Magne Mogstad and Matthew Wiswall

How Linear Models Can Mask Non-Linear Causal Relationships
An Application to Family Size and Children's Education

Abstract:
Many empirical studies specify outcomes as a linear function of endogenous regressors when conducting instrumental variable (IV) estimation. We show that commonly used tests for treatment effects, selection bias, and treatment effect heterogeneity are biased if the true relationship is non-linear. In particular, using linear models can only lead to under-rejection of the null hypothesis of no treatment effects. In light of these results, we re-examine the recent evidence suggesting that family size has no causal effect on children's education. Following common practice, a linear IV estimator has been used, assuming constant marginal effects of additional children across family sizes. We show that the conclusion of no causal effect of family size is an artifact of the specification of a linear model, which masks significant marginal family size effects. Estimating a model that is non-parametric in family size, we find that family size matters substantially for children's educational attainment, but in a non-monotonic way. Our findings illustrate that IV estimation of models which relax linearity restrictions is an important addition to empirical research, particularly when OLS estimation and theory suggests the possibility of non-linear causal effects.

Address: Magne Mogstad, Statistics Norway, Research Department.
E-mail: magne.mogstad@ssb.no

Matthew Wiswall, New York University, Department of Economics,
Email: mwiswall@nyu.edu

1   Introduction

The linear IV estimator, in which an endogenous outcome is a linear function of a potentially endogenous regressor, is a major workhorse in applied economics. When an included regressor takes on multiple values, so called "variable treatment intensity," the linear specification restricts the marginal effects of this regressor to be constant across all margins. There are at least three reasons for this preference for linear specifications in applied research. First, parsimonious linear specifications may in some cases provide a reasonable approximation to a potentially non-linear relationship. Second, there may be insufficient instrumental variation to instrument for multiple endogenous variables. And, third, since many IV estimations suffer from imprecision due to weak instruments, restricting the number of endogenous regressors that need to be instrumented for can improve precision.

Our paper examines the implications for inference from using OLS and IV estimators that assume a linear relationship between the outcome and a possibly endogenous regressors when the true relationship is non-linear. Previous research has shown that the linear OLS and IV estimands can be decomposed into weighted averages of specific marginal effects, where the OLS and IV weights are in general different (Angrist and Imbens, 1995; Yitzhaki, 1996; Angrist and Krueger, 1999; Heckman, Urzua, Vytlacil, 2006). Using these representations, we show that commonly used tests for non-zero causal effects, selection bias, and treatment effect heterogeneity are biased if the true relationship is non-linear. Moreover, we characterize each test's biases in terms of the conditions under which the linear test statistic leads to over- or under-rejection of specific null hypotheses.

We first demonstrate that non-linearities can lead to a conclusion of a zero causal effect using a linear model even if some or all marginal effects are non-zero. Next, we show that because OLS and IV have generally different weights on the marginal effects, non-linearities can also lead to a conclusion of selection bias even when the regressor is exogenous, or lead to a conclusion of no selection bias even when the regressor is endogenous. Emphasizing the distinction between these issues from other commonly cited issues with IV estimation, we show that these results are solely due to the non-linearity of the treatment effects, and not to heterogeneity in the marginal treatment effects or weak instruments. Finally, we demonstrate that comparing linear IV estimators using different instruments to make inferences about treatment effect heterogeneity can be misleading, as non-linearities can mask treatment effect heterogeneity, or lead to an erroneous conclusion that treatment effects are heterogeneous when they are in fact homogeneous.

We demonstrate the empirical relevance of these issues by re-examining the large body of empirical research that estimates the relationship between family size and children's education and tests the Becker and Lewis (1973) quantity-quality model. In our empirical work, we show that the commonly used linear family size model yields erroneous conclusions about the causal effects of family size on children's education. Recent research from several developed countries,

using large data sets, extensive controls for confounding characteristics such as birth order, and instruments for family size, suggest that family size has no causal effect on child education. For instance, a widely cited study by Black, Devereux, and Salvanes (2005, hereafter BDS) based on Norwegian data, concludes that "there is little if any family size effect on child education; this is true when we estimate the relationship with controls for birth order or instrument family size with twin births" (p. 697). Using data from the US, Caceres-Delpiano (2006) comes to a similar conclusion. Other recent studies reporting no effect of family size include Angrist et al. (2006) using data from Israel and Aaslund and Grønqvist (2007) using data from Sweden.

This recent evidence for no causal family size effect is based on a model that is linear in family size, assuming constant marginal effects of additional siblings across family sizes. The most common hypothesis is that additional children reduce parental resources provided to existing children, creating a quantity-quality tradeoff. However, from existing theory and evidence on the influences on child development, there are reasons to suspect a non-linear relationship between family size and children's outcome. Although Becker and Lewis (1973) suggest a quantity-quality tradeoff in family size, it is not necessarily the case that the marginal family size effects are equal at all margins. Indeed, as Rosenzweig and Wolpin (1980) point out, the quantity-quality model allows for strong complementarities between numbers and quality of children. If this is the case, an exogenous increase in family size may actually improve children's outcomes as parents increase their demand for child quality. Additional siblings might also benefit existing children if they stabilize parental relationships,[1] make maternal employment less likely,[2] or there are positive spillover effects among siblings.[3] Overall, the potential negative and positive effects of family size may be relatively stronger or weaker as family size increases, implying that the relationship between family size and children's outcomes could be non-linear and even non-monotonic.

To re-examine the relationship between family size and children's education, we use the same data source as BDS (2005), administrative registers for the entire population of Norway. In addition, we follow BDS (2005) in using twin births as the instruments for family size and in making the same parametric assumptions on an identical set of controls. Our point of departure is to relax the assumption of constant marginal family size, estimating models that are non-parametric in family size.

As in previous studies, our OLS estimate of the linear model indicate an almost zero effect of family size on children's education, after controlling for birth order and other covariates.

---

[1]Becker (1998) argues that children represent a couple-specific investment, implying that the value to the spouses of having children is not fully preserved outside the current relationship. Accordingly, children increase the expected gain from the relationship, discouraging dissolution, which may harm the children. Vuri (2003) and Svarer and Verner (2008) find no stabilizing effect of children on relationships.

[2]Reduction in maternal employment is likely to increase the total time the mother devotes to child care. Empirical results are mixed: some find that maternal employment is detrimental for children (see e.g. Bernal, 2008; Ruhm, 2008), others that it is beneficial (Vandell and Ramanan, 1992).

[3]The impact of interactions among siblings has not received much attention in economics, but is widely studied in the behavioral genetics literature (see e.g. Garcia et al. (2000) for a review) and the development psychology literature, following Bandura's (1977) social learning theory.

However, when estimating a non-parametric model in family size, we find a non-monotonic relationship with statistically significant and sizable marginal effects. Figure 6 provides a sense of how poor of an approximation the linear model is to the empirical relationship. This Figure plots the OLS predicted average education for first born children by their numbers of siblings, ranging from 0 siblings (1 child families) to 5 siblings (6 child families). While the linear OLS estimate yields nearly a flat line and zero family size effect, the non-parametric OLS estimates display an inverse U-shaped pattern with significant marginal family size effects. This evidence of non-linearities in the OLS estimates raises serious doubts of the appropriateness of linear models for the IV estimation.

Like previous studies, our IV results from the linear model using twin births as instruments show a small and imprecisely estimated effect of family size. However, when we estimate a non-parametric model in family size using twins at different birth parities to form multiple instruments, the IV estimates are not significantly different from the sizable OLS estimates. We also employ a number of alternative IV strategies, constructing predicted fertility instruments that are strongly correlated with particular family size margins. Applying the predicted fertility instruments to the non-parametric model in family size, we find large and statistically significant family size effects. For first born children, the causal relationship between their family size and education is clearly non-monotonic. While a third child added to a 2 child family increases the educational attainment of first born children, additional children have a negative marginal effect. The negative effect of family size at higher parities exceeds the birth order effects that BDS (2005) and others have emphasized as large.

To understand why the linear model yields a misleading picture of the relationship between family size and children's education, we estimate the weights attached to the marginal family size effects for the linear OLS and IV estimators. The linear OLS estimator reflects all marginal family size effects and weights them according to the sample distribution of family size, assigning the most weight to marginal effects close to the sample median family size. In comparison, the linear IV estimator only captures the marginal effects at the part of the support shifted by the specific instrument chosen. For example, using twins at second birth as the instrument weights the marginal effect of moving from 2 to 3 children most heavily, assigning far less weight to marginal effects at higher parities. The reasons for the almost zero effect of family size in both OLS and IV estimation of the linear model are that (i) marginal effects at different parities offset each other, and that (ii) the relatively small marginal effects are weighted heavily. Importantly, OLS and IV estimates of the linear family size model assign substantially different weights to the underlying marginal family size effects. Drawing conclusions about the endogeneity of family size by comparing linear OLS and IV estimates, as in previous studies, may therefore be unwarranted. By an analogous argument, comparing linear IV estimates using two different instruments may provide misleading inferences about treatment effect heterogeneity, since different instruments generally assign different weights to the marginal effects.

The general lesson to be drawn is that empirical researchers should be cautious about causal

inference in linear IV estimation, as the linear IV estimator captures only the marginal effects at the part of the support shifted by the specific instrument chosen. While this general point has been made in the theoretical contributions of Angrist and Imbens (1995) and Heckman et al. (2006), the specific forms of misspecification bias that can arise from non-linearities has not been fully appreciated. The fact that these biases can occur even with homogeneous marginal effects and strong instruments underscores their importance.

Section 2 summarizes the result that linear OLS and IV estimands can be expressed as weighted averages of the potentially endogenous regressor. Section 3 discusses how inference can be misleading if we use linear OLS and IV estimators when the true relationship is non-linear. Section 4 describes our data, and Section 5 compares OLS estimates of the linear and non-parametric models in family size. Section 6 presents IV results from the linear family size model, and Section 7 provides IV estimates of the non-parametric model in family size. Section 8 summarizes and concludes with a discussion of policy implications.

