in 0, 1$) that shifts the propensity score from $\Pr(D = 1|X = x, Z_- = 0) = p_0(x)$ to $\Pr(D = 1|X = x, Z_- = 1) = p_1(x)$, the LATE can be written as

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

B. Estimation Procedures

There are two estimation procedures for MTE. The separate estimation approach identifies the components $\mu_j(x) + k_j(p, x)$ separately for the treated and the untreated, whereas LIV works directly with the differences in these components across the two groups.

We first consider identification of MTE with LIV (see, e.g., Heckman and Vytlacil 1999). This approach works with the population mean outcome given X and $P(Z)$. Conditioning on the propensity score and inserting the model for potential outcomes (2), we obtain

$$\begin{aligned} E(Y|P(Z) = p, X = x) = (1 - p)[\mu_0(x) + E(U_0|U_D > p, X = x)] \\ + p[\mu_1(x) + E(U_1|U_D \leq p, X = x)]. \end{aligned} \tag{6}$$

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

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

giving

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

Inserting (7) into (6) gives

$$E(Y|P(Z) = p, X = x) = \mu_0(x) + p[\mu_1(x) - \mu_0(x)] + K(p, x), \tag{8}$$

where

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

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 of $P(Z)$ conditional on X.

As an alternative to LIV, Heckman and Vytlacil (2007) discuss a separate estimation approach to identify the MTE. For each treatment state, this approach works with the population mean outcome given X and $P(Z)$. From (2), we obtain

$$E(Y_j|P(Z) = p, X = x, D = j) = \mu_j(x) + K_j(p, x) \tag{9}$$

for $j = 0, 1$, where

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

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

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

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

With an instrument that generates continuous support of $P(Z)$ from zero to one for each value of X, both LIV and the separate estimation approach nonparametrically identify the MTE over the full unit interval (Heckman and Vytlacil 2007). In practice, however, instruments rarely provide such support. Empirically, it is therefore difficult to apply these procedures without imposing an additional assumption. For example, the empirical analyses of Carneiro and Lee (2009), Carneiro et al. (2011), and Maestas, Mullen, and Strand (2013) assume independence between (U_0, U_1, U_D) and (X, Z). Under this assumption, the MTE is additively separable in X and U_D and therefore identified from the marginal support of $P(Z)$ as opposed to the support of $P(Z)$ given X. An alternative (or ad-

ditional) assumption that is often made in applied research is to impose restrictions on the functional form of k in the MTE model (see, e.g., Moffitt 2008; French and Song 2014). For example, Moffitt (2008) follows Heckman and Robb (1985) in specifying k as a polynomial (linear or quadratic) function of p. Similarly, one could specify the functional form of k_0 and k_1 in the separate estimation approach. A functional form assumption aids identification by allowing interpolation between different values of p in the data or extrapolation beyond the support of $P(Z)$ given X. In the next section, we explore how such additional assumptions provide identification of the MTE when the instrument is discrete.

III. Identifying MTE with a Discrete Instrument

In this section, we show how a discrete instrument can be used to identify the MTE under a functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment.

A. *Identification in a Nonseparable Model*

Throughout subsections III.A and III.B, we invoke assumption 1. 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 can identify a linear MTE model with a single binary instrument.

EXAMPLE 1. The following equations specify a linear MTE model:

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

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.[7]

From these expressions, we derive

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

[7] This specification of the MTE model is consistent with that in Olsen (1980), which proposes a selection correction model based on the linear probability model, assuming that U_D is uniform and that $E(U_j|U_D)$ is linear in U_D.

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

and

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

In this case, the MTE is linear in p and is given by

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

From the expressions above, we get

$$E(Y|P(Z) = p, D = 0) = \mu_0 + \frac{1}{2}\alpha_0 p, \tag{10}$$

$$E(Y|P(Z) = p, D = 1) = \mu_1 + \frac{1}{2}\alpha_1(p - 1), \tag{11}$$

and

$$E(Y|P(Z) = p) = \mu_0 + p(\mu_1 - \mu_0) + \frac{1}{2}p(1 - p)(\alpha_1 - \alpha_0). \tag{12}$$

Suppose that $Z \in \{0, 1\}$ such that $P(1) = p_1$ and $P(0) = p_0$, with $p_1 \in (0, 1)$ and $p_0 \in (0, 1)$. Assume that Z is relevant, so that $p_1 \neq p_0$.

Recall that LIV is based on the integrated MTE in equation (12). Although the MTE model is linear in p, equation (12) is quadratic in p. With a binary instrument, the empirical analogue of $E(Y|P(Z) = p)$ is observed for only two different values of p. Thus, LIV cannot identify the linear MTE model with a binary instrument.

The separate estimation approach is based on equations (10) and (11). Both equations are linear in p. With a binary instrument, the empirical analogues of $E(Y|P(Z) = p, D = 1)$ and $E(Y|P(Z) = p, D = 0)$ are observed for two different values of p. As a result, the separate estimation approach identifies μ_0, μ_1, α_0, and α_1 and thus recovers the MTE with a binary instrument. □

1. Geometry of the Linear MTE Model 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 varies with different quantiles of the unobserved component of selection into treatment.

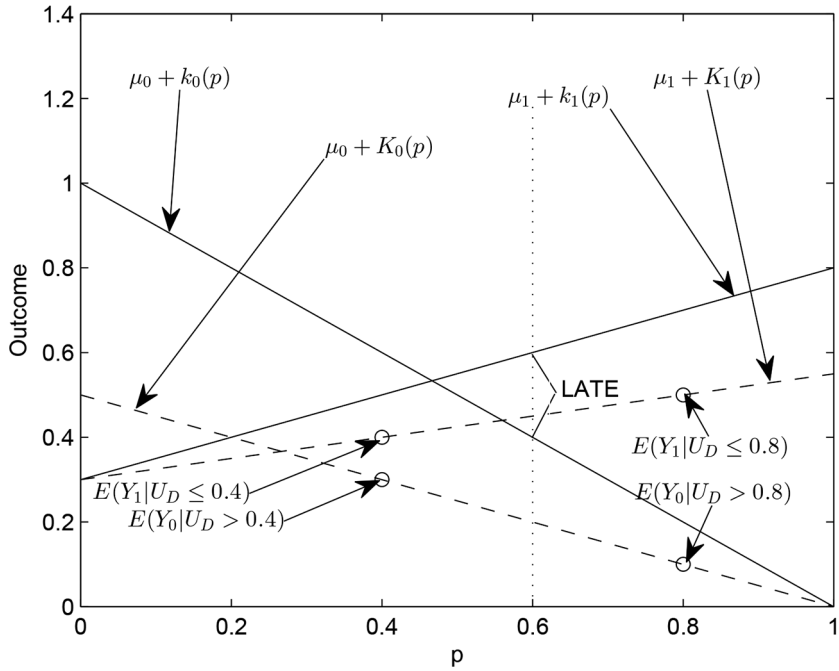

FIG. 1.—This figure shows the geometry of the linear MTE model and LATE. We consider a binary instrument with associated propensity scores $p_0 = 0.4$ and $p_1 = 0.8$. The four circles indicate the expected outcome for each combination of treatment state and instrument value. The dashed line that goes through the two conditional expectations for the treated observations identifies the line $\mu_1 + K_1(p)$. The dashed line that goes through the two conditional expectations 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 $U_D = p$, that is, $\text{MTE}(p) = \mu_1 - \mu_0 + k_1(p) - k_0(p)$. 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. The parameterization values used are $\mu_0 + K_0(p) = 0.5 - 0.5p$ and $\mu_1 + K_1(p) = 0.3 + 0.25p$.

In this example, we consider a binary instrument with associated 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. From these averages we identify the four corresponding conditional expectations, indicated by circles in figure 1. The dashed line that goes through the two conditional expectations for the treated observations identifies $\mu_1 + K_1(p)$. By comparison, the dashed line that goes through the two conditional expectations for the untreated observations identifies $\mu_0 + K_0(p)$. 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 $U_D = p$, that is, $\text{MTE}(p) = \mu_1 - \mu_0 + k_1(p) - k_0(p)$. In this example, the MTE is negative for $U_D < 0.5$ and positive for $U_D > 0.5$. By way of comparison, constant MTE would imply that the solid lines were parallel. The LATE produced by Z_- 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 this interval. 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 nonconstant, different instruments would generally identify different LATEs.

2. Identifying MTE with a Discrete Instrument

Proposition 1 states the general identification result for a discrete instrument under conditional independence.

PROPOSITION 1. Suppose that assumption 1 holds. Assume that $P(Z)$ takes N values, $p_1, \ldots, p_N \in (0, 1)$. Assume further that k, k_1, and k_0 are specified as parametric functions, linear in parameters, with L parameters.

i. Using $E(Y|P(Z) = p, X = x)$, the MTEs can be identified only if $L \leq N - 1$.
ii. Using $E(Y_1|P(Z) = p, X = x, D = 1)$ and $E(Y_0|P(Z) = p, X = x, D = 0)$, the MTEs can be identified only if $L \leq N$.

The proof is given in Appendix A.

The proposition shows what LIV and the separate estimation approach may identify under conditional independence. If we impose restrictions on k_1 and k_0, the separate estimation approach allows identification of richer specifications of the MTE when the instrument is discrete. For instance, example 1 showed that when the instrument is binary, the separate estimation approach can (and LIV cannot) identify an MTE model in which k_1 and k_0 are specified as linear functions of p. Alternatively, a binary instrument could be used in the separate estimation approach to identify an MTE model in which k_0 and k_1 are linear functions of $\Phi^{-1}(p)$, where Φ is the distribution function of the standard normal distribution, as in the normal selection model (see, e.g., Björklund and Moffitt 1987). The parametric specification of k_0 and k_1 can be relaxed with two (or more) instruments or one instrument that takes on multiple values. For example, with an instrument that takes three unique values, the separate estimation approach can identify an MTE model in which k_0 and k_1 are specified as quadratic functions of p or $\Phi^{-1}(p)$.[8]

[8] Proposition 1 shows the maximum number of parameters one can identify under conditional independence. The sufficient conditions for identification are specific to the mod-

B. *Extrapolating and Testing the External Validity of LATE*

Recent work has proposed several tests of the external validity of LATE (see, e.g., Heckman and Schmierer 2010; Heckman et al. 2010; Angrist and Fernandez-Val 2013). All these tests require either two (or more) instruments or one instrument that takes on multiple values. We now propose a complementary test that even with a binary instrument has power against the alternative hypothesis of a nonconstant MTE.

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

$$\begin{aligned} \text{LATE} = & \frac{p_1[\mu_1 + K_1(p_1)] + (1 - p_1)[\mu_0 + K_0(p_1)]}{p_1 - p_0} \\ & - \frac{p_0[\mu_1 + K_1(p_0)] + (1 - p_0)[\mu_0 + K_0(p_0)]}{p_1 - p_0} \end{aligned} \tag{13}$$

because

$$\int_{p_0}^{p_1} k_1(u)du = \int_0^{p_1} k_1(u)du - \int_0^{p_0} k_1(u)du = p_1K_1(p_1) - p_0K_1(p_0)$$

and

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

Equation (13) 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 four 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 linear MTE model can be expressed as

$$\hat{\gamma}^{\text{LATE}} = \frac{[\hat{p}_1\bar{Y}_1(\hat{p}_1) + (1 - \hat{p}_1)\bar{Y}_0(\hat{p}_1)] - [\hat{p}_0\bar{Y}_1(\hat{p}_0) + (1 - \hat{p}_0)\bar{Y}_0(\hat{p}_0)]}{\hat{p}_1 - \hat{p}_0},$$

where $\hat{p}_z$ is the empirical analogue of $P(D = 1|Z = z)$ and $\bar{Y}_j(\hat{p}_z)$ is the empirical analogue of $E(Y|P(Z) = p_z, D = j)$, for $z = 0, 1$ and $j = 0, 1$.

---

eling framework for k_j. If k_j is specified as a global polynomial of p, then N unique values of $P(Z)$ are sufficient to identify an MTE model with a polynomial of order $N - 1$ (e.g., see example 1). By comparison, if k_j is specified as a piecewise linear function of p, it is necessary to have unique values of $P(Z)$ at every line segment to identify the MTE over the full unit interval. We refer to Brinch, Mogstad, and Wiswall (2012) for a discussion of sufficient conditions for identification.

It then follows straightforwardly that $\hat{\gamma}^{\text{LATE}}$ is equal to the standard IV estimator:

$$\hat{\gamma}^{\text{IV}} = \frac{\bar{Y}(\hat{p}_1) - \bar{Y}(\hat{p}_0)}{\hat{p}_1 - \hat{p}_0}.$$

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 model is nonzero so that the MTEs are nonconstant, we reject the external validity of the LATE. In such situations, we may conclude that the LATE estimate is informative only if the instrument-induced effect of treatment is the parameter of interest.

The test for external validity of LATE is simple to implement and does not require estimation of the linear MTE model. In 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 nonconstant MTE is equivalent to testing the null hypothesis

$$\Delta_1 = \Delta_0 \tag{14}$$

versus a two-sided alternative, where

$$\Delta_j = E(Y|D = j, Z = 1) - E(Y|D = j, Z = 0) \quad \text{for } j = \{0, 1\}. \tag{15}$$

It is easily seen from figure 1 that constant MTE in the linear model corresponds to equation (14).[9] To implement this test, one can simply regress Y on D, Z and their interaction and perform a two-sided t-test on the interaction coefficient. If there are covariates in the MTE model, the regression can be performed conditional on X, and it is straightforward to test jointly for whether the MTE is constant for every value of X.