2   OLS and IV Estimation

In this section, we draw on previous work to show how the OLS and IV estimands of a linear model can be decomposed into weighted averages of specific marginal effects. To conform to the empirical analysis, we consider a model of family size and children's outcomes, although the results in Section 2 and 3 hold more generally. To focus on the implications of the linearity assumption, we ignore control variables, but include them in the empirical analysis.

2.1   Potential Outcomes and Marginal Effects

Let s_i denote the number of siblings of individual i: $s_i \in \{0, 1, \ldots, \bar{s}\}$, with $\bar{s}$ finite. We call $f_i(s)$ the effect or potential outcome of having s siblings on individual i. When convenient, we also refer to the effect of "family size", defined as the total number of children in the family: $s_i + 1$. In the most general case, this potential outcome framework allows for outcomes to differ across family size levels for the same individual, $f_i(s) \neq f_i(s')$ for $s \neq s'$, and across individuals with the same family size, $f_i(s) \neq f_{i'}(s)$ for $i \neq i'$. We refer to the first type of heterogeneity as variable treatment intensity and the latter type as heterogeneous treatment effects. For convenience and without loss of generality, we decompose the potential outcomes into components reflecting the mean and the deviation from the mean: $f_i(s) = \mu_s + u_{si}$, with $E[u_{si}] = 0$ for all s.

Using dummy variables constructed as $d_{si} = 1\{s_i \geq s\}$, we can express the observed outcome y_i as

$$y_i = \mu_0 + (\mu_1 - \mu_0 + u_{1i} - u_{0i})d_{1i} + \cdots + (\mu_{\bar{s}} - \mu_{\bar{s}-1} + u_{\bar{s}i} - u_{\bar{s}-1,i})d_{\bar{s}i} + u_{0i}. \qquad (1)$$

This model is non-parametric in family size, as we use dummy variables that fully saturate the support of the family size treatment s_i. Adopting conventional regression notation, we can express (1) as

$$y_i = \mu_0 + \gamma_{1i}d_{1i} + \cdots + \gamma_{\bar{s}i}d_{\bar{s}i} + \epsilon_i, \tag{2}$$

where $\gamma_{si} = \mu_s - \mu_{s-1} + u_{si} - u_{s-1,i}$ and $\epsilon_i = u_{0i}$. The marginal effect on individual i's outcome from being born to a family with s siblings rather than $s-1$ siblings is $f_i(s) - f_i(s-1) = \gamma_{si}$. In the general case, γ_{si} is an heterogeneous individual marginal treatment effect, which we can represent in (2) as random coefficients on the d_{si} family size indicators. Without loss of generality, these individual marginal effects can be decomposed as

$$\gamma_{si} = \gamma_s + \phi_{si},$$

with $\gamma_s = \mu_s - \mu_{s-1}$ and $\phi_{si} = u_{si} - u_{s-1,i}$. As the number of siblings ranges from $0, 1, \ldots \bar{s}$, there are $\bar{s}$ distinct marginal effects for each individual. Given the normalization $E[u_{si}] = 0$ for all s, γ_s is the *marginal* average treatment effect (ATE) of increasing siblings from $s-1$ to s, whereas ϕ_{si} represents heterogeneity in the marginal effect of family size.

2.2 Treatment Effect Heterogeneity and Non-Constant Marginal Effects

Consider three types of restrictions that could be imposed on the general model (2) and the individual marginal effects γ_{si}. First, restricting the marginal family size effects to be homogeneous but allowing them to be non-constant across family sizes restricts the individual marginal effects to $\gamma_{si} = \gamma_s$ for all s and i. In regression notation, this restriction yields a non-parametric model in family size with constant coefficients:

$$y_i = \mu_0 + \gamma_1 d_{1i} + \cdots + \gamma_{\bar{s}} d_{\bar{s}i} + \epsilon_i. \tag{3}$$

Allowing for random coefficients but assuming marginal effects are independent of family size imposes $\gamma_{si} = \beta_i$ for all i and s. Imposing this restriction on (2) yields a regression equation with a random linear slope:

$$y_i = \mu_0 + \beta_i s_i + \epsilon_i. \tag{4}$$

Finally, assuming both homogeneous and constant marginal family size effects restrict the individual marginal effects to be $\gamma_{si} = \beta$ for all i and s. Imposing this restriction on (2) yields the typical linear regression model with constant intercept and slope:

$$y_i = \mu_0 + \beta s_i + \epsilon_i. \tag{5}$$

2.3 OLS Estimation

We briefly review the relationship between OLS estimation of the two constant coefficients models: the linear family size model (5) and the non-parametric model in family size (3).

The OLS estimand for β in (5) is $\beta(OLS) \equiv Cov(y_i, s_i)/Var(s_i)$. The OLS estimands for γ_s marginal effects in (3) are

$$\gamma_s(OLS) \equiv E[y_i|s_i = s] - E[y_i|s_i = s - 1].^4$$

Angrist and Krueger (1999), drawing on results from Yitzhaki (1996), show that the OLS estimand for the linear model (5) is a weighted average of the OLS estimands of the marginal effects from model (3) which is non-parametric in the regressor:

$$\beta(OLS) = \sum_{s=1}^{\bar{s}} \gamma_s(OLS)w_s(OLS), \tag{6}$$

where the linear OLS weights are

$$w_s(OLS) = \frac{q_s}{\sum_{k=1}^{\bar{s}} q_k},$$

with

$$q_s = (E[s_i|s_i \geq s] - E[s_i|s_i < s])(pr(d_{si} = 1)(1 - pr(d_{si} = 1)).$$

The linear OLS weights $w_s(OLS)$ are non-zero and sum to 1. If the marginal effects are non-constant ($\gamma_s(OLS) \neq \gamma_{s'}(OLS)$ for $s \neq s'$), then it follows from (6) that the OLS estimator for the linear model depends on the sample distribution of family size. In particular, marginal effects that are close to the sample median family size receive the most weight in the linear OLS estimator. Depending on the weights, the linear OLS estimand will range between the maximum and minimum $\gamma_s(OLS)$.[5]

2.4 IV Estimation

Much of the recent discussion concerning IV estimation focuses on interpretation of the linear IV estimator in the presence of heterogeneous treatment effects and variable treatment intensity (see e.g. Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist et al., 2000; Heckman et al., 2006; Moffitt, 2008). To examine these issues, consider the case of a single binary

---

[4]Substituting the outcome equation from (5), we have $\gamma_s(OLS) = \gamma_s + E[\epsilon_i|s_i = s] - E[\epsilon_i|s_i = s - 1]$. This expression indicates that OLS estimation of the marginal effects identifies the sum of the marginal ATE, γ_s, and a selection bias term, $E[\epsilon_i|s_i = s] - E[\epsilon_i|s_i = s - 1]$.

[5]It is important to note that this weighting decomposition not only holds for the probability limit estimand of the linear OLS estimator, but also holds for the estimator itself. The sample analog linear OLS estimator $\hat{\beta}(OLS)$ has this exact weighting: $\hat{\beta}(OLS) = \sum_{s=1}^{\bar{s}} \hat{w}_s(OLS)\hat{\gamma}_s(OLS)$, where $\hat{\beta}(OLS)$, $\hat{w}_s(OLS)$, and $\hat{\gamma}_s(OLS)$ are the sample analog estimators.

instrument $z_i \in \{0, 1\}$. Define $S_i(q)$ as the number of siblings if child i is exposed to $z_i = q$. Following Angrist and Imbens (1995), we make the following assumptions:

A1 (Independence): $\{S_i(1), S_i(0), f_i(0), f_i(1), \dots, f_i(\bar{s})\} \perp z_i$.

A2 (First Stage): $pr(S_i(1) - S_i(0)) \neq 0$.

A3 (Monotonicity): $S_i(1) \geq S_i(0)$ for all i.

These assumptions imply that the instrument is independent of potential outcomes and of potential treatment assignments, has some effect on family size (analogous to the usual rank condition for identification), and affects everyone in the same way, if at all.

The linear IV estimator uses z_i to instrument for siblings. Angrist and Imbens (1995) show that the IV estimand for β in (5), $\beta(z)$, can be decomposed into a weighted average:

$$\beta(z) \equiv \frac{E[y_i|z_i = 1] - E[y_i|z_i = 0]}{E[s_i|z_i = 1] - E[s_i|z_i = 0]} = \sum_{s=1}^{\bar{s}} w_s(z)\gamma_s(z), \tag{7}$$

where $\gamma_s(z) = E[\gamma_{si}|Q_{si}(z)]$, $w_s(z) = \frac{pr(Q_{si}(z))}{\sum_{k=1}^{\bar{s}} pr(Q_{ki})}$, and $Q_{si}(z)$ is compact notation for the event $S_i(1) \geq s > S_i(0)$.

Adapting the local average treatment effect (LATE) terminology of Angrist and Imbens (1994), we call $\gamma_s(z)$ the sth *marginal* LATE. The marginal LATEs are the marginal ATEs for the particular complier group whose treatment status is shifted by the instrument z (that is, all i such that $S_i(1) \geq s > S_i(0)$).

The IV weights $w_s(z)$ are non-negative and sum to one. We denote the weights $w_s(z)$ and the IV estimand $\beta(z)$ as a function of the particular instrument z_i in order to emphasize that other instruments can lead to different weights and different estimands. Depending on the weights, the linear IV estimand will range between the maximum and minimum $\gamma_s(z)$. The IV weights can be directly estimated using the sample analog of the expression above.