In the above test for external validity, we specified k_0 and k_1 as linear functions of p and then tested constant MTE versus the alternative of nonconstant MTE. In Appendix B, we consider two ways in which the linearity restrictions in k_0 and k_1 can be relaxed. We first provide a test in which k_0 and k_1 are specified as monotonic quadratic functions of p. Next, we consider a test in which all we assume is that k_0 and k_1 are monotonic functions of p. In both cases, the null hypothesis of constant MTE provides testable restrictions on (Δ_0, Δ_1). The choice of test involves a trade-off between power and sensitivity to the specification of k_0 and k_1;

[9] By 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. It is also evident from fig. 1 that the linear model does not impose any restrictions on (Δ_0, Δ_1) without further restrictions such as constant MTE.

weaker parametric assumptions on k_0 and k_1 provide weaker testable restrictions.[10]

C. *Identification with Separability*

In the absence of additional assumptions, proposition 1 shows that the parametric specification of k_1 and k_0 will need to be restrictive when the discrete instrument takes few values. This subsection demonstrates how an auxiliary assumption allows us to relax parametric restrictions on k_0 and k_1 using the separate estimation approach. The auxiliary assumption is as follows.

ASSUMPTION 2.

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

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 the MTE is additively separable in X and U_D:[11]

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

Assumption 2 is weaker than additive separability between D and X, as imposed in equation (1), which is a standard auxiliary assumption in applied work using IV.[12] Furthermore, assumptions 1 and 2 are implied by (but do not imply) full independence ($Z, X \perp U_1, U_0, U_D$), a common assumption in applied work estimating MTE (see, e.g., Carneiro and Lee 2009; Carneiro et al. 2011; Maestas et al. 2013).

A natural question is whether the separability assumption can be motivated by or is consistent with what is known or typically assumed about technology or preferences. Consider, for example, the literature on production function estimation. Suppose we study a production function for which Y_j is an output and X and U_j are input factors. Assumption 2 is then implied by perfect substitutability between the X and U_j inputs. By comparison, if Y_j, X, and U_j are measured in logs, assumption 2 is implied by unit elasticity between observable and unobservable inputs, as in the

[10] If the instrument takes more than two values, the parametric specification of k_1 and k_0 can be further relaxed. See also Heckman and Schmierer (2010) and Heckman et al. (2010) for tests in situations with continuous instruments.

[11] We could also relax the separability assumption to a subset of observable characteristics, so that the conditional expectation function of $U_1 - U_0$ as a function of U_D is allowed to depend on some (but not all) of the X variables. This would be achieved by performing the MTE estimation conditional on certain observable characteristics.

[12] Assumption 2 is weaker because it allows the treatment effects to vary by X and U_D, though not by the interaction of the two terms. By comparison, most applied work using IV assumes additive separability between D and X, as in eq. (1), which constrains the treatment effects to be the same for all values of X. This assumption implies not only that $E(Y_j|X, U_D) = \mu_j(X) + E(U_j|U_D)$ but also that $\mu_1(X) = \mu_0(X)$.

Cobb-Douglas technology. More generally, assumption 2 is compatible with a production technology in which unobserved productivity differences across agents are factor neutral, which is a standard assumption in methods of production function estimation (see, e.g., Olley and Pakes 1996; Levinsohn and Petrin 2003).[13]

In settings in which the separability assumption can be justified, we can use the separate estimation approach to relax the parametric restriction on k_1 and k_0. Example 2 illustrates what can and cannot be identified under assumptions 1 and 2. For simplicity, we consider a case with a single binary instrument, a single binary covariate, and k_0 and k_1 specified as quadratic functions of p.

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

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

and

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

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

From these expressions, we derive

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

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

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 identifies the quadratic MTE model.

Suppose we introduce a single binary covariate to the model, $X \in \{0, 1\}$. With a binary instrument $Z_- \in \{0, 1\}$, assumptions 1 and 2 give us four values of p for the treated and the untreated: $P(1,1) = p_1$, $P(0,1) = p_2$, $P(1,0) = p_3$, and $P(0,0) = p_4$, where $P(z_-, x)$ denotes $\Pr(D = 1 | Z_- = z_-, X = x)$. At the same time, we have additional parameters that we need to estimate since the model allows the intercepts μ_0 and μ_1 to

[13] The idea of treating errors as unobserved (factor-neutral) productivity at the firm level goes back at least to Marschak and Andrews (1944).

vary with X. Assume that Z_- is relevant for each value of X, implying that $p_1 \neq p_2$ and $p_3 \neq p_4$.

The LIV approach is based on the equation

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

In this equation, we have six parameters (μ_{00}, μ_{01}, $\mu_{10} - \mu_{00}$, $\mu_{11} - \mu_{01}$, $\alpha_{11} - \alpha_{01}$, and $\alpha_{12} - \alpha_{02}$) but only four values of p. This means that LIV cannot identify a quadratic MTE model with a binary Z_- and a binary X. In fact, the inclusion of X does not allow for identification of even a linear MTE model.

The separate estimation approach is based on the equations

$$E(Y|X = x, P(Z) = p, D = 0) = \mu_{00} + \mu_{01}x + \frac{1}{2}\alpha_{01}p + \frac{1}{3}\alpha_{02}p(p + 1) \tag{16}$$

and

$$E(Y|X = x, P(Z) = p, D = 1) = \mu_{10} + \mu_{11}x + \frac{1}{2}\alpha_{11}(p - 1) + \frac{1}{3}\alpha_{12}(p^2 - 1). \tag{17}$$

In each equation, we have four parameters and data that allow us to evaluate the expectation for four values of p. This means that the separate estimation approach identifies a quadratic MTE model with a binary Z_- and a binary X.

There is one exception to the conclusion in the above paragraph. Explicit specification of the linear equation system necessary to solve for the parameters in (16) and (17) shows that the parameters are uniquely identified if

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

The system will then have a unique solution, except if $p_1 + p_2 = p_3 + p_4$ (which is testable). In practice, a unique solution is likely in applications in which X is relevant given Z_-, so that $p_1 \neq p_3$ and $p_2 \neq p_4$. □

1. Geometry of MTE with Separability

Under assumptions 1 and 2, the MTE is additively separable in X and U_D and therefore identified from the marginal support of $P(Z)$ as opposed to the support of $P(Z)$ given X. Figure 2 uses the notation introduced in example 2 to illustrate how these assumptions allow the separate estimation approach to trace out the MTE across different observable characteristics X. As in example 2, we consider the case with a single binary instrument, a single binary covariate, and k_j specified as a quadratic function of p.

In the data, we observe the average outcome for each combination of treatment state, the instrument, and the covariate. From these averages we identify the corresponding conditional expectations, displayed for the treated as circles in panel A of figure 2. From the conditional expectations, we know the average derivative of K_1 in the interval (0.2, 0.4) and the average derivative of K_1 in the interval (0.5, 0.8). Because k_1 is a quadratic function, the derivative of K_1 is linear in p and we know the derivative at the midpoint of each interval. Panel B of figure 2 displays the derivative of K_1 at these midpoints. In the separate estimation approach, the dashed line that goes through the two points identifies α_{11} and α_{12}. In the same way, this approach identifies α_{01} and α_{02} from the average outcomes associated with the four values of the propensity score for the untreated. To identify the quadratic MTE model, we solve for the parameters in (16) and (17) for each value of X and recover $(\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11})$.

2. Identifying MTE with a Discrete Instrument

Example 2 showed how the separate estimation approach allows us to identify a quadratic MTE model with a single binary instrument and a single binary covariate. Proposition 2 states the general identification result for a discrete instrument under separability and conditional independence.

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 $P(Z)$, $p_1, \ldots, p_{MN} \in (0, 1)$. Assume further that k, k_1, and k_0 are specified as parametric functions, linear in parameters, with L parameters.

i. Using $E(Y|P(Z) = p, X = x)$, the MTEs can be identified only if $L \leq (N - 2)M + 1$.
ii. Using $E(Y_1|P(Z) = p, X = x, D = 1)$ and $E(Y_0|P(Z) = p, X = x, D = 0)$, the MTEs can be identified only if $L \leq (N - 1)M + 1$.

The proof is given in Appendix A.

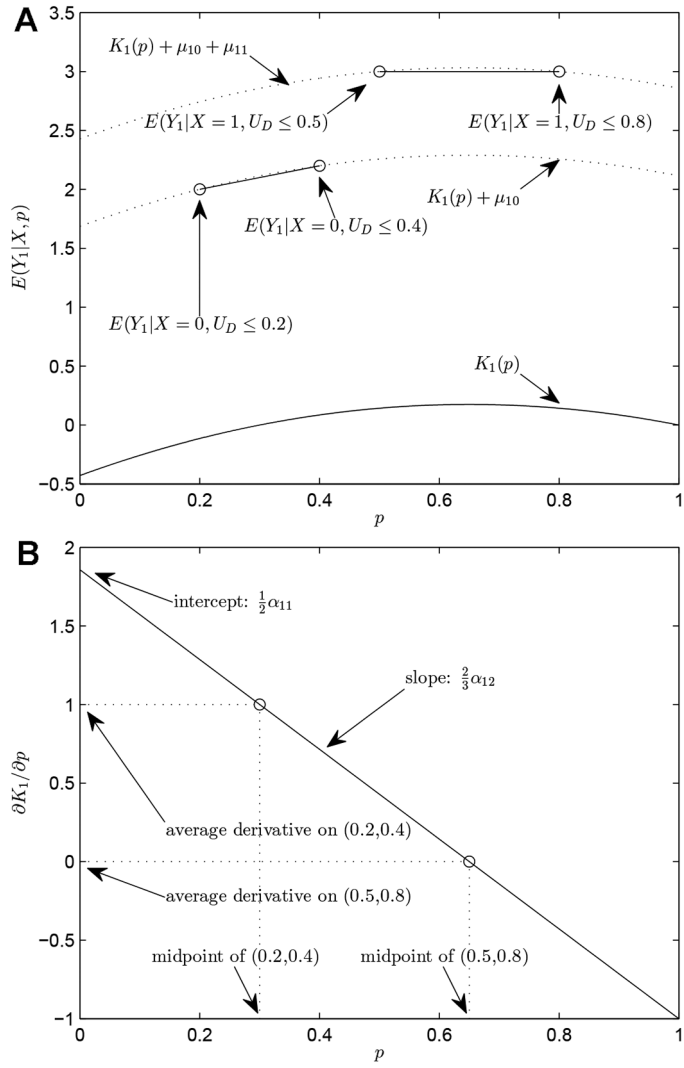

FIG. 2.—This figure illustrates the identification of the quadratic MTE model with a binary instrument and a binary covariate. Panel A displays the expected outcomes associated with the four values of the propensity score for the treated. Indicated by circles are the four conditional expectations $E(Y_1|X = j, P(Z) = p)$. From these points, we know the average derivative of K_1 in the interval (0.2, 0.4) and the average derivative of K_1 in the interval (0.5, 0.8). Because k_1 is a quadratic function, the derivative of K_1 is linear in p and we know the derivative at the midpoint of each interval. Panel B displays the derivative of K_1 at these midpoints. In the separate estimation approach, the dashed line that goes through the two points identifies α_{11} and α_{12}. The parameterization values used are $E(Y_1|X = 0, P(Z) = 0.2) = 2$, $E(Y_1|X = 0, P(Z) = 0.4) = 2.2$, $E(Y_1|X = 1, P(Z) = 0.5) = 3$, and $E(Y_1|X = 1, P(Z) = 0.8) = 3$.

The proposition shows what LIV and the separate estimation approach can identify under conditional independence and separability. If we impose restrictions on k_1 and k_0, the separate estimation approach can identify richer specifications of the MTE when the instrument is discrete. The reason is that the separability assumption allows us to combine information about the shapes of k_1 and k_0 across different values of X. In situations in which X or Z takes many values, the specification of the MTE may be relatively flexible as k_1 and k_0 may take a large number of parameters. However, even in cases in which $P(Z)$ is continuous because X contains a continuous component, it may be necessary to specify k_1 and k_0 as parametric functions taking a finite number of parameters. The reason is that it is also necessary to control for X in equation (9). By comparison, if $P(Z)$ is continuous because Z is continuous for each X, then the MTE can be nonparametrically identified (Heckman and Vytlacil 1999).

D. Discussion

The above propositions are useful in making precise what can and cannot be identified with discrete instruments. At the same time, they raise a number of questions. What choices have to be made to estimate the MTE with a discrete instrument, and how can one examine the sensitivity of results to these choices? Are there restrictions on the joint distribution of (U_0, U_1, U_D) that will imply a specific parametric MTE model? Can the specified MTE model be interpreted as an approximation as opposed to being exactly correct?

1. Empirical Specification and Sensitivity Checks

In thinking about model specification and sensitivity checks, it is useful to note that without any normalization of U_D, k_j can be written as

$$k_j(p) = E(U_j|F_{U_D}(U_D) = p) \equiv E(U_j|U_D = F^{-1}_{U_D}(p)),$$

where F_{U_D} is the strictly increasing distribution function of U_D. Define the conditional expectation of U_j as a function of U_D as $g_j(u) = E(U_j|U_D = u)$. Now

$$k_j(p) = g_j(F^{-1}_{U_D}(p)). \tag{18}$$

For any choice of F_{U_D}, one can obtain any function k_j by setting $g_j(u) = k_j(F_{U_D}(u))$. This means that the specification of U_D is simply a normalization that has no implications for statements we would make about the MTE.

What matters for the empirical MTE model is how we specify k_j. In the empirical analysis, we examine the sensitivity of the results to different

choices of k_j. For example, our baseline model under separability specifies k_j as local polynomials of p. We probe the stability of the baseline results to several alternative specifications. We report results in which k_j is specified as different functions of p, including splines, global polynomial series, and global Fourier series. Additionally, we examine how the results change if k_j is specified as a polynomial function of $\Phi^{-1}(p)$, where Φ is the distribution function of the standard normal distribution. It is reassuring to find that our results are robust to the empirical specification of the MTE model.

2. Parametric Structure and Distribution of Unobservables

Following Heckman and Robb (1985), empirical studies often specify the MTE model as a global polynomial function of p. A natural question is then, What restrictions on the joint distribution of (U_0, U_1, U_D) imply that k_j will be a global polynomial of a given order?

The class of distributions of (U_0, U_1, U_D) implying that k_j is a polynomial in $F_{U_D}^{-1}(p)$ is characterized by

$$U_j = \sum_{i=0}^{M} \lambda_{ij} F_{U_D}^{-1}(p)^i + W_j, \quad j = 0, 1,$$

where W_0 and W_1 are random variables that are mean independent of U_D. To see this, specify g_j in equation (18) to be a global polynomial function, giving

$$k_j(p) = \sum_{i=0}^{M} \lambda_{ij} F_{U_D}^{-1}(p)^i, \quad j = 0, 1. \tag{19}$$

Different subclasses give different specifications of k_j. One example is the joint distributions of (U_0, U_1, U_D) that belong to the subclass of elliptically symmetric distributions (Fang, Kotz, and Ng 1990). In this case, k_j will be linear in $F_{U_D}^{-1}(p)$. One member of this subclass is jointly normally distributed (U_0, U_1, U_D), in which case k_j is linear in $F_{U_D}^{-1}(p) = \Phi^{-1}(p)$, as in the normal selection model. Another example is the subclass of distributions for which $U_j = \alpha_j p + W_j$. In this case, k_j is linear in $F_{U_D}^{-1}(p) = p$, as in example 1.

3. Approximating Model

In some settings with discrete instruments, it may be useful to think of the empirical specification of the MTE model being an approximation as opposed to being exactly correct.[14]

[14] Standard approximation results for linear regression models cannot be directly applied (e.g., White 1980). To see this, recall that we are estimating $E(Y_1|U_D \leq p, X = x)$ and

In Appendix C, we consider how the analysis is affected if the empirical specification of k_j is an approximating model. In particular, suppose that the true model of k_j is a polynomial in $F_{U_D}^{-1}(p)$ of order M_2, whereas the empirical model of k_j is a polynomial in $F_{U_D}^{-1}(p)$ of order $M_1 < M_2$. This appendix provides a limit statement of the distance between the true model and the probability limit of the empirical model. For example, if $k_1(p) = \alpha_0 + \alpha_1 p + \alpha_2 p^2$ in the true model and we estimate a linear model with $k_1(p) = \beta_0 + \beta_1 p$, then the probability limits of the estimators of β_0 and β_1 approach α_0 and α_1 at the same rate as α_2 goes to zero. This means that if the MTE is approximately linear, as in the empirical analysis of Carneiro et al. (2011), then the approximation error from specifying k_0 and k_1 as linear functions of p will be small.

IV. Empirical Analysis

In this section, we apply the separate estimation approach to perform an MTE analysis of how family size affects children's educational attainment.

A. *Data and Summary Statistics*

As in Black et al. (2005), our data are based on administrative registers from Statistics Norway covering the entire resident population of Norway who were between ages 16 and 74 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 contain 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 et al.'s study 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 et al. (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 at first parity 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 five siblings as well as a handful of families in which the mother had a birth before she was age 16 or after she was 49.

$E(Y_0|U_D > p, X = x)$ in order to recover $E(Y_1|U_D = p, X = x)$ and $E(Y_0|U_D = p, X = x)$. The best linear approximations of the former functions are generally not the best linear approximations of the latter functions (and therefore not of the MTE).

In addition, we exclude a few children for whom their own or their mother's education is missing. Rather than dropping the larger number of observations for which information on fathers is missing, we include a separate category of missing for father's education and father's age.

As in Black et al. (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 firstborn child from being in a family with two or more siblings rather than one 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, while parental education 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 (second- and third-born children are twins). The validity of 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. We refer to Black et al. (2005) and Angrist et al. (2010) for empirical evidence in support of the validity of the instruments.

Our sample consists of 514,004 firstborn children with at least one sibling. Table 1 displays basic descriptive statistics. In 50 percent of the sample, there are at least three children in the family, and the average family size is 2.7 children. There are a few noticeable differences between children from a family with only one sibling and those with two or more siblings. As expected, 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. Furthermore, the second- and third-born children are twins in about 1 percent of the families. It is also evident that firstborn children with one sibling have higher educational attainment than firstborn children with two or more siblings, pointing to a negative association between family size and child quality. However, parents with more children tend to be less educated and younger at first birth, suggesting that we need to be cautious in giving the association between family size and child quality a causal interpretation.

TABLE 1
DESCRIPTIVE STATISTICS: ESTIMATION SAMPLE

	MEAN		
	All	Two Children	Three+ Children
Outcome:			
Years of schooling	12.3	12.5	12.1
Instruments:			
Same sex, first and second children	.501	.471	.529
Twins at second birth	.010	0	.019
Endogenous regressor:			
At least three children	.502	0	1
Covariates:			
Female	.473	.475	.471
Age in 2000	39.5	38.0	40.9
Mother's age at first birth	24.0	24.6	23.3
Father's age at first birth	26.8	27.1	26.4
Mother's years of schooling	10.0	10.1	9.9
Father's years of schooling	10.1	10.2	10.0
Observations	514,004	255,933	258,071

NOTE.—Descriptive statistics are for 514,004 children. All children are firstborn 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.

B. Fertility Decision Model

Table 2 presents estimates of the average marginal effects from a logit model for the choice of having three or more children (instead of two children). In terms of the choice model defined by (3), I_D represents the net benefit from having another child, 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 covariates are exogenous; what we assume is that the instruments are independent of the unobservables conditional on the covariates. The twins instrument is interacted with all other variables to ensure that the propensity score is one in the event of a twin birth.

The instruments are (individually and jointly) strong predictors of family size. The average effect of a twin birth is almost 0.52. This means that nearly 48 percent of mothers with twins at second parity 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. Mother's age at first birth is also predictive of family size: The propensity score decreases by 2.2 percentage points if the mother is 1 year older at the first birth.

TABLE 2
FERTILITY DECISION MODEL: AVERAGE DERIVATIVES

	Average Effect
Covariates:	
Age in 2000	.0126 (.0001)
Mother's age at first birth	−.0215 (.0002)
Father's age at first birth	−.0019 (.0002)
Mother's years of schooling	.0017 (.0008)
Father's years of schooling	−.0066 (.0015)
Female	.0013 (.0015)
Instruments:	
Same sex, first and second children	.0567 (.0012)
Twins at second parity	.5179 (.0007)

NOTE.—This table reports the average partial effect (average treatment effect for binary variables) from a logit model. The dependent variable is the probability of being in a family with two or more siblings rather than one sibling. All covariates in the table enter the model linearly. In addition, we add second- and third-order terms in age in 2000, mother's age at first birth, father's age at first birth, mother's years of schooling, and father's years of schooling. We also include interactions between the first-order terms of all covariates. The instrument same sex, first and second children enters the model without interaction terms. The instrument twins at second 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 parentheses are computed by nonparametric bootstrap with 100 bootstrap replications.

C. *IV Estimates*

Following Black et al. (2005), we specify the following outcome equation (second stage):

$$Y = \mu + \beta D + X'\delta + \epsilon, \tag{20}$$

where Y denotes the child's years of schooling, X is a vector of controls for (predetermined) child and parental characteristics, and ϵ is the error term. In line with much of the previous literature, throughout the empirical analysis we will focus on the treatment effect on a firstborn child from being in a family with two or more siblings ($D = 1$) rather than one 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. In every specification, we include the same vector of observables $Z = (X, Z_{-})$ in the first-stage equations. What we change is the instrument(s) excluded from the outcome equation (Z_{-}).

In column 1, we follow Carneiro et al. (2011) in performing linear IV with $P(Z)$ as the instrument. We construct $P(Z)$ using the parameter esti-

TABLE 3
OLS AND IV ESTIMATES

	$P(Z)$ as Instrument (1)		Z_- as Instrument (2)
IV:			
Same-sex instrument	−.208		.174
	(.105)		(.115)
Twins instrument	−.065		.050
	(.060)		(.062)
Both instruments	−.015		.076
	(.053)		(.055)
OLS		−.052	
		(.007)	

NOTE.—This table reports OLS and IV estimates of the effect of family size on the educational attainment of firstborn children. Column 1 reports linear IV estimates with $P(Z)$ as instrument. We construct $P(Z)$ using the parameter estimates from the logit model with average derivatives reported in table 2. Column 2 reports standard linear IV estimates with Z_- as instrument. We use the same specification for the covariates as reported in table 2. The first row excludes the same sex, first and second children instrument from the second stage, the second row excludes the twins at second parity instrument from the second stage, and the third row excludes both instruments from the second stage. The OLS estimate of the second-stage specification (20) is reported in the fourth row. Standard errors in parentheses are robust to heteroskedasticity.

mates from the logit model, for which average marginal effects are reported in table 2. When excluding the same-sex instrument from the outcome equation, we estimate that being in a family with two or more siblings rather than one sibling lowers the educational attainment of firstborn children by 0.208 year. If instead we exclude the twins instrument from the outcome equation, we still find a negative point estimate but cannot reject no effect of family size at conventional significance levels. When we exclude both instruments from the outcome equation, the IV estimate is close to zero. Indeed, the LATE based on both instruments is significantly different from the LATE based on the same-sex (twins) instrument at the 5 (10) percent significance level.[15] The fact that the LATEs vary significantly with the choice of excluded instrument indicates nonconstant MTEs.

In column 2, we follow Black et al. (2005) in using a standard linear IV procedure with Z_- as the excluded instruments. While the effect of family size induced by twins is only 0.050, the effect based on the same-sex instrument is as large as 0.174. The IV estimates in columns 1 and 2 are consistent under the same assumptions (Carneiro et al. 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

[15] Tests comparing different coefficients in table 3 are performed using nonparametric bootstrap with 1,000 bootstrap replications.

same as the LATE of a Z_{-} shift in D. Indeed, the IV estimates differ substantially between columns 1 and 2.

D. *MTE Weights of Treatment Parameters*

As a first step toward understanding why the IV estimates vary so much with the choice of excluded instrument, we estimate the distribution of weights across the MTEs. 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 (average treatment effect on the untreated). The expressions and interpretations for these weights are provided in Heckman and Vytlacil (2007). 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 three or more children ($D = 1$) rather than two

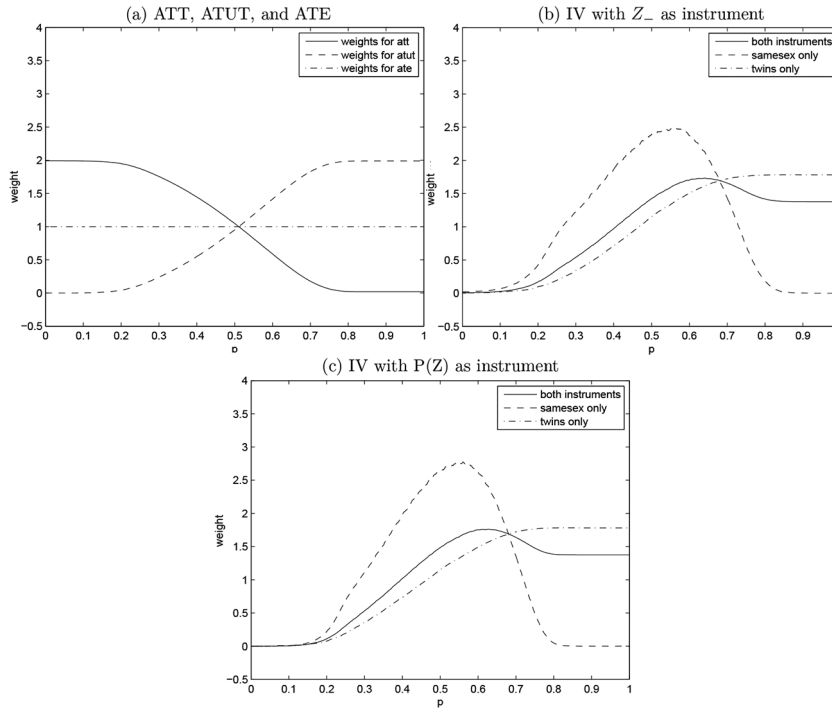

FIG. 3.—These figures show the weights of MTE for treatment effects parameters and instruments. Panel a graphs MTE weights associated with the ATT, the ATE, and the ATUT. Panels b and c graph MTE weights associated with the IV estimates presented in table 3. The formulas for these weights are provided in Heckman and Vytlacil (2007). 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 three or more children rather than two children. A high value of p means that a family is less likely to have three or more children.

children ($D = 0$). Recall that a high value of U_D means that a family is less likely to have three 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 IV estimates based on the same-sex instrument. This pattern is quite intuitive: With twin births, 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 preference for mixed-sex sibship induces them to have a third child. Indeed, there are no never-takers with twins, which gives support on $p = 1$. This explains why the twins instrument assigns weights to individuals with high values of U_D.