Angrist and Imbens (1995) label the linear IV estimand (7) the *average causal response*. As Angrist and Imbens (1995) point out, there are two types of averaging underlying the average causal response. First, there is averaging over individuals indexed i. Only the individuals whose family size is affected by the instrument are included in the complier group and the average causal response. Second, there is averaging across marginal LATEs. The weights $w_s(z)$ place more weight on the marginal effects where the cumulative distribution function of family size is more affected by the particular instrument. An important feature of the linear IV estimand is that some of the IV weights $w_s(z)$ on the marginal effects can be zero if the chosen instrument z_i does not shift family size at this margin. In comparison, the OLS estimand places positive weight on every marginal effect in the empirical support of sample.[6]

---

[6]Heckman, Urzua, Vytlacil (2006) provide a detailed analysis of what various instruments identify when there is heterogeneity in treatment effects and variable treatment intensity. In their terminology, the twin birth instruments

3 Inference Using Linear Estimators

This section uses the above expressions to discuss how inference can be misleading if we use linear OLS and IV estimators when the true relationship is non-linear. We consider a number of commonly used tests, including tests for i) treatment effects, ii) selection bias, and iii) heterogeneous treatment effects. For each test, we formulate the implicit null hypothesis in terms of the marginal effects, and discuss the conditions under which using linear test statistics leads to under- or over-rejection of the null hypothesis. To illustrate the consequences of non-linearities for inference using misspecified linear estimators, we use a Monte Carlo study.

3.1 Simulation Example

Our Monte Carlo study is based on a simple example where the treatment takes on 3 values: $s_i \in \{0, 1, 2\}$. Potential outcomes $f_i(s)$ are specified as: $f_i(0) = \epsilon_i$, $f_i(1) = \gamma_{1i} + \epsilon_i$, $f_i(2) = \gamma_{1i} + \gamma_{2i} + \epsilon_i$. The marginal treatment effects are then $f_i(1) - f_i(0) = \gamma_{1i}$ for the 0 to 1 sibling margin, and $f_i(2) - f_i(1) = \gamma_{2i}$ for the 1 to 2 sibling margin. In regression notation, the observed outcome is then

$$y_i = \gamma_{1i} d_{1i} + \gamma_{2i} d_{2i} + \epsilon_i,$$

where $d_{1i} = 1\{s_i \geq 1\}$, and $d_{2i} = 1\{s_i \geq 2\}$. There are two binary instruments, $z_{1i} \in \{0, 1\}$ and $z_{2i} \in \{0, 1\}$, which are constructed to satisfy Assumptions A1-A3. The Simulation Appendix provides additional details on the data generating process.

Figure 1 presents a simulation using 500 replications of 5,000 observations from the data generating process in which we construct the treatment to be exogenous of the potential outcomes, $d_{si} \perp f_i(0), f_i(1), f_i(2)$ for $s = 1, 2$. In addition, the potential outcomes are homogeneous, $\gamma_{si} = \gamma_s$ for all i and s. We simulate three different linear estimators:

1) Linear OLS:

$$\beta(OLS) = Cov(y_i, s_i)/Var(s_i).$$

2) Linear IV using z_{1i} as the instrument:

$$\beta(z_1) = Cov(y_i, z_{1i})/Cov(z_{1i}, s_i).$$

---

are "transition specific" instruments that affect a specific family size margin (e.g. from 0 to 1 sibling), with the caveat that twin birth instruments may affect fertility at higher parities. An important contribution of Heckman et al. (2006) is that they provide the instrument-specific weights on the heterogeneous marginal treatment effects (MTE). It should be noted, however, that the term "marginal" in their MTE context refers to the effect of the treatment for heterogeneous individuals who are at specific utility margins, rather than treatment margins (e.g. moving from 0 to 1 sibling) as in our context. For our purposes, Angrist and Imbens' (1995) decomposition of the linear IV estimator in terms of marginal LATEs is convenient. Neither Heckman et al. (2006) nor Angrist and Imbens (1995) examine the consequences of non-linear treatment effects for commonly used tests for zero-treatment effects, selection bias, and treatment effect heterogeneity.

3) Linear IV using z_{2i} as the instrument:

$$\beta(z_2) = Cov(y_i, z_{2i})/Cov(z_{2i}, s_i).$$

In Figure 1, we plot the average value of the linear estimates across the 500 replications of the data generating process.[7] For each replication, we set $\gamma_1 = 1$ and vary the other marginal effect γ_2. As we vary γ_2 away from 1, we increase the degree of non-linearity. At $\gamma_1 = \gamma_2 = 1$, the marginal effects are constant and the linear model is correctly specified. At this point, the three estimators intersect and produce the same estimate (modulo sampling error): $\beta(OLS) = \beta(z_1) = \beta(z_2) = 1$. Recall that in this simulation the treatment is exogenous, and hence the OLS estimators are consistent.

Figure 1 shows that as we move γ_2 away from γ_1 and introduce non-linearity in the treatment effects, the three estimators diverge from each other. The reason is that each of the estimators weights the marginal effects differently. Table A-1 reported in the Simulation Appendix provides the weights on the marginal effects. The linear OLS estimator places approximately 46 percent weight on the γ_1 marginal effect and the 54 percent weight on the γ_2 marginal effect. In contrast, the two IV estimators have substantially different weighting of the marginal effects, reflecting the strength of the instruments on each treatment margin. In our simulation, z_{1i} is constructed to affect mainly the first treatment margin, while instrument z_{2i} affects exclusively the second margin. Given these particular instruments, the linear IV estimator using instrument z_{1i} places 2/3 weight on γ_1 and 1/3 weight on γ_2. In comparison, the linear IV estimator using instrument z_{2i} places 0 weight on γ_1 and all weight on γ_2. For this reason $\beta(z_1)$ is the flattest line in Figure 1 as this linear estimator has relatively little weight placed on γ_2, while $\beta(z_2)$ is a much steeper line as this linear estimator weighs γ_2 more heavily.

3.2 Testing for Treatment Effects

Consider testing for whether family size affects children's outcomes using IV estimation. The relevant null hypothesis is that each of the marginal LATEs identified by the instrument is zero: $\gamma_s(z) = 0$ for all s. Suppose we follow the previous literature in using the linear IV estimator $\beta(z)$ from (7) to test this null hypothesis.

There are two cases. In the first case, the null hypothesis is false and at least one of the marginal LATEs are non-zero: $\gamma_s(z) \neq 0$ for some s. Under-rejection of the false null hypothesis occurs when $\beta(z) = 0$ and we fail to reject the false null hypothesis of no treatment effect. One possibility for under-rejection is when the linear IV estimands of the marginal effects are

---

[7]Since each of these linear estimators have the form $\beta = \sum_s w_s \gamma_s$, we can write them as a function of a particular marginal effect γ_j:

$$\beta(\gamma_j) = \sum_{s \neq j} w_s \gamma_s + w_j \gamma_j.$$

The intercept is $\sum_{s \neq j} w_s \gamma_s$ and slope is w_j. In the case of the 1-2 sibling margin, this line is given by $\beta(\gamma_2) = w_1 \gamma_1 + w_2 \gamma_2$.

non-monotonic and cancel each other out. In fact, the linear IV estimate can suggests no family size effect even if *all* marginal LATEs are non-zero. For example, if the marginal LATE of moving from 0 to 1 siblings is positive and the marginal LATEs at higher parities are negative, then $\beta(z) = 0$ if $w_1(z)\gamma_1(z) = -\sum_{s=2}^{\bar{s}} w_s(z)\gamma_s(z)$. Figure 1 presents an example of this case. For the linear IV estimator using z_1 as the instrument, $\beta(z_1) = 0$ when $\gamma_2 = -2$. In this case, the linear estimator is zero, even though both marginal treatment effects are non-zero.

Another possibility for under-rejection occurs when the IV weights associated with non-zero marginal LATEs are zero. Since the linear IV estimand captures only the marginal effects at the part of the support shifted by the specific instrument chosen, it is possible that the range of variation in family size induced by the instrument has no effect on children's education, when in fact there are non-zero marginal effects outside the support of the instrument. As an example, assume that the instrument z only shifts family size in one part of the support of the family size distribution, say, from 0 to 1 siblings so that $pr(Q_{1i}(z)) = 1$ and $pr(Q_{si}(z)) = 0$ for $s > 1$. Then, the linear IV estimand is equal to the marginal LATE at this point, $\beta(z) = \gamma_1(z)$. If the marginal LATE is zero at this margin, $\gamma_1(z) = 0$, yet non-zero at other margins such that $\gamma_s(z) \neq 0$ for some $s > 1$, then the linear IV estimator under-rejects the false null hypothesis of no effect of family size. Returning to the example in Figure 1, the instrument z_2 is constructed to affect only the second margin γ_2. When $\gamma_2 = 0$, the linear IV estimator using z_2 indicates a zero effect of family size, despite the non-zero marginal effect of going from 0 to 1 siblings.

In the second case, the null hypothesis is true and over-rejection occurs when $\beta(z) \neq 0$. Because the linear IV estimand is a weighted average of the underlying marginal LATEs with non-negative weights, as shown in (7), over-rejection of the true null hypothesis is not possible. Under the true null hypothesis, the linear IV estimator correctly imposes the constant marginal treatment effects restriction. As a consequence, the bias in using linear IV estimators when the marginal effects are non-constant is one-sided, and can only lead to under-rejection of the null hypothesis of no treatment effect.

By the same token, linear OLS estimation can only lead to under-rejection of the null-hypothesis of no treatment effects when the marginal effects are non-constant. Under-rejection occurs when the OLS estimands of the marginal effects are non-monotonic and offset each other. Over-rejection is not possible, since the linear OLS estimand is a weighted average of the underlying marginal OLS estimands with positive weight at each margin, as shown in (6).