Second, both ATT and ATE assign much more weight to families who are likely to have another child 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 for the twin birth compliers should therefore be equal to the ATUT. This explains why the distribution of weights with the twins instrument closely mirrors the distribution of weights for the ATUT.

E. Nonparametric Bounds

The sensitivity of the LATEs of family size to the choice of excluded instrument raises concerns about their policy-relevant and external validity. One possible response is to abandon the goal of point identification and instead construct bounds on conventional treatment parameters. In Appendix E, we derive nonparametric bounds on the ATE and the ATT of family size.

A natural starting point is the worst-case (or no-assumption) bounds (see Manski 1989). To construct these bounds, we exploit that the dependent variable—years of schooling—is bounded between 6 and 21. As shown in the first row of Appendix table E1, the worst-case bounds are very wide and therefore not informative about the effects of family size. In the other rows of table E1, we impose the IV assumptions of exclusion, monotonicity, and relevance.[16] Under these assumptions, we can estimate

[16] Alternatively, one could invoke the monotone IV assumption (Manski and Pepper 2000). This assumption is a weakened form of the exclusion restriction. As a result, it will necessarily produce wider bounds than the standard IV assumptions. Indeed, the monotone IV assumption does little to tighten the worst-case bounds in our setting.

the LATE, the average potential outcomes of always-takers if treated, and the average potential outcomes of never-takers if nontreated. Taken together, these components allow us to construct bounds on the ATE and the ATT. Unfortunately, the bounds are too wide to be informative about the effects of family size on children's schooling.

In principle, one could try to get tighter bounds by invoking the assumptions of monotone treatment response (Manski 1997) or monotone treatment selection (Manski and Pepper 2000). The former assumption requires that increasing family size does not raise children's schooling; the latter assumption states that children with fewer siblings would have more schooling at each level of family size than children with more siblings. We have chosen not to invoke these assumptions, as theory is not informative about the expected sign of the selection to or effects of additional siblings.

V. Empirical Analysis of Marginal Treatment Effects

We now examine the scope for moving beyond the LATEs of family size by adopting a functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment. We begin by using the separate estimation approach to estimate a linear MTE model and use it to test the external validity of LATE. We next invoke the additional assumption of additive separability between observed and unobserved heterogeneity in treatment effects and estimate a flexible specification of the MTE model. For now, we exclude only the same-sex instrument from the outcome equation, but we will later provide estimates excluding both instruments.

In the separate estimation approach, the empirical MTE model is determined by the specification of k_1 and k_0. The MTEs can then be recovered from the empirical analogues of $E(Y|P(Z) = p, X = x, D = 1)$, $E(Y|P(Z) = p, X = x, D = 0)$, and $P(Z)$. Regardless of whether parametric or nonparametric methods are used, it is necessary to specify a modeling framework for k_0 and k_1. Two key choices have to be made. The first is to select a class of functions for k_0 and k_1. The second is to select the number of parameters to estimate. This amounts to choosing bandwidth in a local polynomial regression or selecting the number of terms in a series estimation.

In the first part of our empirical analysis, we assume conditional independence but not additive separability. As a result, the MTE is identified from the support of $P(Z)$ given X as opposed to the marginal support of $P(Z)$. With a binary instrument, we then need to specify k_1 and k_0 as parametric functions of p with no more than two parameters in each function.

Our baseline specification normalizes U_D to be unit uniform and specifies k_0 and k_1 as linear functions of p.

In the second part of our empirical analysis, we assume conditional independence and separability. As a result, $P(Z)$ takes a large number of unique values for both the treated and untreated. We can therefore be very flexible in the specification of k_0 and k_1. Our baseline specification normalizes U_D to be unit uniform and uses a local quadratic estimator in which the bandwidths are selected using the "leave one out" cross-validation method.[17] In Section V.C, we perform a number of sensitivity checks, showing that our results are robust to alternative specifications of k_j.

A. *Linear MTE Model and External Validity of LATE*

Table 4 displays results from an MTE model without covariates in which k_0 and k_1 are linear in p and (U_0, U_1, U_D) is independent of Z. Panel A shows estimates of the intercept and the slope of the linear MTE 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 the standard linear IV procedure.

The results in table 4 illustrate that in situations with a binary instrument, the separate estimation approach gives an estimate of LATE from the linear MTE model that is numerically equivalent to the estimate of LATE from standard IV estimation. By applying the standard IV estimator, we obtain the following estimate of the LATE of family size on children's education:

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

By comparison, the separate estimation approach recovers the LATE from estimating

$$\hat{\mu}_1 + \hat{K}_1(p) = 11.720 + 0.773p,$$
$$\hat{\mu}_0 + \hat{K}_0(p) = 12.780 - 0.216p,$$

[17] See Ichimura and Todd (2007) for a discussion of choice of bandwidth. If the propensity score takes enough values to identify the true MTE model, cross-validation will pick the correct model with probability approaching one as the sample size grows. The standard t-tests will therefore be asymptotically valid, even after model selection. However, the usual finite sample issues of inference after model selection still apply, as the asymptotic distribution may not approximate the finite sample distribution well. See Leeb and Pötscher (2005) for a discussion.

TABLE 4
ESTIMATES OF LINEAR MTE MODEL AND LATE BASED ON SAME-SEX INSTRUMENT

	$p = .473$	$p = .531$	Intercept	Slope
	A. Estimates of Linear MTE Model and Its Components			
Linear MTE model:				
$\mu_1 + K_1(P) = E(Y_1 \mid U_D < p)$	12.086	12.131	11.720	$+.775p$
	(.008)	(.007)	(.095)	(.188)
$\mu_0 + K_0(P) = E(Y_0 \mid U_D > p)$	12.462	12.450	12.564	$-.216p$
	(.007)	(.008)	(.091)	(.181)
$\mu_1 + k_1(p) = E(Y_1 \mid U_D = p)$	12.453	12.542	11.720	$+1.550p$
	(.084)	(.105)	(.095)	(.376)
$\mu_0 + k_0(p) = E(Y_0 \mid U_D = p)$	12.576	12.551	12.780	$-.432p$
	(.101)	(.080)	(.272)	(.0362)
$\text{MTE}(p) = E(Y_1 - Y_0 \mid U_D = p)$	$-.123$	$-.008$	-1.006	$+1.981p$
	(.129)	(.130)	(.290)	(.529)
	B. LATE from IV and Linear MTE Model			
Instrumental variables:				
$[E(Y \mid \Pr(D) = .531) - E(Y \mid \Pr(D) = .473)]/(.531 - .473)$				$-.065$
				(.129)
Linear MTE model:				
$\int_{.471}^{.531} \text{MTE}(p) = \text{MTE}[(.531 + .471)/2]$				$-.065$
				(.129)

NOTE.—This table displays LATE and linear MTE estimates of family size on the educational attainment of firstborn children. Panel A reports estimates from the linear MTE model with same sex, first and second as the excluded instrument. Panel B reports estimates of LATE from the IV estimator and the linear MTE model, with same sex, first and second as the excluded instrument. We do not include any covariates in the MTE estimation or the IV estimation. Standard errors in parentheses are computed by nonparametric bootstrap with 100 bootstrap replications.

and

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

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

In the separate estimation approach, the last step to derive the LATE is

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

By comparison to the standard IV estimator, estimates of the linear MTE model from the separate estimation approach offer 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 informative only about the same-sex-induced effect of family size. If we extrapolate the linear MTE estimate, we obtain an ATE of −0.016 and an ATT of 0.48. However, we do not put much stock in these extrapolations, given the nonlinear pattern that we find in the analysis with separability.

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

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

versus a two-sided alternative. To perform the test, we regress Y on D, Z and their interaction. On the basis of the two-sided t-test on the interaction coefficient, we reject the null hypothesis of a constant MTE at the 1 percent significance level (p-value .0001).

There are two caveats to the rejection of the external validity of the LATE. The first is that the same-sex instrument may be correlated with variables other than family size. If these variables affect children's education, then Z depends on (U_0, U_1, 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. Although most of the LATEs are too imprecisely estimated to draw firm 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 .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.

The second caveat is that the test of external validity relies on k_1 and k_0 being specified as linear functions of p. As usual, the test is not valid if the model is misspecified (i.e., k_1 or k_0 is not a linear function of p). Misspecification can lead us to falsely reject the null hypothesis of constant MTE, and it can make us fail to correctly reject the null hypothesis of constant MTE. In Appendix B, we consider two ways in which the linearity restrictions in k_0 and k_1 can be relaxed. We first provide a test in which k_1 and k_0 are specified as monotonic quadratic functions of p. Our results show that in this case we also can reject the null hypothesis of constant MTE in favor of the alternative of nonconstant MTE at the 1 percent significance level (p-value .0016). Next, we consider a test in which we assume that k_1 and k_0 are only monotonic functions of p. In this case, there is not enough power to reject the null hypothesis of constant MTE at conventional levels of significance.

B. *Flexible MTE Model with Separability*

If we are willing to assume only that (U_0, U_1, U_D) is independent of Z given X (assumption 1), then a binary instrument identifies a linear MTE model only. This means that unless one is willing to use the linear MTE model 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 additional assumption of additive separability between observed and unobserved heterogeneity in treatment effects (assumption 2).

Figure 4 shows the empirical support of $P(Z) \equiv \Pr(D = 1|Z)$ under assumptions 1 and 2, using same-sex as the excluded 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 $D = 0$. As for the IV estimates reported in table 3, we construct $P(Z)$ using the parameter estimates from the logit model for which average marginal effects are reported in table 2. We see that assumptions 1 and 2 yield substantial support in the interval (0.20, 0.75). We do not, however, obtain substantial support outside this interval, which implies that we cannot identify MTE as U_D approaches zero or one. Following Carneiro et al. (2011), we trim our data by dropping observations for which the estimated $P(Z)$ is below 0.209 or above 0.751, which correspond to the 0.01 and 0.99 percentiles in the distributions of $P(Z)$ for the treated and untreated, respectively.[18]

[18] This sample restriction barely moves our MTE estimates.

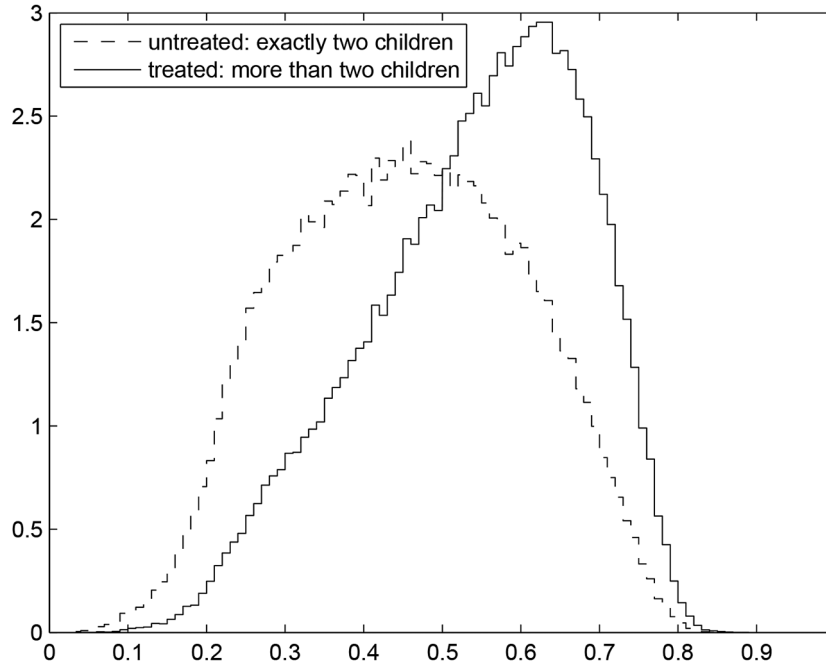

FIG. 4.—Histogram of propensity scores for the treated (solid) and the untreated (dotted), with same-sex instrument as the excluded instrument. This histogram shows the empirical support of $P(Z) \equiv \Pr(D = 1|Z)$ under assumptions 1 and 2, with same-sex as the excluded instrument. The common support 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 specified in the note to table 2.

Under assumptions 1 and 2, the MTE is additively separable in X and U_D and therefore identified from the marginal support of $P(Z)$ as opposed to the support of $P(Z)$ given X. This allows us to be very flexible in the specification of k_0 and k_1. In particular, we have 32,446 (33,093) unique values for the pairs of $\Pr(D = 1|X = x, Z_{\text{samesex}} = 0)$ and $\Pr(D = 1|X = x, Z_{\text{samesex}} = 1)$ for the treated (untreated). As a result, in principle we can identify an MTE model with more than 32,000 parameters in the specification of both k_1 and k_0. We therefore think of $P(Z)$ as taking enough values that we can interpret the local quadratic model as an exact specification with no approximation error.