3.3 Testing for Selection Bias

Following Hausman (1978), a standard test of selection bias is to compare the linear OLS and IV estimates. The idea is that if family size is exogenous, the OLS and IV estimates would differ only by sampling error. For example, Caceres-Delpiano's (2006) study of family size effects and children's outcomes concludes that "the two-stage least-squares estimates are statistically distinguishable from OLS estimates, indicating an omitted variables bias in the single equation

model" (p. 738). In the general case of treatment effect heterogeneity, the linear OLS and IV estimators can differ even if there is no selection bias because the estimators capture the responses of different sub-groups (Heckman and Vytlacil, 2006). As shown below, however, even with homogeneous treatment effects the Hausman test can be misleading if the marginal effects are non-constant.

Assume homogeneous treatment effects ($\gamma_{si} = \gamma_s$ for all i and s), and consider testing the null hypothesis that family size is exogenous, which we can write as the OLS estimator for each marginal effect being equal to the marginal ATE: $\gamma_s(OLS) = \gamma_s(z) = \gamma_s$ for all s. As before, suppose we use linear OLS and IV estimators to test the null hypothesis. We reject the null hypothesis if $\beta(OLS) \neq \beta(z)$, and fail to reject otherwise. Assuming homogeneous treatment effects, the difference between the linear estimand (6) and the linear IV estimand (7) is

$$\beta(OLS) - \beta(z) = \sum_{s=1}^{\bar{s}} w_s(OLS)\gamma_s(OLS) - w_s(z)\gamma_s. \tag{8}$$

There are two cases. In the first case, the null hypothesis is true and $\gamma_s(OLS) = \gamma_s$ for at all s. Under the true null hypothesis, (8) becomes

$$\beta(OLS) - \beta(z) = \sum_{s=1}^{\bar{s}} (w_s(OLS) - w_s(z))\gamma_s.$$

Over-rejection of the true null hypothesis occurs when $\beta(OLS) - \beta(z) \neq 0$. If the marginal effects are constant ($\gamma_s = \gamma$ for all s) or the OLS and IV weights assigned to non-constant marginal effects are the same ($w_s(OLS) - w_s(z) = 0$), then $\beta(OLS) = \beta(z)$ and the Hausman test is a valid test of endogeneity in family size. However, in the general case of non-constant marginal effects, $\beta(OLS)$ may differ from $\beta(z)$ even when family size is exogenous if the OLS and IV weights are different: $(w_s(OLS) - w_s(z)) \neq 0$ for some s. Intuitively, the Hausman test over-rejects the null hypothesis because it confuses selection bias with differences in the linear OLS and IV estimators due to different weighting of non-constant marginal effects.

Figure 1 presents an example of this case. Recall that this figure is constructed from a simulation imposing the null hypothesis of no selection bias. When the linear model is correct ($\gamma_1 = \gamma_2 = 1$), the linear OLS and IV estimators provide the same estimate. However, as we introduce non-linearities, these estimators diverge because of different weighting of the marginal effects. To provide a sense of the over-rejection that is possible given non-linear treatment effects, we estimate the P-value of the Hausman selection test for each data sample and instrument. Figure 2 plots the average of the P-values, where we have normalized the P-value when the linear model is correct at 1.[8] This figure illustrates that the P-value for the Hausman selection bias test falls as we increase the level non-linearity. As we move away from the linear model, the fall in the P-value indicates that we are over-rejecting the true null

---

[8]The level of the P-value is not informative for this simulation example since it can be manipulated by changing the sample size or the degree of dispersion in the data.

hypothesis of no selection bias. The graph of P-values suggests that even modest degrees of non-linearity can change the probability of rejection of this commonly used test for selection bias.

Next, consider a second case in which the null hypothesis is false and there is selection bias: $\gamma_s(OLS) \neq \gamma_s$ for at least one s. Under-rejection occurs when $\beta(OLS) = \beta(z) \neq 0$. This occurs if selection bias at different birth parities offset each other:

$$\sum_{s=1}^{\bar{s}} w_s(OLS)\gamma_s(OLS) = \sum_{s=1}^{\bar{s}} w_s(z)\gamma_s.$$

Figure 3 graphs the three linear estimators maintaining the assumption of homogeneous marginal treatment effects but constructing the treatments to be endogenous: $Cov(d_{si}, \epsilon_i) \neq 0$ for $s = 1, 2$. As described in the Simulation Appendix, we have constructed positive selection bias which shifts the linear OLS estimate up. In Figure 3, the linear OLS estimator intersects with each of the linear IV estimators. For this simulation, the linear IV estimator using instrument z_1 is equal to the linear OLS estimator, $\beta(z_1) = \beta(OLS)$, when $\gamma_2 = -2$. In comparison, $\beta(z_2) = \beta(OLS)$ when $\gamma_2 = 2.5$. These two intersection points indicate under-rejection of the null hypothesis, as the linear OLS and linear IV estimators are equal even though the treatments are endogenous.

In Figure 4, we provide the relative P-values for this simulation. Because the two linear IV estimators are equal to the linear OLS estimate at different degrees of non-linearity, the peak of the P-value graphs is in different locations on the γ_2 axis. Notice that for the linear estimator using z_1, values of $\gamma_2 < 1$ lead to under-rejection of the false null hypothesis and higher P-values than at the point where the linear model is correct. An interesting aspect of this figure is that the P-value for the selection bias test falls below the level for the linear model at some points along the γ_2 axis. This is because at some values of γ_2, both the selection bias and the non-linearities push the linear IV estimators away from the linear OLS estimators.

3.4 Testing for Heterogeneous Treatment Effects

Given that different instruments define different linear IV estimands, Angrist et al. (2006) argue that using various instruments to estimate the same linear model can be used as a test for the presence of heterogeneous treatment effects and provide evidence on the external validity of the IV estimates. Angrist et al. (2006) construct several IV estimators exploiting various combinations of family size instruments and other included covariates to form "multiple natural experiments." The idea behind their test is that with treatment effect heterogeneity, the IV estimates should differ since each IV estimator defines the LATE for a different complier group. Because they generally find no precise effect of family size on children's outcomes when varying the instruments in their linear model, Angrist et al. (2006) conclude that there is strong case for a homogeneous zero effect of family size.

A difficulty in interpreting this test for heterogeneous treatment effects is that varying the instrument shifts not only the complier population but also the weights assigned to the potentially different marginal treatment effects. From (7), the difference between the IV estimand using instrument z and instrument z' can be expressed as:

$$
\begin{aligned}
\beta(z) - \beta(z') &= \sum_{s=1}^{\bar{s}} (w_s(z) - w_s(z'))\gamma_s \\
&+ \sum_{s=1}^{\bar{s}} \{w_s(z)E[\phi_{si}|Q_{si}] - w_s(z')E[\phi_{si}|Q_{si}]\} \\
&\equiv \Delta_\gamma(z, z') + \Delta_\phi(z, z').
\end{aligned} \tag{9}
$$

The first (second) term after the equality represents the first (second) term after the identity. As shown in (9), the difference between two IV estimands using different instruments consists of two parts: i) $\Delta_\gamma(z, z')$, a difference due to the different weights the instruments place on the marginal ATEs, and ii) $\Delta_\phi(z, z')$, a difference due to population heterogeneity in the treatment effects.

While $\beta(z) = \beta(z')$ could indicate a homogeneous effect of family size, as argued by Angrist et al. (2006), this need not be the case. To see this, consider the null hypothesis of homogeneous family size effects: $E[\phi_{si}|Q_{si}] = E[\phi_{si}] = 0$ for all s. If the null hypothesis is true, $\Delta_\phi(z, z') = 0$. Under an assumed linear model, which imposes $\gamma_s = \beta$ for all s, it follows that $\Delta_\gamma(z, z') = 0$. In this case, comparing $\beta(z)$ and $\beta(z')$ is a valid test of heterogeneous treatment effects. However, if the marginal effects are non-constant and $\Delta_\gamma(z, z') \neq 0$, $\beta(z)$ may differ from $\beta(z')$ even when the family size effects are homogeneous. In this case, the test statistic derived from comparing linear IV estimators $\beta(z) - \beta(z')$ leads to over-rejection of the true null hypothesis.

Figure 1 and Figure 3 provide simulations in which the null hypothesis of homogeneous treatment effects is maintained. Figure 1 assumes exogenous treatment, whereas Figure 3 imposes endogenous treatment. In both of these figures the linear IV estimators diverge from each other as we increase the degree of non-linearity, illustrating a case in which the true null hypothesis is over-rejected using linear IV estimators. Figure 5 provides the relative P-values for the test of equality between the two estimators for Figure 3 with endogenous but homogeneous treatment effects. Mirroring the divergence in the IV estimators in Figure 3, the P-value falls as we increase the degree of non-linearity.

On the other hand, non-linearities may also lead to under-rejection of the null hypothesis. Different instruments can produce the same linear IV estimate or average causal response if the differences owing to heterogeneous treatment effects are offset by the differences due to non-linearities, that is, when $\Delta_\gamma(z, z') = -\Delta_\phi(z, z')$. Consequently, the interpretation of $\beta(z) = \beta(z')$ as evidence for homogeneous family size effects, like in Angrist et al. (2006), rests on the assumption of a linear causal relationship.

If the researcher has more than one instrument available, Hausman (1978) proposed a test

for whether the additional instruments are valid in the sense that they are uncorrelated with the error term. Specifically, he suggested comparing the linear IV estimator using all instruments to the linear IV estimator using a single instrument. The idea is that if all instruments are valid, the estimates should differ only as a result of sampling error. As pointed out by Heckman et al. (2006) for example, this test rests on the assumption of homogeneous treatment effects, as different instruments generally identify different LATEs. Analogous to the arguments made above for the test for heterogeneous treatment effects, it is also necessary to assume constant marginal effects. Otherwise, the linear IV estimators using all instruments can differ from the linear IV estimator using a single instrument, even with homogenous marginal treatment effects at all margins, because of the different weighting of the marginal effects.

4 Data

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

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

The only difference between our sample selection and that in BDS (2005) is that we exclude a small number of families with more than 6 children.[9] The final sample includes 1,429,126 children from 625,068 families (98 % of the full sample of all families). Table 1 displays the basic descriptive statistics for this sample. In all respects, there are only minor differences between our sample and that of BDS (2005). Moreover, we cannot detect any difference between

---

[9]Our main reason for excluding large families is that the estimates of the marginal birth order effects and the marginal family size effects are unstable and imprecise for families with more than 6 children. We discuss these findings below.

the characteristics of the full sample and our sample of families with 6 or fewer children. About 48 percent of the children in the sample are female and a twin birth occurs in about 1.4 percent of families. The age of the child, the mother, and the father are measured in year 2000. The child's education is also collected from year 2000, and the education of the parents is measured at age 16 of the child. As expected, fathers are, on average, slightly older and more educated than mothers.

As in BDS (2005), our measure of family size is the number of children born to each mother. In the sample of families with 6 or fewer children, the average family size is 2.9 children. Table 2 provides the distribution of family sizes. Nearly 8 percent of the sample were only children, 33 percent were from 2 child families, and 32 percent were from 3 child families. The remaining 27 percent of the sample consists of children born to families with 4, 5, or 6 children.

5   OLS Estimates

This section compares OLS estimates of the linear and non-parametric models in family size, illustrating the sensitivity of the OLS results to the assumption of constant marginal effects.

5.1   Results for Full Sample

Table 3 reports the OLS estimates of the linear model (5) and non-parametric models in family size (3). This table replicates Table IV in BDS (2005, p 679), except we exclude children from families with more than 5 siblings. The first column of Table 3 shows that the OLS estimate of β in the linear family size model is -0.20, indicating that each additional sibling reduces average education of all the children in the family by 0.2 years.

The second column of Table 3 estimates the non-parametric model in family size. This model includes 5 sibling dummy variables: $d_{1i}, \ldots, d_{5i}$, where $d_{1i} = 1$ if child i has 1 or more siblings (family size of 2 children or more), $d_{2i} = 1$ if 2 or more siblings (family size of 3 or more children), and so on. The coefficient estimates of these dummy variables are the OLS estimates of the marginal effects of increasing family size by 1 additional sibling. Estimates of this non-parametric model indicate a non-monotonic relationship between family size and children's education. Moving from a 1 child family to a 2 child family is estimated to increase education by 0.37 years. In contrast, the marginal effects of additional siblings at higher birth parities are negative.

The remaining columns of Table 3 add control variables (the same as BDS 2005) to the linear and non-parametric models in family size. Columns 3 and 4 add dummy variables for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Including these variables reduces (in absolute value) both the linear and the non-parametric estimates of the effect of family size on children's education, suggesting that OLS estimation could be biased due to endogenous family size.

Columns 5 and 6 add a set of dummy variables for birth order. Like the non-parametric model in family size, the dummy variables for birth order are constructed to provide marginal effects of birth order. We include 5 birth order dummy variables: $b_{2i}, \ldots, b_{6i}$, where $b_{2i} = 1$ if child i was born second or higher in the birth order (and $b_{2i} = 0$ otherwise), $b_{3i} = 1$ if born third or higher in the birth order (and $b_{3i} = 0$ otherwise), and so on. We find, as BDS (2005), that the linear effect of family size in the model that controls for birth order and other demographic variables is very small, around -0.01.

As is evident from Column 6, relaxing the linearity assumption in family size reveals much larger marginal family size effects. In this specification, the birth order and family size dummy variables fully saturate the support of both variables, with the reference or omitted category specified as first born children in families with 1 child (only children). The estimates then indicate the marginal effect of increasing family size by 1 child (e.g. from 1 child family with 0 siblings to a 2 child family with 1 sibling) or being 1 birth parity later in the birth order (e.g from first to second born). Even controlling for birth order in this specification, the only child penalty is still strong, as the marginal effect of moving from a 0 to 1 siblings is 0.22 additional years of education. The marginal effect of moving from a 1 to 2 siblings is estimated to be small and positive at 0.02. However, the marginal effects of additional siblings at higher parities are between -0.073 and -0.089, 7 to 8 times larger than what indicated by the linear family size model.[10]

As emphasized by BDS (2005), the birth order effects are large. The estimates in Column 6 of Table 3 indicate that moving from first to second in the birth order lowers average education by 0.37 years, and moving from second to third in the birth order lowers average education by a further 0.22 years. However, the marginal effects of family size at higher parities are actually *larger* than the marginal effects of birth order. Adding a 4th sibling reduces children's education by 0.089 years, whereas moving from 4th to 5th in the birth order reduces children's education by about half as much, 0.04 years. Similarly, adding a 5th sibling reduces completed education by 0.084 years but the marginal effect of moving from 5th to 6th in the birth order reduces attainment by 0.06 years. It should be noted that given the standard errors of these estimates, we could not reject the hypothesis at the 5 percent level that the family size and birth order marginal effects are the same at these higher parities. But it is instructive that the OLS

---

[10]We construct the dummy variables as marginal effects to focus attention on the constant marginal effects restriction imposed by linearity. BDS (2005, p. 679, Table IV, Column 6) report similar OLS results using a non-parametric model in family size, where the dummy variables are constructed as total effects *relative to 1 child families*, i.e. dummy variables for whether child i is born to a family with 1 sibling or not, 2 siblings or not, and so on. This difference in the construction of dummy variables does not affect the estimation of the treatment effects since both specifications fully saturate the empirical support. However, given that the marginal effects are non-monotonic, caution should be used in interpreting these different dummy variable constructions. In particular, since many of the total effects *relative to 1 child families* are positive, one might conclude that there are no negative effects of additional children. Figures 6 and 7 show the total effects relative to only children are positive and increasingly small as the number of siblings increases, while the marginal effects (the slopes in these figures) are negative after the first sibling.

estimates of the birth order and family size marginal effects are similar in magnitude.[11]

5.2 OLS Weights

Table 4 reports the estimated weights for the linear OLS estimator. Given the distribution of family sizes in Norway, where most families have between 2 to 3 children, the OLS estimator places much more weight on the 1 to 2 sibling and 2 to 3 sibling marginal effects than on the other margins. The non-monotonic distribution of marginal family size effects and these particular OLS weights yield the near zero linear OLS estimate.

One implication of the dependence of the linear OLS estimator on the sample distribution of treatments is that the linear OLS estimate can vary from sample to sample as the distribution of family sizes changes, even if the OLS estimates of the marginal effects are the same. Consequently, the conclusion in previous research, like BDS (2005), of similar effects of family size across different samples rests on the assumption of constant marginal effects. By the same token, caution is called for when comparing linear OLS estimates across countries. To illustrate this point, we construct a linear OLS estimate combining the marginal family size effects from Norway reported in Column 6 of Table 3 with OLS weights based on the actual distribution of family sizes in Indonesia.[12] Given that larger family sizes are much more common in Indonesia compared to Norway, the Indonesian linear OLS estimator places more weight on the negative marginal effects. Re-weighting the Norwegian marginal effects estimates by the Indonesian linear OLS weights produces a linear estimate of -0.052. This constructed Indonesian linear estimate is several orders of magnitude larger than our linear OLS estimate for Norway reported in Column 5 of Table 3. On the other hand, re-weighting the Norwegian marginal effects for a country which has many 1 child families, as in modern China, would likely produce a positive linear estimate as much more weight would be placed on the positive 0 to 1 sibling marginal effect. This dependence of linear OLS estimators on the distribution of family sizes suggests that we cannot immediately interpret cross-country variation in linear estimates as evidence of different underlying relationships between family size and child outcomes.

5.3 Results by Birth Order

Table 5 reports results from the linear family size model (5) and the non-parametric model in family size (3), when estimated separately by birth order. Every model estimated in this

---

[11]We have also estimated the model in Column 6 of Table 3 for the sample of children from families with 1-10 children, including a full set of family size and birth order dummy variables. For the families with 7-10 children, the estimated marginal family size effects at these parities are negative but imprecise. At these higher parities, the estimated marginal birth order effects are more precise but unstable, alternating between positive and negative marginal effects. Estimated marginal family size effects (standard errors in parentheses): 6th sibling, -0.041 (0.032); 7th sibling -0.054 (0.051); 8th sibling -0.023 (0.077); 9th sibling -0.084 (0.11). Estimated marginal birth order effects (standard errors in parentheses): born 7th, -0.077 (0.040); born 8th, 0.18 (0.064); born 9th, -0.29 (0.10); born 10th, 0.097 (0.167).

[12]See Maralani (2008, Table 1). We use the family size distribution for the 1967-1977 cohorts, excluding children from families with more than 5 siblings, as with the Norwegian sample.

table includes the full set of demographic controls. The top panel of Table 5 estimates the linear family size model, whereas the bottom panel estimates the non-parametric model in family size. Contrasting the estimates from the two types of models for each birth order, indicates the extent to which the linear model approximates the underlying relationship between family size and child education. Figures 6 and 7 graph the predicted average child education from the models using the regression estimates reported in Table 5. The figures present educational attainment relative to only children, whose average educational attainment is normalized to 0.