Our baseline model uses an estimation procedure that follows closely the approach used in Heckman et al. (2006) and Carneiro et al. (2011). The first step is to estimate $P(Z)$. We obtain $\hat{P}(Z)$ from the parameter estimates of the logit model reported in table 2. Our specification is quite flexible, and alternative functional form specifications for the choice model (e.g., probit or linear probability model) produce results sim-

ilar to the ones reported here. The second step is to estimate $\mu_0(X)$, $\mu_1(X)$, K_0, and K_1. To this end, we use the double residual regression method of Robinson (1988) as modified by Heckman, Ichimura, and Todd (1997). This amounts to first estimating $\mu_1(X)$ and $\mu_0(X)$, using the same specification as in the IV estimation of table 3, and next estimating the functions K_1 and K_0 using local quadratic regression of $Y_1 - \hat{\mu}_1(X)$ and $Y_0 - \hat{\mu}_0(X)$ on $\hat{P}(Z)$. The bandwidths are selected using the "leave one out" cross-validation method.

Figure 5 displays how the MTE depends on U_D, with 95 percent confidence intervals computed from a nonparametric bootstrap.[19] The MTE estimates are evaluated at mean values of X. Our estimates suggest that the effects of family size vary in magnitude and even sign (i.e., β is heterogeneous) and that families act as if they possess some knowledge of their idiosyncratic return (β is correlated with D). Specifically, our estimates show that an increase in family size raises the average educational attainment of firstborn children in families with U_D less than 0.40. This means that firstborn 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 trade-off in families for whom preferences for mixed-sibling sex composition play a more important role in the decision to have another child. For values of U_D above 0.62, the estimated MTEs are positive. This means that in families unlikely to have a third child (possibly because unobserved psychic or financial costs are too high), the educational attainment of firstborn would benefit from an increase in family size.[20]

As in Carneiro, Heckman, and Vytlacil (2003), the MTE estimates show a U shape and the magnitude of heterogeneity is substantial. This pattern could not be recovered by the traditional approach to estimating the model of equations (2) and (3), which assumes that (U_0, U_1, U_D) are jointly normally distributed and independent of Z (see, e.g., Björklund and Moffitt 1987).[21] In particular, the normal selection model may mask heterogeneity in the effects of family size if the population is segmented in preferences or constraints. For example, preference for mixed-sex sibships is

[19] Heckman et al. (1997) show that 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 reestimate $P(Z)$ so all standard errors account for the fact that $\hat{P}(Z)$ is itself an estimated object.

[20] The QQ model of fertility 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; Mogstad and Wiswall 2016). See also the discussion in App. D.

[21] The normal selection model assumes that the MTE is either constant, monotonically declining, or monotonically increasing in U_D; the MTE tends toward $\pm\infty$ as U_D tends toward zero or one (unless the MTE is constant); and 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 $U_D = 1 - u$.

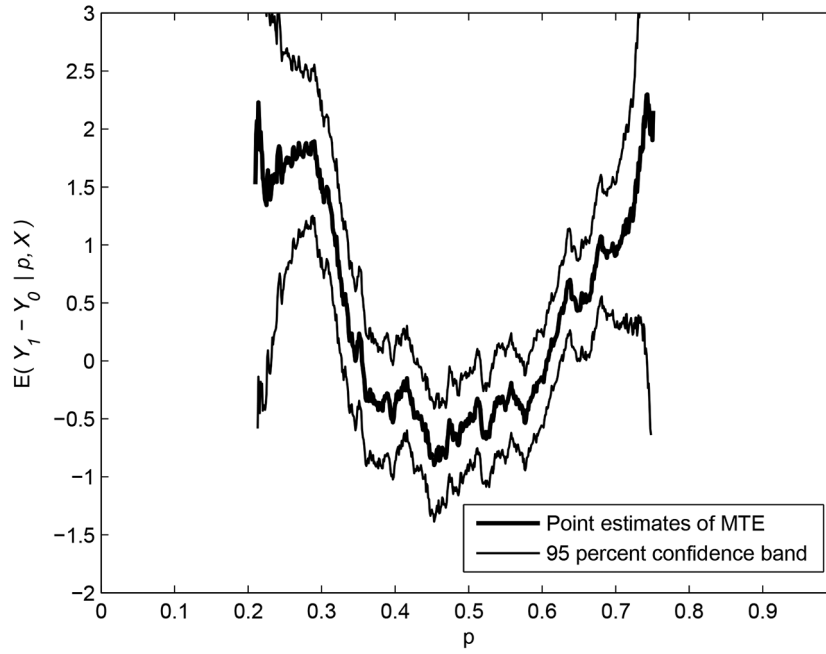

FIG. 5.—This figure displays the MTE estimates based on assumptions 1 and 2. We use same-sex as the excluded instrument. The MTE estimates are evaluated at the mean values of the covariates. We construct $P(Z)$ using the parameter estimates from the logit model specified in the note to table 2. The MTE estimates are based on a double residual regression separately for the treated and nontreated, using a local quadratic regression with uniform kernel and a bandwidth of 0.0615. The 95 percent confidence band (dashed lines) is computed from a nonparametric 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 three or more children rather than two children. A high value of p means that a family is less likely to have three or more children.

unlikely to be manifested with equal force by all groups in the population. Mixture distributions arise naturally when the population contains two or more distinct subpopulations.

In Appendix D, we present a simple example of MTE in a mixture model with two subpopulations of equal size. The population distribution of MTE that is 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.

We have already discussed the rejection of the hypothesis that MTE is constant in U_D on the basis of the estimates of the linear MTE model, but

without invoking assumption 2. With this separability assumption, we can test whether the MTE is constant in U_D or not. We evaluate the MTE in five intervals equally spaced between 0.22 and 0.72. As in Carneiro et al. (2011), we construct pairs of intervals and compare the mean of the MTEs for each pair. Table 5 reports the outcome of these comparisons. For example, column 1 reports

$$E(Y_1 - Y_0 | X = \bar{X}, 0.22 \leq U_D \leq 0.27)$$
$$-E(Y_1 - Y_0 | X = \bar{X}, 0.31 \leq U_D \leq 0.36) = 1.102,$$

with a p-value of .034 for no difference in these LATEs. Table 6 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 on the basis of heterogeneous returns.

C. *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 et al. (2011) in rescaling the weights so that they integrate to one over the region of common support.

TABLE 5
COMPARING LATES ACROSS DIFFERENT INTERVALS OF THE PROPENSITY SCORE

	LATE OVER INTERVALS				
	(.22, .27) − (.31, .36)	(.31, .36) − (.40, .45)	(.40, .45) − (.49, .54)	(.49, .54) − (.58, .63)	(.58, .63) − (.67, .72)
Point estimate	1.102	1.011	.046	−.413	−1.006
Standard error	.521	.307	.257	.241	.301
p-value	.034	.001	.859	.087	.001
p-value of joint test			.000		

NOTE.—This table reports tests of constant MTE of family size on the educational attainment of firstborn children. The MTE estimates are based on assumptions 1 and 2, with same sex, first and second as the excluded instrument (see fig. 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 nontreated, using local quadratic regression with uniform kernel and bandwidth of 0.0615. 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
SUMMARY MEASURES OF TREATMENT EFFECTS

	ATE	ATT
Rescaled support parameters	.474	.654
	(.136)	(.210)

NOTE.—This table reports rescaled support estimates of ATE and ATT of family size on the educational attainment of firstborn children. We use estimates of MTE in the interval (0.209, 0.751) and rescale the weights to integrate to one over this region. The MTE estimates are based on assumptions 1 and 2, with same sex, first and second as the excluded instrument (see fig. 5). We construct $P(Z)$ using the parameter estimates from the logit model specified in the note to table 2. The MTE estimates are based on double residual regression separately for the treated and nontreated, using local quadratic regression with uniform kernel and bandwidth of 0.0615. Standard errors are based on nonparametric bootstrap (of each estimation stage), with 100 bootstrap replications.

We use the MTE estimates reported in figure 5 to construct rescaled estimates on the ATE and the ATT. Table 6 displays the lower rescaled support estimates. The rescaled support estimates are 0.474 for the ATE and 0.654 for the ATT. 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. This illustrates the need to be cautious in going from the mean impact for compliers to the effects for other parts of the population.

D. Sensitivity Analysis

1. Model Validation Using the Twins Instrument

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

Figure 6 compares estimates of the MTE based on the same-sex instrument to those using both the same-sex and the twins instrument. In each case, we use the estimation procedure as described above. In particular, the specification of the choice model is fixed; what we change is the instrument(s) excluded from the outcome equation. In our baseline specification, the dummy variable for twin birth is included in both stages and the same-sex dummy variable is the only excluded instrument. In the model

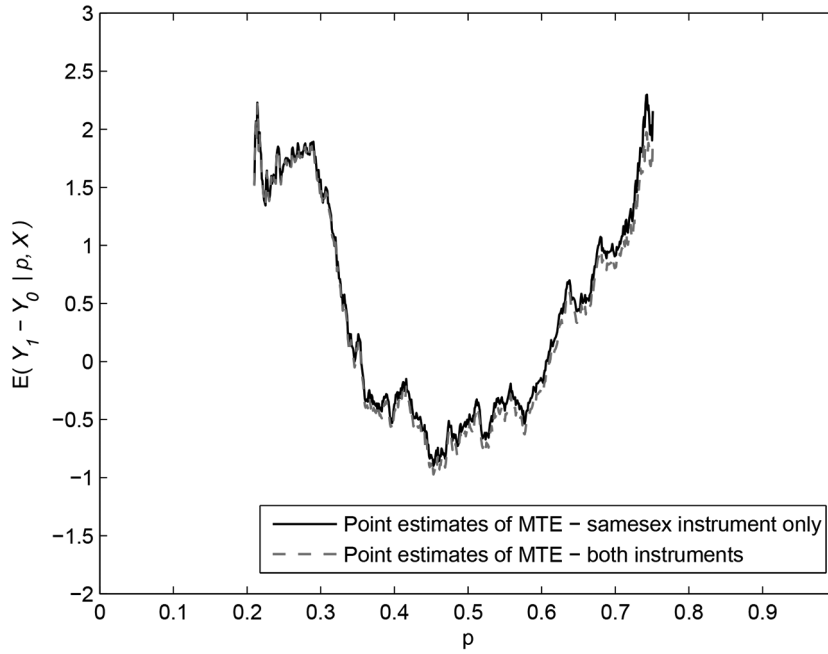

FIG. 6.—This figure displays the MTE estimates based on assumptions 1 and 2 with the same-sex instrument only and with both same-sex and twins instruments. We show estimates with same sex, first and second as the only excluded instrument (solid line) and when both the same sex, first and second instrument and the twins at second parity instrument are excluded from the outcome equation (dashed line). The MTE estimates are evaluated at the mean values of the covariates. We construct $P(Z)$ using the parameter estimates from the logit model specified in the note to table 2. The MTE estimates are based on a double residual regression separately for the treated and nontreated, using a local quadratic regression with uniform kernel and a bandwidth of 0.0615. 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 three or more children rather than two children. A high value of p means that a family is less likely to have three or more children. Color version available as an online enhancement.

validation, we treat twin birth as an additional excluded instrument. It is reassuring to find that the two MTE estimates display the same U-shaped pattern. Indeed, the point estimates are very similar in magnitude. This finding suggests that the differences in the IV estimates by the choice of excluded instrument occur because of different weighting of the MTE rather than invalidity of the instruments (or assumption 2).

A limitation with the validation exercise presented in figure 6 is that the same-sex instrument could be driving both MTE estimates. To address this concern, it would be useful to estimate the MTE separately for each excluded 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 since there are both always-takers and compliers. Figure 7 shows how the estimates of the function k_1 vary with the choice of excluded instrument. For each instrument, we use the same estimation procedure as described above, keeping the specification of the choice model fixed. The similarity in the estimates of k_1 gives credibility to the MTE estimates reported in figure 5.