For each of the birth order sub-samples, the coefficients on the main diagonal of Table 5 indicate the marginal effect of the first sibling on the youngest child in the family (e.g. the marginal effect on the first born child moving from 0 to 1 siblings, the marginal effect on the second born from moving from 1 to 2 siblings, and so on). The OLS estimates indicate that this marginal next child has a positive effect on first and second born children and a small negative (but insignificant) effect for later born children.[13] For each of the birth orders, the linear family size specification underestimates the negative effect of additional children beyond the marginal next child. Examining Figure 6, it is clear that the contrast between the linear and non-parametric specifications is particularly stark for the sub-sample of first born children. While the linear OLS specification predicts that additional children have a zero impact on first born children (linear estimate of 0.0001), the non-parametric specification predicts significant negative effects of having more than 1 sibling. Adding a 3rd sibling is estimated to reduce educational attainment of first born children by 0.086 years, adding a 4th sibling reduces education an additional 0.16 years, and a 5th sibling child an additional 0.11 years. These marginal effects are several orders of magnitude larger than the predictions from the linear model.

6   Linear IV Estimates

This section presents IV results from the linear family size model. Furthermore, we show how different weighting of the marginal family size effects lead to differences in the OLS and IV estimates of the linear model, even if family size is exogenously determined.

6.1   Twin Birth Instruments

Like previous studies, we use twin births as a source of exogenous variation in family size. The rationale for using twins as instruments is that for some families, twin births increase the number of siblings beyond the desired family size.[14] We follow BDS (2005) and Angrist et al.

---

[13]One interpretation of this result for first and second born children is that the birth of an additional child benefits the existing youngest child because this child learns from interacting with or teaching the younger sibling. Another interpretation is that parents are uncertain about the quality of their children and the realization of a high quality child makes them to choose to have an additional child.

[14]Following Rosenzweig and Wolpin (1980), twins births have been frequently used as an exogenous shock to family size. See for example BDS (2005) for results supporting the internal validity of twin birth as instrument for family size. Angrist et al. (2006) also use sex composition of the children as an instrument for family size.

(2006) in estimating the effect of family size separately for sub-samples of families with 1 or more, 2 or more, and 3 or more siblings. Specifically, for the sample of first born children in families with 1 or more siblings, we use the following second and first stages to construct the IV estimator for the linear family size model:

$$y_i = \beta s_i + X_i'\delta + \epsilon_i, \tag{10}$$

$$s_i = \lambda twin_{ci} + X_i'\rho + \eta_i, \tag{11}$$

where the vector of included covariates X_i includes a constant, and $twin_{ci}$ is a dummy variable equal to 1 if the cth birth in child i's family was a twin birth (implying the c and $c + 1$ births are twins).[15] The IV estimators for other birth orders are formed similarly. Under standard regularity conditions, a sufficient assumption for consistent estimation of the β parameters in (10) is a mean-independence assumption $E[\epsilon_i|X_i, Z_i] = 0$, where Z_i is the vector of excluded twin birth instruments.[16]

We also provide results based on the Angrist et al. (2006) strategy of estimating the linear specification using different instruments to increase the range of variation used in the IV estimations. In particular, we maintain (10) as the second stage, but change the first stage using twins at different birth parities as the instrument. A complication in using $twin_{ci}$ as an instrument is that this instrument is only defined for individuals born to families with at least c births. That is, for individuals with a completed family size of 2 children, $twin_{3i}$ is undefined. Angrist et al. (2006) prove how to form valid instruments for the full sample from twin birth instruments. We follow their procedure in constructing our twin instruments.

Table 6 presents IV results for the linear family size model, estimated separately by birth order. The first stage for each of the two-stage least-squares (2SLS) estimators is reported in the Appendix, Table C-1. As found in the previous literature, the twin birth instruments are strongly correlated with completed family size, so there is little concern about weak instruments here. The first column of Table 6 shows the OLS results for each sub-sample, whereas Columns 2-5 present 2SLS results using different twin instruments. The last column reports the over-identified IV estimator using all of the twin birth instruments. Moving across the columns of Table 6, we see that the second stage estimate of the linear family size effect changes as we use different twin birth instruments. For first born children in families with at least two children, the linear family size effect varies from a gain of 0.05 years of education from an additional child

---

However, recent evidence suggests that sex composition may have a direct effect on children's outcomes, implying that it may not be a valid instrument for family size (see e.g. Dahl and Moretti, 2008).

[15]To avoid including the endogenously selected outcomes of children born after the twin birth, we follow BDS (2005) and Angrist et al. (2006) and restrict the sample to children born before the twin birth. Our specification differs from BDS (2005) in that we estimate each of the models separately by birth order rather than pool birth order samples and include birth order dummy variables.

[16]As we include covariates X_i in these specifications, Assumption A1 on $twin_{ci}$ can be weakened to independence conditional on X_i. See Angrist and Imbens (1995) for a discussion of the interpretation of the IV estimator when covariates are included. An alternative strategy to estimating LATEs with covariates is found in Abadie (2003).

using twins on second birth to a loss of 0.04 years using twins on the fifth birth. For the other birth order sub-samples, the IV estimates display similar variation in second stage estimates as we change the instrument. In general, this non-constancy of the IV estimates may be because of treatment heterogeneity as different instruments identify different complier groups or because of non-linearities as the weighting of the marginal effects changes with the choice of instrument.

Like many previous studies, all of these 2SLS estimates of the linear family size effect are imprecise. Below the estimates and the standard errors, we report 95 confidence intervals in brackets. For all but 2 of the 12 2SLS estimates, the 95 percent confidence intervals cover the OLS estimates. Moreover, we cannot reject the hypothesis that almost all the effects of additional siblings exceed 0.1 years of education. The estimates from the over-identified IV model using all the instruments reported in the last column generates a gain in precision relative to the use of each instrument separately. Yet, the 95 percent confidence intervals still cover the OLS estimates, except for the one estimate using twins on the second birth for the sample of first born in families with at least 2 children.

6.2 Weights for the Linear IV Estimators

Table 7 calculates the IV weights, defined by (7), for β in (10) for the sample of first, second, and third born children. Table 7 also reports the corresponding OLS weights from (6). For simplicity, these weights are calculated omitting the covariates used in the empirical implementation of the OLS and IV estimators. As expected, using twins on the second birth as the instrument ($twin_{2i}$) weights the 1 to 2 children margin most heavily and has a small impact on changes in family size at higher parities. For the sub-sample of first born children in families with 1 or more siblings, 76 percent of the linear IV weight is on the 1 to 2 child marginal effect, whereas only 1.6 percent of the weight on the 4 to 5 child margin. A similar pattern is evident for the other twin birth instruments and other birth order samples. As discussed in detail above, the difference in the weighting of the marginal effects as we shift the instrument implies that we need to be cautious about interpreting the difference in the second stage results as reflecting treatment effect heterogeneity. By the same token, comparing the multiple linear IV estimates that use different instruments is likely to be a misleading test of the validity of certain instruments, as in a standard over-identification test.

Importantly, the distribution of OLS and IV weights are quite different. For the sample of first born children, the linear OLS estimator places 44 percent weight on the 1-2 child margin, far less than the twins at second birth instrument, which places 76 percent of the weight on this marginal effect. As discussed above, a standard test of endogeneity is to compare OLS and IV estimate from a linear model. However, the validity of this test rests on the assumption of constant marginal effects. Otherwise, OLS and IV estimates of a linear model may differ even when there is no selection bias, solely because of the different weighting of the marginal family size effects. The difference in weighting implies that the discrepancy in the OLS and

IV estimates reported in Table 6 may not necessarily indicate selection bias. This point is ignored in previous studies of the effects of family size on children's outcomes, which have concluded that significant differences between linear OLS and IV results suggest that family size is endogenously determined (see e.g. Caceres-Delpiano (2006)).

6.3 Predicted Fertility Instruments

In this sub-section, we re-estimate the linear family size model using predicted fertility to instrument for number of siblings. Our goal in using these alternative instruments is to more closely isolate particular treatment effect margins and to improve the precision of the IV estimates. Our estimation procedure consists of two steps. First, we estimate the predicted fertility instruments. In the second step, we apply the conventional 2SLS procedure to estimate the linear family size model, using these predicted fertility propensity scores as instruments for the number of siblings. The IV estimators based on the predicted fertility instruments are consistent under the same assumptions as the IV estimators using the excluded twin instruments directly.[17]

Given the vector of covariates X_i and the vector of twin birth instruments Z_i, we construct predicted fertility instruments at each sibling margin s as functions of these instruments: $p_s(X_i, Z_i) = pr(s_i \geq s | X_i, Z_i)$. As an example, consider the sample of first born children in families with 2 or more children, where the $p_2(X_i, Z_i)$ instrument is constructed as

$$p_2(X_i, Z_i) = \begin{cases} 1 & \text{if } twin_{2i} = 1 \\ f_2(X_i, \theta_2) & \text{if } twin_{2i} = 0 \end{cases} \tag{12}$$

This functional form recognizes that if there is a twin birth on the second birth, then the probability that child i has at least 2 siblings is 1. For a child from a family with a singleton birth on the second birth, the predicted probability that he or she has 2 or more siblings is specified as a non-linear function of the included covariates: $f_2(X_i, \theta_2) \in [0, 1]$. We construct the other instruments similarly. However, we cannot use future twin births (twin births on third birth, etc.) since part of this sub-sample does not have a third or higher birth. For example, for individuals with a completed family size of 2 children, $twin_{3i}$ is undefined. As above, we solve this complication by constructing the twin instruments as suggested by Angrist et al (2006). The remaining predicted fertility instruments are then defined as $p_s(X_i, Z_i) = f_s(X_i, Z_i, \theta_s)$ for $s = 3, 4, 5$, where Z_i includes these constructed twin instruments. The predicted fertility instruments are constructed analogously for other sub-samples.

After estimating the θ_s parameters using standard probit model estimation, we construct the predicted fertility instruments, $\hat{p}_s(X_i, Z_i)$. Results using a logit model are similar. In the second step, we apply the conventional 2SLS procedure to estimate the linear family size model

---

[17]Similar to our strategy, Wooldridge (2001) and Carnerio et al. (2003) provide examples of treatment effect analysis using propensity scores as instruments. In both applications, they find a substantial improvement in the precision of the IV estimates using the predicted instruments over the IV estimates using the excluded instrument directly.