2. Parametric Specification

The empirical MTE model is determined by the specification of k_0 and k_1. The estimates in figure 5 are based on a model that specifies these func-

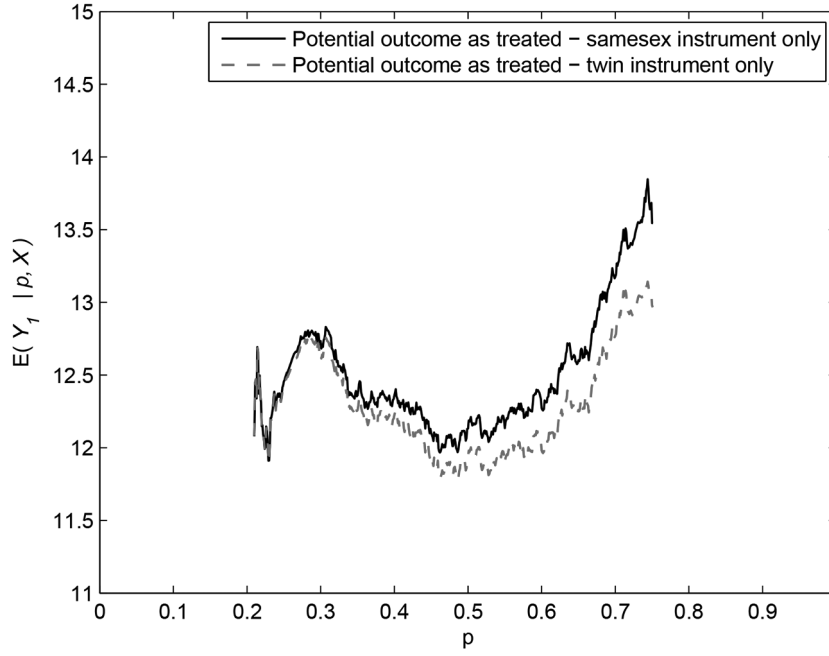

FIG. 7.—This figure uses the sample of treated and displays the estimates of the expected outcome as treated $\mu_1(X) + k_1(p)$, for each instrument, based on assumptions 1 and 2. We show estimates with same sex, first and second as the only excluded instrument (solid line) and with twins at second parity as the only excluded instrument (dashed line). The estimates are evaluated at the mean values of the covariates. We construct $P(Z)$ using the parameter estimates from the logit model specified in the note to table 2. The estimates are based on a double residual regression on the sample of treated, using a local quadratic regression with uniform kernel and a bandwidth of 0.0615. The y-axis measures the outcome in years of schooling, whereas the x-axis represents the unobserved component of parents' net gain from having three or more children rather than two children. A high value of p means that a family is less likely to have three or more children. Color version available as an online enhancement.

tions as local polynomials of p. We now probe the stability of our baseline estimates to alternative specifications of the empirical MTE model.

In figure 8, we show sensitivity of the MTE estimates to specifying k_0 and k_1 as different functions of p. Panel a compares our baseline MTE estimates from figure 5 with a tenth-order global polynomial, a fifth-order

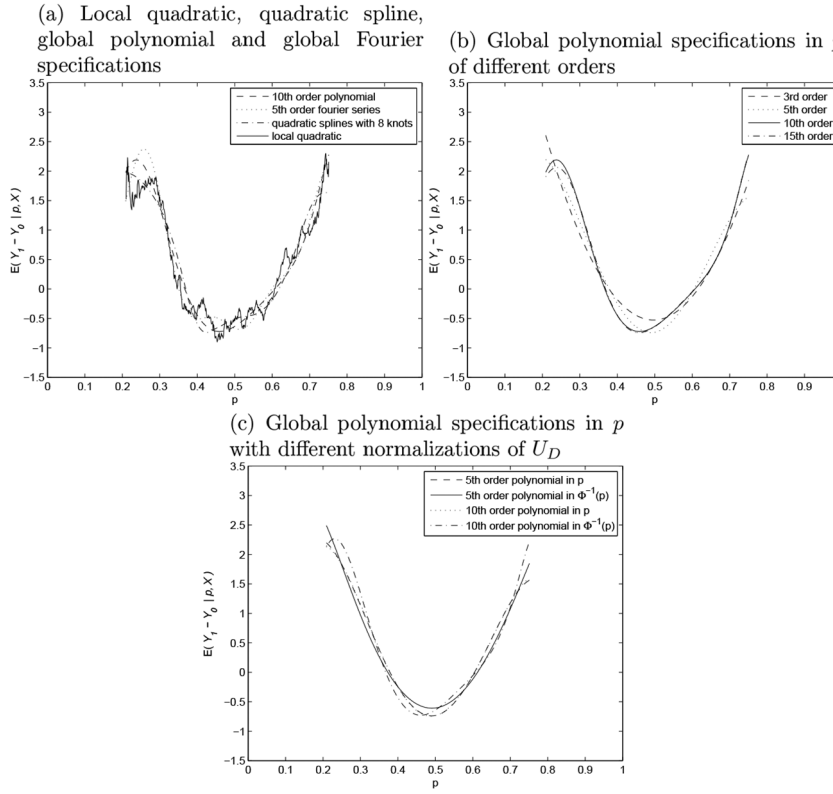

FIG. 8.—These graphs show specification checks for the MTE estimates with same-sex as the excluded instrument. These graphs display MTE estimates based on assumptions 1 and 2. We use same-sex as the only excluded instrument. The MTE estimates are evaluated at the mean values of the covariates. We construct $P(Z)$ using the parameter estimates from the logit model specified in the note to table 2. 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 three or more children rather than two children. A high value of p means that a family is less likely to have three or more children. In graph a, the MTE estimates in the solid line are based on a double residual regression separately for the treated and nontreated, using a local quadratic regression with uniform kernel and a bandwidth of 0.0615 as in figure 5. The three other MTE estimates are based on series specifications of the functions k_1 and k_0, with a tenth-order polynomial (dashed curve) and a fifth-order Fourier series (dotted curve) and a quadratic spline specification with eight knots (dashes and dots). In graph b, the MTE estimates come from specifications of the functions k_1 and k_0 as global polynomials in p of different orders. In graph c, the MTE estimates are based on global polynomial specifications of the functions k_1 and k_0 in p or $\Phi^{-1}(p)$.

Fourier series, and a quadratic spline with eight knots[22] (each specification uses 11 parameters for k_1 and k_0). The MTE estimates from these alternative estimators are very similar to the baseline results. In panel b, we show results from the global polynomial specification with different orders. It is reassuring to find that the MTE estimates are very similar for polynomial specifications with third- and higher-order terms. This suggests that the power series have a good approximation rate in our setting. In panel c, we specify k_0 and k_1 as polynomials in $\Phi^{-1}(p)$, providing a generalization of the normal selection model. This change to the empirical MTE model barely moves the estimates. Taken together, these specification checks suggest that the MTE estimates are robust to changes in the modeling framework for k_0 and k_1.

VI. Conclusions

The interpretation of IV estimates as LATE of instrument-induced shifts in treatment raises concerns about their external validity and policy relevance. One road to go down in the search for external validity and policy relevance is to estimate MTE, which have clear economic interpretation and summarize all conventional treatment parameters. However, full identification of the MTE requires some form of parametric or functional restriction when the instrument is discrete, as it often is in applied research.

In this paper, we showed how a discrete instrument can be used to identify the MTE under functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment. One key result is that this separate estimation approach 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 flexible MTE model under the additional assumption of additive separability between observed and unobserved heterogeneity in treatment effects.

We applied 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 found 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 rejected the external validity of the LATEs of family size at conventional significance levels. When comparing the MTE weights asso-

[22] We place these knots at equally spaced quantiles of $P(Z)$.

ciated with the IV estimates to the MTE weights associated with the ATE and the ATT, we found that the latter treatment parameters assign much more weight to positive MTEs. This explains why the ATE and the ATT of family size are sizable and positive, while the LATEs are smaller and sometimes negative. Taken together, our findings point to the importance of moving beyond LATE in situations with a discrete instrument.

Appendix A

Proofs

Proof of Proposition 1

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

$$E(Y|P(Z) = p) = \mu_0 + p(\mu_1 - \mu_0) + K(p; \theta), \tag{A1}$$

where θ is a vector of L parameters. Because k is linear in θ, it follows that K is linear in θ and $E(Y|P(Z) = p)$ is linear in (μ_0, μ_1, θ).

The linear equation system (A1) has N equations and $L + 2$ parameters (μ_0, μ_1, θ). Thus, if $L > N - 2$, the system is underdetermined and the model is not identified.

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

$$E(Y|P(Z) = p, D = j) = \mu_j + K_j(p; \theta_j), \quad j = 0, 1, \tag{A2}$$

where θ_j is a vector of L parameters. Because k_j is linear in θ, it follows that K_j is linear in θ and $E(Y|P(Z) = p, D = j)$ is linear in (μ_j, θ_j).

For each value of j, the linear equation system (A2) has N equations and $L + 1$ parameters (μ_j, θ_j). Thus, if $L > N - 1$, the system is underdetermined and the model is not identified. QED

Proof of Proposition 2

Suppose that assumptions 1 and 2 hold. Assume that X takes on M different values and Z_- takes on N values for each X, giving MN combinations $(P(Z), X)$, which we label $(p_1, x_1), \ldots, (p_{MN}, x_N)$.

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

$$E(Y|P(Z) = p, X = x) = \mu_0(x) + p[\mu_1(x) - \mu_0(x)] + K(p; \theta), \tag{A3}$$

where θ is a vector of L parameters. Because k is linear in θ, K is linear in θ and $E(Y|P(Z) = p, X = x)$ is linear in (μ_0, μ_1, θ).

The linear equation system (A3) has MN equations and $2M + L$ parameters (μ_0, μ_1, θ). Thus, if $L > M(N - 2)$, the system is underdetermined and the model is not identified.

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

$$E(Y|P(Z) = p, X = x, D = j) = \mu_j(x) + K_j(p), \quad j = 0, 1, \tag{A4}$$

where θ_j is a vector of L parameters. Because k_j is linear in θ, K_j is linear in θ and $E(Y|P(Z) = p, D = j)$ is linear in (μ_j, θ_j).

For each value of j, the linear equation system (A4) has MN equations and $M + L$ parameters (μ_j, θ_j). Thus, if $L > M(N - 1)$, the system is underdetermined and the model is not identified. QED

Appendix B

Tests of Constant MTE

Subsection III.B provided a test of constant MTE, assuming that k_1 and k_0 are linear functions of p. Under assumption 1, testing the null hypothesis of constant MTE versus the alternative hypothesis of nonconstant MTE is then equivalent to testing the null hypothesis $\Delta_0 = \Delta_1$ versus a two-sided alternative, where

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

Panel A of figure B1 illustrates the testable prediction for (Δ_0, Δ_1) under the null hypothesis of constant MTE. The dark (white) area gives the set of (Δ_0, Δ_1) that are (not) consistent with the null hypothesis of constant MTE. To implement the test, we regress Y on D, Z and their interaction and perform a two-sided t-test on the interaction coefficient. In the first row of table B1, we report the results of this test. We can reject the null hypothesis of constant MTE at the 1 percent significance level.

A limitation of this test is that it relies on k_1 and k_0 being specified as linear functions of p. As usual, the test is not valid if the model is misspecified (i.e., k_1 or k_0 is not a linear function of p). Misspecification can lead us to falsely reject the null hypothesis of constant MTE, and it can make us fail to correctly reject the null hypothesis of constant MTE.

We now consider two ways in which to relax the parametric assumption on k_0 and k_1. We first provide a test in which k_1 and k_0 are specified as monotonic quadratic functions of p. Next, we consider a test in which all we assume is that k_1 and k_0 are monotonic functions of p. In both cases, the null hypothesis of constant MTE provides testable restrictions against the alternative of nonconstant MTE. However, weaker parametric assumptions on k_0 and k_1 provide weaker testable restrictions on (Δ_0, Δ_1). Therefore, the choice of test involves a trade-off between power and sensitivity to the specification of k_0 and k_1.[23]

[23] The parametric assumption about k_1 and k_0 (linear, quadratic, or monotonic in p) does not in itself restrict the possible values of (Δ_0, Δ_1). To see this, consider the case in which k_1 and k_0 are linear functions of p with slope coefficients α_1 and α_0, respectively. Without further assumptions, (Δ_0, Δ_1) can take any values because there are no constraints on the difference between α_1 and α_0. This means that the specified structure on k_1 and k_0 does not, in itself, impose testable restrictions on (Δ_0, Δ_1). However, under the null hypothesis of constant MTE, $\alpha_1 = \alpha_0$, and, as a result, (Δ_0, Δ_1) must be on the 45-degree line of fig. B1. By comparison,

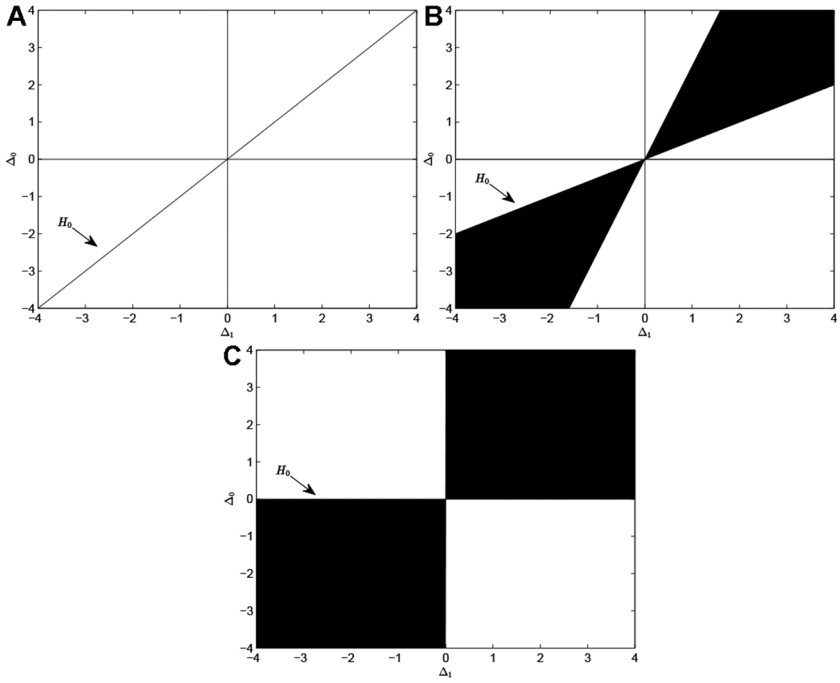

FIG. B1.—This figure illustrates the testable predictions for (Δ_0, Δ_1) under the null hypothesis of constant MTE for various specifications of the MTE model. The dark (white) area gives the set of (Δ_0, Δ_1) that are (not) consistent with the null hypothesis of constant MTE, given the specification of k_1 and k_0. In each graph, we assume that assumption 1 holds. Graph A assumes that k_0 and k_1 are linear in p. Graph B assumes that k_0 and k_1 are monotonic and quadratic functions of p. To ease the presentation, this graph also imposes that $p_0 + p_1 = 1$. Graph C assumes that k_0 and k_1 are monotonic functions of p.