(10), using the predicted fertility $\hat{p}_s(X_i, Z_i)$ as an instrument for the number of siblings s_i. Specifically, for the sample of first born children in families with 2 or more children, we use the same linear second stage model (10), but replace the 2SLS first stage from (11) with a 2SLS first stage specified as

$$s_i = \delta \hat{p}_s(X_i, Z_i) + X_i'\rho + \eta_i. \tag{13}$$

In general, the consistency of the IV estimator is unaffected by misspecification of the $p_s(X_i, Z_i)$ function and the asymptotic variance of the IV estimator is unaffected by the initial estimation of $p_s(X_i, Z_i)$. However, the small sample properties of the IV estimator may depend on whether we use the predicted fertility instruments or the twin instruments directly (see the discussion in Newey, 1990, 1993). Like Angrist et al (2006) who interact excluded instruments with covariates, the use of predicted fertility instruments generates an over-identified IV estimator which may exacerbate the small sample bias in IV estimation. We therefore choose a parsimonious specification of $f_s(\cdot)$.[18]

The first stage results are reported in the Appendix, Table C-2. Given our large samples and first-stage results showing that the predicted fertility instruments are strongly correlated with family size, the literature on small sample bias of the IV estimator suggests that this number of over-identifying restrictions should be of little concern (e.g. Staiger and Stock, 1997). Our simulation experiment reported in Appendix B supports this conjecture. The simulation experiment shows that the small sample bias and small sample variance of the IV estimator using the predicted fertility instruments are both smaller than for the IV estimator using the twin birth instruments directly.

Table 8 reports the second-stage results using the predicted fertility instruments in the linear IV model. For the sample of first in families with at least 2 children, the linear IV estimates clearly depend on the particular instrument chosen. Using the 2 or more siblings predicted fertility instrument $pr(s_i \geq 2)$ yields a small and imprecisely estimated linear coefficient of -0.0036. However, using the other instruments $pr(s_i \geq 3), \ldots, pr(s_i \geq 5)$ yield precise linear IV estimates between -0.46 and -0.63 for first born children. Using all of the instruments to instrument for the number of siblings in an over-identified model yields a linear coefficient of -0.41, which is between the lowest and highest linear estimates using each instrument separately.

To shed light on why different instruments may lead to different linear estimates, Table 9 provides the correlation of the predicted fertility instruments with each of the treatment margins for sample of first born children. The higher correlation on the diagonal indicates that each of these instruments is mostly correlated with certain treatment margins, although they have non-

---

[18] $f_2(\cdot)$ uses the following specification of covariates: i) linear and quadratic in child's own age, mother's age, and father's age, ii) 6 intercepts for each level of father's education and 6 intercepts for each level of mother's education, iii) an intercept for missing father's age, and iv) an intercept for child's sex. Adding the common intercept, this specification includes 21 unknown parameters. $f_s(\cdot)$ for $s = 3, 4, 5$ includes a linear function of each of the constructed twin instruments that occur after the first birth, in addition to the other 21 covariates. We also include in $f_5(\cdot)$ an intercept for whether child i's family had a twin on the second birth.

zero correlation with all treatment margins. If the marginal family size effects are constant (and homogeneous), the linear IV estimator based on different instruments should not yield different estimates. The fact that the linear estimate varies with the instrument chosen is suggestive of non-constant marginal LATEs, although we cannot rule out heterogeneous treatment effects as the only reason.

For the other birth order samples, a similar pattern emerges. For second born and third born, the instruments that are highly correlated with fertility at the lower birth parities result in significantly lower linear IV estimates. For each of these samples, the number of siblings has a statistically significant negative effect on children's completed education.

7 Non-Parametric IV Estimates

The sensitivity of the OLS results to the choice between a linear and a non-parametric model in family size underscores the need to be cautious when interpreting the IV estimates of the linear model. We directly address this issue by using twin births at each parity to form multiple instruments and estimate the non-parametric model in family size.

7.1 Twin Birth Instruments

For the sample of first born children in families of at least 2 children (1 or more siblings), the second and first stages of the 2SLS estimator are

$$y_i = \gamma_2 d_{2i} + \cdots + \gamma_5 d_{5i} + X_i'\delta + \epsilon_i, \tag{14}$$

$$d_{si} = \delta_2 twin_{2i} + \cdots + \delta_5 twin_{5i} + X_i'\rho + \eta_i \text{ for } s = 2, \ldots, 5, \tag{15}$$

where, as in the OLS estimation, $d_{si} = 1$ if child i has s or more siblings (total family size of $s + 1$ children or more).

The 2SLS estimator is formed using (15) as the first stage for (14). As with the linear IV estimator, a sufficient assumption for consistent estimation is mean-independence: $E[d_{si}|X_i, twin_{2i}, \ldots, twin_{5i}]$ for $s = 2, \ldots, 5$. We form the IV estimators for other birth orders similarly. It should be noted that a limitation in using twin births to instrument for family size is that we have no valid instrument for the 0 to 1 sibling margin since we cannot use twins on the first birth without considering the outcomes for first born children who are twins. Given the stark differences in the marginal effects between the 0 to 1 margin and the 1 to 2 sibling margin in the OLS results, we are reluctant to use a linear model to extrapolate our IV results to the 0 to 1 margin.[19]

---

[19] Alternative instruments using policy variation that induce families to increase family size from 1 to 2 children could be used to instrument for this margin. See, for example, Qian (2008) using the non-uniform application of the One Child policy in China. Interestingly, she finds a positive effect on first born children of an increase in family size from 0 to 1 siblings, which conforms with our OLS results.

Table 10 reports IV results for the non-parametric model in family size, using twin births at each parity to form multiple instruments, whereas the first stage for each 2SLS estimator is reported in the Appendix, Table C-3. Like the 2SLS estimates for the linear model using the twin birth instruments directly , the 2SLS estimates for the non-parametric specification are imprecisely estimated. For each of the 2SLS estimates, the 95 percent confidence intervals cover the non-parametric OLS estimates. It is interesting to note that for the higher parities (3rd born children in families with 4 or more children and 4th born children in families with 5 or more children) the 2SLS estimates show larger negative family size effects than the corresponding OLS estimates.

7.2 Predicted Fertility Instruments

We next turn to using the predicted fertility instruments. These instruments are attractive for IV estimation of the non-parametric model in family size, as they more closely isolate particular treatment effect margins and improve the precision of the estimates.

Depending on the sample (1st born children in families with 2 or more children, 2nd born children in families with 3 or more children, etc.), there are a number of endogenous regressors. For example, for the sample of first born children in families with 2 or more children, there are 4 endogenous family size dummy variables $d_{2i}, \ldots, d_{5i}$. The predicted fertility instruments $p_s(X_i, Z_i) = pr(s_i \geq s|X_i, Z_i)$ exploit two features of our particular application: twin births unequivacably increase family size by at least 1 child and the $p_s(X_i, Z_i)$ functions are in fact non-linear with range limited to the unit interval. Assuming a linear probability model, as is implicit in the specification using twin birth directly (15), ignores this non-controversial structure. In addition, because large families are relatively rare in Norway (about 4.5 percent of the families have 5 or more children), relaxing the linear probability model for the d_{si} variables seems to be especially warranted at higher parities.[20]

As with the linear IV estimation using the predicted fertility instruments, estimation takes two steps. First, we estimate the predicted fertility instruments using a probit model. As before, the results are robust to using a logit model. In the second step, we apply the conventional 2SLS procedure to estimate the non-parametric family size model (14), using the estimated instruments $\hat{p}_s(X_i, Z_i)$ to instrument for each of the endogenous number of siblings indicators $d_{2i}, \ldots, d_{5i}$. Specifically, for the sample of first born children in families with 2 or more children, we use the same non-parametric second stage model (14), but replace the first stage from

---

[20]From Kelejian (1971) we know that consistency of 2SLS does not rely on the correct specification of the first stage. However, the properties of the IV estimator are influenced by the instruments used, and in particular how correlated the instruments are with the endogenous variables.

(15) with a first stage specified as[21]

$$d_{si} = \delta_2 \hat{p}_2(X_i, Z_i) + \cdots + \delta_5 \hat{p}_5(X_i, Z_i) + X_i'\rho + \eta_i. \tag{16}$$

As with the linear IV estimator, the consistency and asymptotic distribution of the IV estimator is unaffected by the use of the predicted fertility instrument. To the extent that the predicted fertility instruments better approximate the relationship between family size and the X and Z variables, our predicted fertility instruments may have lower asymptotic variance.[22] In Appendix B, a Monte Carlo simulation shows that the finite sample properties of the non-parametric IV estimator using the predicted fertility instruments is superior to that using the twin birth instruments directly. In these simulations, we construct the predicted fertility instruments using a probit model, as discussed above. Both when the simulation assumes a Normal distribution for the error terms and when using a Gamma distribution with a high degree of skewness, the non-parametric IV estimators based on the predicted fertility instruments have lower average absolute value of bias and lower variance across simulations, relative to the non-parametric IV estimator using twin birth instruments directly. These simulation results provide some evidence that predicted fertility instruments do not increase the small sample bias of the IV estimation.

Returning to our Norwegian data, consider the first stage results presented in the Appendix. For each first stage regression, the total F-statistics is higher when using the predicted fertility instruments (Table C-4) than when we apply the twin instruments directly (Table C-3). The gains in the first stage fit are particularly large for the small probability events. For example, the F-statistic for the first stage for d_{4i} variable in the first born sample is 20 percent larger using the predicted fertility instruments than using twin birth instruments directly (269 vs. 225). The F-statistic for the first stage for 5 siblings (d_{6i}) in the first born sample is 30 percent larger using the predicted fertility instruments (185 vs. 142).