Quadratic and Monotonic k_1 *and* k_0

For simplicity, consider the case with a binary instrument and no covariates. The following equations specify a quadratic MTE model:

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

and

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

where the constant terms ensure that the marginal expectations of U_1 and U_0 are zero. Under the null hypothesis of constant MTE, $\alpha_{01} = \alpha_{11}$ and $\alpha_{02} = \alpha_{12}$, implying that

under the alternative hypothesis of nonconstant MTE, $\alpha_1 \neq \alpha_0$, implying that (Δ_0, Δ_1) can be anywhere in fig. B1 except on the 45-degree line.

TABLE B1
TESTS OF NULL HYPOTHESIS OF CONSTANT MTE: TESTABLE PREDICTIONS AND p-VALUES

Modeling Framework	Testable Predictions of Constant MTE	p-Value
Linear k_1 and k_0	$\Delta_0 = \Delta_1$	.0001
Monotonic quadratic k_1 and k_0	$\frac{p_1+p_0}{1+p_1+p_0} \leq \frac{\Delta_1}{\Delta_0} \leq \frac{3-p_1-p_0}{2-p_1-p_0}$	.0032
Monotonic k_1 and k_0	Δ_0 and Δ_1 have same sign	.2404

$$\Delta_0 = \frac{1}{2}\alpha_{11}(p_1 - p_0) + \frac{1}{3}\alpha_{12}(p_1^2 - p_0^2) + \frac{1}{3}\alpha_{12}(p_1 - p_0),$$
$$\Delta_1 = \frac{1}{2}\alpha_{11}(p_1 - p_0) + \frac{1}{3}\alpha_{12}(p_1^2 - p_0^2).$$

Solving for α_{12} and α_{11} gives

$$\alpha_{12} = -3\frac{\Delta_1 - \Delta_0}{p_1 - p_0}$$

and

$$\alpha_{11} = \frac{2}{p_1 - p_0}\Delta_1 - \frac{2}{3}\alpha_{12}(p_1 + p_0).$$

Monotonicity of k_0 and k_1 in p on (0, 1) is equivalent to either $\alpha_{11} \geq 0 \cup \alpha_{11} + 2\alpha_{12} \geq 0$ or $\alpha_{11} \geq 0 \cup \alpha_{11} + 2\alpha_{12} \geq 0$. Simple calculations give

$$\alpha_{11} \gtreqless 0 \Leftrightarrow \Delta_1 \gtreqless \frac{p_1 + p_0}{1 + p_1 + p_0}\Delta_0 \tag{B1}$$

and

$$\alpha_{11} + 2\alpha_{12} \gtreqless 0 \Leftrightarrow \Delta_1 \lesseqgtr \frac{3 - p_1 - p_0}{2 - p_1 - p_0}\Delta_0. \tag{B2}$$

When k_1 and k_0 are specified as monotonic quadratic functions of p, constant MTE therefore requires that

$$\frac{p_1 + p_0}{1 + p_1 + p_0} \leq \frac{\Delta_1}{\Delta_0} \leq \frac{3 - p_1 - p_0}{2 - p_1 - p_0}.$$

Panel B of figure B1 illustrates the testable predictions for (Δ_0, Δ_1) under the null hypothesis of constant MTE. The dark (white) area gives the set of (Δ_0, Δ_1) that are (not) consistent with the null hypothesis of constant MTE. As expected, assuming quadratic and monotonic k_1 and k_0 (instead of linear) gives weaker testable restrictions on (Δ_0, Δ_1) from the null hypothesis of constant MTE.

Monotonic k_1 *and* k_0

Consider the case with a binary instrument and no covariates. If k_1 and k_0 are monotonic functions of p, so are K_1 and K_0. Under the null hypothesis of constant

MTE, monotonicity of K_1 and K_0 implies that Δ_1 and Δ_0 should have the same sign. If k_1 and k_0 are specified as monotonic functions of p, we can reject the null hypothesis of constant MTE if Δ_1 and Δ_0 have different signs.

Panel C of figure B1 illustrates the testable predictions for (Δ_0, Δ_1) under the null hypothesis of constant MTE. The dark (white) area gives the set of (Δ_0, Δ_1) that are (not) consistent with the null hypothesis of constant MTE. As expected, assuming monotonic k_1 and k_0 (instead of quadratic and monotonic) gives weaker testable restrictions on (Δ_0, Δ_1) from the null hypothesis of constant MTE.

Implementation of Tests

In the test with monotonic k_1 and k_0, the null hypothesis of constant MTE implies that (Δ_0, Δ_1) is an element in a subset of $\mathbb{R}^2$ that is the union of two convex subsets (see panel C of fig. B1). We proceed by first providing p-values for the tests of whether (Δ_0, Δ_1) is an element in each of the convex subsets. Finally, we construct a p-value for the joint test of the null hypothesis that (Δ_0, Δ_1) is an element in the union of the two convex subsets versus a complementary alternative hypothesis, by using the intersection-union method of Berger (1982).

Specifically, our procedure consists of three steps. First, we test the null hypothesis that (Δ_0, Δ_1) is an element in the lower-left quadrant of panel C of figure B1 versus a complementary alternative. This is a joint test of two inequality constraints. We produce a p-value for this test by testing each inequality constraint separately with one-sided t-tests and applying a Bonferroni correction to the result. The Bonferroni-corrected p-value for the test of the null hypothesis that both inequality constraints are satisfied versus a complementary alternative is two times the smallest of the p-values from the two t-tests. Second, we use the same procedure to compute a p-value for the test of the null hypothesis that (Δ_0, Δ_1) is an element in the upper-right quadrant of panel C of figure B1. Finally, we apply the intersection-union method to construct a p-value for the joint test of the null hypothesis that (Δ_0, Δ_1) is an element in the union of the two convex subsets versus the complementary alternative. Berger (1982) shows that the p-value for this joint test is the lower of the two p-values from the first two steps.[24]

The main advantages of this testing procedure are that it is simple to implement and that it also applies to dependent tests. We can therefore apply it directly to test for constant MTE when k_1 and k_0 are quadratic and monotonic. The main drawback of the procedure is that it is not efficient because of the conservative nature of Bonferroni corrections, so the reported p-values should be interpreted as upper bounds.

In the second row of table B1, we report the p-value of the test for constant MTE (against the alternative hypothesis of nonconstant MTE), assuming that k_0 and k_1 are quadratic and monotonic functions of p. Our findings suggest that we can re-

[24] Let p_1 and p_2 be p-values associated with each of these two tests. Under the null hypothesis, at least one of these p-values (the one associated with the set containing the true parameter values) will have a distribution that dominates a uniform distribution, such that $\Pr(p_i < q) \leq q$. The intersection-union method suggests using $\max(p_1, p_2)$ as the p-value of the full test. This is a valid p-value because $\Pr(\max(p_1, p_2) < q) \leq \Pr(p_i < q)$ for both $i = 1$ and $i = 2$, and since one of these is bounded above by q, we have $\Pr(\max(p_1, p_2) < q) \leq q$.

ject the null hypothesis of constant MTE at the 1 percent significance level given this specification of k_0 and k_1. In the third row of table B1, we report the p-value of the test for constant MTE (against the alternative hypothesis of nonconstant MTE), assuming that k_0 and k_1 are monotonic functions of p. Our findings suggest that we cannot reject the null hypothesis of constant MTE at conventional significance levels with such a flexible model specification.

Appendix C

An Approximation Result

In this appendix, we consider how the analysis is affected if the empirical specification of k_j is an approximating model. Consider an empirical model of $k_j(p)$ that is a polynomial in $F_{U_D}^{-1}(p)$ of order M_1. To estimate the parameters in $k_j(p)$, the separate estimation approach uses that

$$E(Y_j|P(Z) = p, D = j) = \mu_j + K_j(p), \quad j = 0, 1,$$

where

$$\begin{aligned} K_1(p) &= E(U_1|U_D < F_{U_D}^{-1}(p)) \\ &= \frac{1}{p}\int_0^p \sum_{i=0}^{M_1} \lambda_{1i} F_{U_D}^{-1}(u)^i du \\ &= \sum_{i=0}^{M_1} \lambda_{1i} \frac{1}{p}\int_0^p F_{U_D}^{-1}(u)^i du \end{aligned}$$

and

$$\begin{aligned} K_0(p) &= E(U_0|U_D > F_{U_D}^{-1}(p)) \\ &= \frac{1}{1-p}\int_p^1 \sum_{i=0}^{M_1} \lambda_{0i} F_{U_D}^{-1}(u)^i du \\ &= \sum_{i=0}^{M_1} \lambda_{0i} \frac{1}{1-p}\int_p^1 F_{U_D}^{-1}(u)^i du. \end{aligned}$$

Given a consistent estimate of $P(Z)$, the λ_1's can be estimated from the following ordinary least squares (OLS) regression on the sample of treated:

$$Y = \int_{i=0}^{M_1} \lambda_{1i} x_{1i} + \eta, \tag{C1}$$

where

$$x_{1i} = \frac{1}{p}\int_0^p F_{U_D}^{-1}(u)^i du,$$

and the λ_0's can be estimated from the following OLS regression on the sample of untreated:

$$Y = \int_{i=0}^{M_1} \lambda_{0i} x_{0i} + \eta, \tag{C2}$$

where

$$x_{0i} = \frac{1}{1-p} \int_p^1 F_{U_D}^{-1}(u)^i du.$$

Our interest is centered on how the analysis is affected by the empirical specification of $k_j(p)$ being an approximation as opposed to being exactly correct.[25] In particular, suppose that the true models of $k_j(p)$ are polynomials in $F_{U_D}^{-1}(p)$ of order $M_2 > M_1$. In this case, the OLS estimates of the λ_1's and λ_0's suffer from omitted variable bias. Let δ_i^j be a vector of the probability limits of the slope coefficients in the regression of x_{ji} on $[x_{j0}, ..., x_{jM_1}]$ for $i = M_1 + 1, ..., M_2$ and $j = 0, 1$. The probability limit of the estimator of λ_{ji} based on the approximating model is then

$$\tilde{\lambda}_{j,i} = \lambda_{j,i} + \sum_{k=M+1}^{M_2} \delta_i^j(k) \lambda_{j,k}$$

for $0 < i \leq M_1$. Define $\tilde{\lambda}_{j,i} = 0$ for $i > M_1$. The constant terms have the probability limits

$$\tilde{\lambda}_{j,0} = \lambda_{j,0} + \sum_{i=1}^{M_2} (\lambda_{j,i} - \tilde{\lambda}_{j,i}) \bar{x}_{ji}.$$

A polynomial MTE model may be represented by a vector $\Lambda = (\lambda_{0,0}, \lambda_{1,0}, \ldots, \lambda_{1,M_2})$. We measure the distance between two models by the standard Euclidean vector norm of order $2(M_2 + 1)$. The following proposition provides a limit statement of the distance between the true models and the probability limits of the empirical models of $k_j(p)$.

PROPOSITION. Let Λ^n be a sequence of polynomial MTE models of order M_2, indexed by n, that converges to a model of order $M_1 < M_2$ at the rate $O(n^{-1})$. Let $\tilde{\Lambda}^n$ be the sequence of probability limits of the estimator based on equations (C1) and (C2) when the data are generated by Λ^n. Then the difference between Λ^n and $\tilde{\Lambda}^n$ converges at the rate $O(n^{-1})$.

Proof. Note first that the vector norm converges at a rate $O(n^{-1})$ if and only if each element converges at the rate $O(n^{-1})$: Let $(\gamma_1, ..., \gamma_m)^n$ be a sequence of vectors with Euclidean norms γ^n. If there exist constants C and n_0 such that $\gamma^n \leq C/n$ for $n > n_0$, then also $|\gamma_i^n| \leq C/n$ for $i = 1, \ldots, m$, $n > n_0$. Conversely, if $|\gamma_i^n| \leq C/n$ for $i = 1, \ldots, m$, for $n > n_0$, then we also have $\gamma^n \leq \sqrt{\sum_{i=1}^m C_i^2}/n$.

The element-wise differences between the probability limits of the estimator of the approximating model and the true model are

$$\tilde{\lambda}_{j,i} - \lambda_{j,i} = \sum_{k=M+1}^{M_2} \delta_i^j(k) \lambda_{j,k,}$$

[25] Note that we cannot apply standard best linear approximation results for OLS, since our parameters of interest are the MTE and not the conditional expectation functions in the linear regressions.

for $i = 1, ..., M_1$,

$$\tilde{\lambda}_{j,0} - \lambda_{j,0} = \sum_{i=1}^{M_2} (\lambda_{j,i} - \tilde{\lambda}_{j,i}) \bar{x}_{ji},$$

and

$$\tilde{\lambda}_{j,i} - \lambda_{j,i} = -\lambda_{j,i}$$

for $i \in \{M_1 + 1, ..., M_2\}$. Clearly, $\tilde{\lambda}_{j,i} - \lambda_{j,i}$ are proportional to $\lambda_{j,k}$ for $k \in \{M_1 + 1, ..., M_2\}$. Since there exist constants n_0 and C_i, for $i \in \{M_1 + 1, ..., M_2\}$, such that $|\lambda_{j,i}| \leq C_i/n$ for $n > n_0$, there also exist constants D_i for $i = 0, ..., M_2$, with the D_i's being linear combinations of the C_i's such that $|\tilde{\lambda}_{j,i} - \lambda_{j,i}| \leq D_i/n$ for $n > n_0$. QED

Appendix D

Heterogeneity in the Effects of Family Size

This appendix describes 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., Björklund and Moffitt 1987). Although this normal selection model restricts the shape of the MTE model, 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 toward $\pm\infty$ as U_D tends toward zero or one (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 heterogeneity in the effects of family size if the population is segmented by preferences 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 when the population contains two or more distinct subpopulations. In figure D1, 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 three or more children be generated from 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 two. Individuals in the first subgroup have a constant MTE of one, while individuals in the second subgroup have a constant MTE of negative one. Although each subgroup has a constant MTE, figure D1 shows that the population distribution of MTE derived from this mixed model has a U shape. Individuals from the subgroup with high MTE are overrepresented in the tails, whereas individuals from the subgroup with a low MTE tend to be in the middle ranges of U_D. The reason is

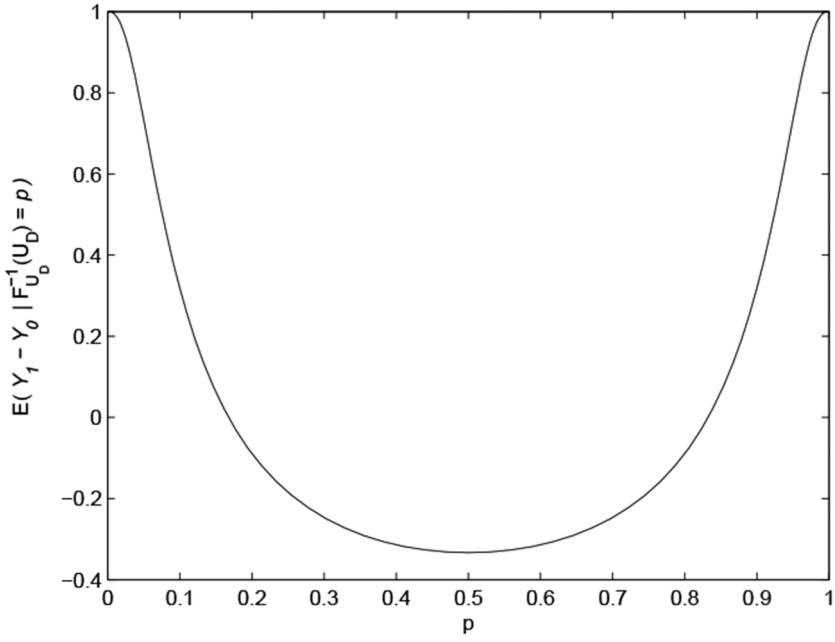

FIG. D1.—Example of MTE generated from a mixture model. This figure displays the population distribution of MTE, which is generated from a mixture of two normal selection models. The population consists of two equally sized subgroups: one with constant MTEs equal to one and the other with constant MTEs equal to negative one. In the selection equation, both groups enter treatment if a random variable exceeds a threshold of zero. The group with negative MTEs has random variables that are standard normal, while the group with positive MTEs has random variables that are normal with mean zero and variance two. The *y*-axis measures the value of the MTE in the population (i.e., the weighted average of the MTEs in the two subgroups), 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.

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.

Finally, 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 quantity-quality 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; Mogstad and Wiswall 2016). Also other theories, outside the Becker and Lewis model, suggest heterogeneity in the effects of family size on child outcome. In particular, for some families, additional siblings may benefit existing children if they stabilize the parental relationship (see, e.g., Becker 1998), if they make maternal employment less

likely (see, e.g., Ruhm 2008), or if there are positive spillover effects among siblings (see, e.g., Bandura 1977).

Appendix E

Bounds for ATE and ATT

This appendix presents nonparametric bounds on the ATE and the ATT of family size. We first construct worst-case bounds (see Manski 1989), exploiting that the dependent variable is bounded. In an attempt to tighten these bounds, we impose the standard IV assumptions of exclusion, monotonicity, and relevance.[26]

Worst-Case Bounds

The ATE can be written as

$$\begin{aligned} E(Y_1 - Y_0) &= E(Y_1 - Y_0|D = 1)\Pr(D = 1) + E(Y_1 - Y_0|D = 0)\Pr(D = 0) \\ &= [E(Y_1|D = 1) - E(Y_0|D = 1)]\Pr(D = 1) \\ &\quad + [E(Y_1|D = 0) - E(Y_0|D = 0)][1 - \Pr(D = 1)], \end{aligned}$$

while the ATT is defined as

$$E(Y_1 - Y_0|D = 1) = E(Y_1|D = 1) - E(Y_0|D = 1).$$

In data, we observe the empirical analogue of $\Pr(D = 1)$, $E(Y_1|D = 1)$, and $E(Y_0|D = 0)$, whereas $E(Y_0|D = 1)$ and $E(Y_1|D = 0)$ are unobserved. However, the dependent variable—years of schooling—is bounded between 6 and 21. This gives us the inequalities $6 \le E(Y_0|D = 1) \le 21$ and $6 \le E(Y_1|D = 0) \le 21$, allowing us to construct worst-case bounds. For the ATE, we obtain a lower bound by setting $E(Y_0|D = 1) = 21$ and $E(Y_1|D = 0) = 6$, whereas an upper bound is constructed by setting $E(Y_0|D = 1) = 6$ and $E(Y_1|D = 0) = 21$. For the ATT, we construct a lower bound by setting $E(Y_0|D = 1) = 21$ and an upper bound by setting $E(Y_0|D = 1) = 6$. The worst-case bounds are reported in the first row of table E1. They are not informative about the effects of family size. For example, the worst-case bounds on the ATE tell us that the effect of another sibling is somewhere between -7.7 and 7.3 years of schooling.

Bounds Based on Instrumental Variables

In an attempt to tighten the bounds, we impose the standard IV assumptions of exclusion, monotonicity, and relevance. Under these assumptions, we can use the sharp bounds provided in Heckman and Vytlacil (2001). Rows 2 and 3 of table E1

[26] Alternatively, one could invoke the monotone IV assumption (Manski and Pepper 2000). This assumption is a weakened form of the exclusion restriction. As a result, it will necessarily produce wider bounds than the standard IV assumptions. Indeed, the monotone IV assumption does little to tighten the worst-case bounds in our setting.

TABLE E1
NONPARAMETRIC BOUNDS FOR ATE AND ATT

	ATE Bounds		ATT Bounds	
	Lower	Upper	Lower	Upper
Worst-case bounds	−7.7	7.3	−8.9	6.1
Bounds with IV assumptions:				
Same-sex instrument	−7.2	6.9	−8.4	5.7
Twins instrument	−4.3	3.1	−8.8	6.0

NOTE.—This table provides point estimates of nonparametric bounds for the average treatment effect of family size and the average treatment effects on the treated of family size. The dependent variable is years of schooling.

present point estimates of these bounds. Unfortunately, they are too wide to be informative about the effects of family size.

References

Angrist, Joshua D., and Ivan Fernandez-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics: Tenth World Congress*, vol. 3, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 401–36. Cambridge: Cambridge Univ. Press.

Angrist, Joshua D., Guido W. Imbens, and Donald D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *J. American Statis. Assoc.* 91 (434): 444–72.

Angrist, Joshua D., Victor Lavy, and Analia Schlosser. 2010. "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *J. Labor Econ.* 28 (4): 773–824.

Bandura, Albert. 1977. *Social Learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Becker, Gary S. 1998. *A Treatise on the Family*. Enl. ed. Cambridge, MA: Harvard Univ. Press.

Becker, Gary S., and H. Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *J.P.E.* 81 (2): 279–88.

Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics* 24 (4): 295–300.

Björklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *Rev. Econ. and Statis.* 69 (1): 42–49.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2005. "The More the Merrier? The Effects of Family Size and Birth Order on Children's Education." *Q.J.E.* 120 (2): 669–700.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2012. "Beyond LATE with a Discrete Instrument." Discussion Paper no. 703, Statistics Norway, Oslo.

Caceres-Delpiano, Julio. 2006. "The Impacts of Family Size on Investment in Child Quality." *J. Human Resources* 41 (4): 738–54.

Carneiro, Pedro, James J. Heckman, and Edward Vytlacil. 2003. "Understanding What Instrumental Variables Estimate: Estimating the Average and Marginal Return to Schooling." Working paper, Univ. Chicago.

———. 2011. "Estimating Marginal Returns to Education." *A.E.R.* 101 (6): 2754–81.

Carneiro, Pedro, and Sokbae Lee. 2009. "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality." *J. Econometrics* 149 (2): 191–208.
Fang, Kai-Tai, Samuel Kotz, and Kai Wang Ng. 1990. *Symmetric Multivariate and Related Distributions*. London: Chapman & Hall.
French, Eric, and Jae Song. 2014. "The Effect of Disability Insurance Receipt on Labor Supply." *American Econ. J.: Econ. Policy* 6 (2): 291–337.
Hanushek, Eric A. 1992. "The Trade-Off between Child Quantity and Quality." *J.P.E.* 100 (1): 84–117.
Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *J. Econ. Literature* 48 (2): 356–98.
Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "How Details Make a Difference: Semiparametric Estimation of the Partially Linear Regression Model." Manuscript, Univ. Chicago.
Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labor Market Data*, edited by James J. Heckman and Burton Singer, 39–110. Cambridge: Cambridge Univ. Press.
Heckman, James J., and Daniel Schmierer. 2010. "Tests of Hypotheses Arising in the Correlated Random Coefficient Model." *Econ. Modelling* 27 (6): 1355–67.
Heckman, James J., Daniel Schmierer, and Sergio Urzua. 2010. "Testing the Correlated Random Coefficient Model." *J. Econometrics* 158 (2): 177–203.
Heckman, James J., and Sergio Urzua. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *J. Econometrics* 156 (1): 27–37.
Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Rev. Econ. and Statis.* 88 (3): 389–432.
Heckman, James J., and Edward Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proc. Nat. Acad. Sci.* 96 (8): 4730–34.
———. 2001. "Local Instrumental Variables." In *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics; Essays in Honor of Takeshi Amemiya*, edited by Cheng Hsiao, K. Morikune, and James L. Powell, 1–46. New York: Cambridge Univ. Press.
———. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.
———. 2007. "Econometric Evaluation of Social Programs, Part II." In *Handbook of Econometrics*, vol. 6, edited by James J. Heckman and Edward E. Leamer, 4875–5143. Amsterdam: Elsevier.
Ichimura, Hidehiko, and Petra E. Todd. 2007. "Implementing Nonparametric and Semiparametric Estimators." In *Handbook of Econometrics*, vol. 6, edited by James J. Heckman and Edward E. Leamer, 5369–5468. Amsterdam: Elsevier.
Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
Kirkebøen, Lars, Edwin Leuven, and Magne Mogstad. 2016. "Field of Study, Earnings, and Self-Selection." *Q.J.E.* Electronically published May 3.
Leeb, Hannes, and Benedikt M. Pötscher. 2005. "Model Selection and Inference: Facts and Fiction." *Econometric Theory* 21 (1): 21–59.
Levinsohn, James, and Amil Petrin. 2003. "Estimating Production Functions Using Inputs to Control for Unobservables." *Rev. Econ. Studies* 70 (2): 317–41.
Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *A.E.R.* 103 (5): 1797–1829.

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *J. Human Resources* 24 (3): 343–60.
———. 1997. "Monotone Treatment Response." *Econometrica* 65 (6): 1311–34.
Manski, Charles F., and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68 (4): 997–1010.
Marschak, Jacob, and William H. Andrews. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica* 12 (3–4): 143–205.
Moffitt, Robert. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Annales d'Economie et de Statistique* (91–92): 239–61.
Mogstad, Magne, and Matthew Wiswall. 2016. "Testing the Quantity-Quality Model of Fertility: Estimation Using Unrestricted Family Size Models." *Quantitative Econ.* 7 (1): 157–92.
Olley, G. Steven, and Ariel Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64 (6): 1263–97.
Olsen, Randall J. 1980. "A Least Squares Correction for Selectivity Bias." *Econometrica* 48 (7): 1815–20.
Robinson, P. M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* 56 (4): 931–54.
Rosenzweig, Mark R., and Kenneth I. Wolpin. 1980. "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica* 48 (1): 227–40.
Ruhm, Christopher J. 2008. "Maternal Employment and Adolescent Development." *Labour Econ.* 15 (5): 958–83.
Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1): 331–41.
White, Halbert. 1980. "Using Least Squares to Approximate Unknown Regression Functions." *Internat. Econ. Rev.* 21 (2): 149–70.