Table 11 provides the second stage estimates of the marginal effects of family size on children's education using the predicted fertility instruments. The main finding is that there are significant and large family size effects on children's education. Furthermore, the results indicate a non-monotonic causal relationship between family size and children's education. For

---

[21]To see the difference between the non-parametric IV estimator using twins directly and that using predicted fertility instruments, note that these two 2SLS estimators are equivalent: i) using (15) as the first stage or ii) using the first stage: $d_{si} = \delta_2 \tilde{p}_{2i} + \cdots + \delta_5 \tilde{p}_{5i} + X_i'\rho + \eta_i$, where $\tilde{p}_{si} = \hat{\kappa}_s twin_{ci}$ and $\hat{\kappa}_s$ is the OLS estimate from the regression of d_{si} on $twin_{ci}$. The difference between using the twin instruments and the predicted fertility instruments is the model used to predict the endogenous family size variables d_{si}. When using twin births directly, i.e. using (15) as the first stage, a linear probability model is assumed: $d_{si} = \kappa_s twin_{ci}$. In contrast, the IV estimator using the predicted fertility instruments (12) exploits the non-controversial structure discussed above.

[22]The optimal (lowest asymptotic variance) instruments are in general unknown. Newey (1990, 1993) discusses a number of non-parametric estimators for optimal instruments. As an alternative, we estimate our predicted fertility instruments assuming a particular parametric functional form to address the concern that a higher level of small sample bias may be introduced using more general non-parametric methods. However, as discussed below, we have estimated non-parametric instruments as well. If our parametric functional form is correct, our predicted fertility instruments are the optimal instruments. Even if our parametric functional form is misspecified, the IV estimator using these predicted fertility instruments is still consistent under the same mean-independence assumption used to justify the IV using twin instruments directly.

first born in families with at least 2 children, a third child is estimated to increase completed education by 0.15 years. This estimate is significantly different from the OLS estimate reported in Table 10 but within the 95 percent confidence interval for the IV estimators reported in Table 10 which use the twin instruments directly. Table 11 shows that additional children are estimated to reduce completed education by 0.47 years for a fourth child, another 0.8 years for a fifth child, and an additional 0.79 years for a sixth child. These estimates are several times larger than the corresponding OLS estimates presented in Table 10 and outside the lower bound of the 95 percent confidence intervals for the IV estimates reported in Table 10. It should also be noted that these marginal family size effects exceed the birth order effects that BDS (2005), Conley and Glauber (2006), and Price (2008) emphasize as large.

For later born children, the IV estimates in Table 10 provide no statistically significant effect on existing children's education from the birth of the next marginal child: 4th born child for 3 child families in Column (2), 5th born child for 4 child families in Column (3), and 6th born child for 5 child families in Column (4). However, we find statistically significant negative effects of additional children on educational attainment. For second born children in 3 child families, the birth of a 5th child reduces educational attainment by 0.57 years, and a sixth child by an additional 0.5 years. Similarly, for third born children in 4 child families, the birth of a sixth child reduces educational attainment by 0.52 years.

We have also estimated non-parametric optimal instruments, as suggested by Newey (1993). Specifically, we estimated $E[d_{si}|X_i, Z_i]$ for each permissible X_i and Z_i cell (both X_i and $twin_{ci}$ have discrete supports). Using the estimated $E[d_{si}|X_i, Z_i]$ instruments generated precise IV estimates of the non-parametric model in family size, with coefficient estimates similar to those for the non-parametric OLS. However, we are reluctant to report these results, since the large number of cells (over $100,000$ depending on the sample and endogenous variable) implies that this procedure uses a very large number of over-identifying restrictions, which could increase the small sample bias of the IV estimator considerably. Our approach here of using a particular non-linear model and a parsimonious parametric function of the X_i variables is intended to achieve a more reasonable tradeoff between bias and variance of the IV estimator. For an in-depth discussion of this issue, see Donald and Newey (2001).

8 Conclusions

Many empirical studies specify outcomes as a linear function of a potentially endogenous regressor when conducting IV estimation. These models restrict the marginal effects to be constant across all margins. In this paper, we examine the implications for inference from using OLS and IV estimators that assume a linear relationship between the outcome and the potentially endogenous regressors when the true relationship is non-linear. We find that non-linear treatment effects biases commonly used tests for non-zero treatment effects and selection bias. That these issues exist even in otherwise ideal conditions for IV estimation – strong instruments

and homogenous treatment effects – makes them especially important for empirical researchers. Finally, we demonstrate that comparing linear IV estimators using different instruments to make inferences about treatment effect heterogeneity can be misleading, as non-linearities can mask treatment effect heterogeneity or lead to an erroneous conclusion that treatment effects are heterogeneous when they are in fact homogeneous.

We demonstrate the empirical relevance of these results by re-examining the large body of empirical research that estimates the relationship between family size and children's education and tests the Becker and Lewis (1973) quantity-quality model. Much of the early literature that tested the quantity-quality model found that larger families reduced child quality, such as educational attainment (e.g. Rosenzweig and Wolpin, 1980; Hanushek, 1992). However, recent studies from several countries, using large data sets, extensive controls for confounding characteristics such as birth order, and instrumental variables for family size, have challenged this model and concluded that family size has no causal effect on children's outcomes (Black et al, 2005; Caceres-Delpiano, 2006; Angrist et al, 2006; Grønqvist, 2007).

All of these studies employ a linear family size model, although economic theory suggests that the relationship between family size and children's outcomes is likely to be non-linear and perhaps even non-monotonic. We show that the conclusion of no effect of family size is an artifact of the specification of a linear family size model. OLS and IV estimates of a non-parametric model in family size suggest that family size matters substantially, but in a non-monotonic way. Our IV estimates of negative effects of family size at higher parities exceed the birth order effects that previous studies have emphasized as large.

An understanding of the relationship between family size and children's outcomes can be important from a policy perspective. Most developed countries have a range of policies affecting fertility decisions, including publicly provided or subsidized child care as well as welfare and tax policies, such as maternity leave laws, family allowances, single parent benefits, and family tax credits. In fact, families with children receive special treatment under the tax and transfer provisions in twenty-eight of the thirty OECD countries (OECD, 2002). Many of these policies are designed such that they reduce the cost of having one child more than the cost of having additional children, in effect promoting smaller families.[23] If a policy goal is to slow or reverse the unprecedented fertility decline most developed countries have experienced over the last 30 years, the effects of family size on children's outcomes become ever more important. Accepting the recent findings of no effect of family size on existing children suggests that

---

[23]For example, welfare benefits are, in many cases, reduced or even cut off after reaching a certain number of children. In the United States, a recipient of the Earned Income Tax Credit program could in 2007 receive a maximum credit of USD $2,900$ if he or she had one qualifying child; for two or more qualifying children, the maximum credit was only USD $4,700$. In addition, a number of US states have implemented family cap policies, providing little or no increase in cash benefits when a child is born to a mother who is on welfare. Another example is from Norway, offering generous benefits to single parents. However, the benefit amount received is independent of the number of dependent children. Some developing countries have implemented far more radical policies to promote smaller families, such as China's One Child Policy and an aggressive public promotion of family planning in Mexico and Indonesia. See Feyrer et al. (2008) and Del Boca and Wetzels (2008) for recent reviews of the literature for developed countries.

policies promoting larger families would have few negative externalities on the human capital development of existing children. Our finding of a non-monotonic relationship between family size and children's education with large negative effects at higher parities suggests that we need to be cautious in this conclusion.

References

AASLUND, O., AND H. GRØNQUIST (2007): "Family Size and Child Outcomes: Is There Really No Trade-Off," IFAU Working Paper 15.

ABADIE, A. (2003): "Semi-Parametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–63.

ANGRIST, J., AND A. KRUEGER (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics, Vol. 3A*, ed. by O. Aschenfelter, and D. Card. North Holland, New York.

ANGRIST, J. D., K. GRADDY, AND G. W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67(3), 499–527.

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

ANGRIST, J. D., V. LAVY, AND A. SCHLOSSER (2006): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *MIT Working Paper, 06-26*.

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

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

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

BERNAL, R. (2008): "The Effect of Maternal Employment and Child Care on Children's Cognitive Development," *International Economic Review*, 49(4), 1173–1209.

BJØRKLUND, A., T. ERIKSSON, M. JANTTI, O. RAAUM, AND E. OSTERBACKA (2004): "Family Structure and Labor Market Success: The Influence of Siblings and Birth Order on the Earnings of Young Adults in Norway, Finland, and Sweden," in *Generational Income Mobility in North America and Europe*, ed. by M. Lorak. Cambridge University Press, MA.

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

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

CARNEIRO, P., J. HECKMAN, AND E. VYTLACIL (2003): "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," *working paper*.

CONLEY, D., AND R. GLAUBER (2006): "Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility," *Journal of Human Resources*, 41(4), 722–737.

DAHL, G., AND E. MORETTI (2008): "The Demand for Sons," *Review of Economic Studies*, 75(4), 1085–1120.

DEL BOCA, D., AND C. WETZELS (2008): *Social Policies, Labour Markets and Motherhood: A Comparative Analysis of European Countries*. Cambridge University Press.

FEYRER, J., B. SACERDOTE, AND A. STERN (2008): "Will the Stork Return to Europe and Japan? Understanding Fertility within Developed Nations," *Journal of Economic Perspectives*, 22(3), 3–22.

GARCIA, M., D. SHAW, E. WINSLOW, AND K. YAGGI (2000): "Destructive sibling conflict and the development of conduct problems in young boys," *Developmental Psychology*, 36(1), 44–53.

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

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

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

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

KELEJIAN, H. H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*, 66(334), 373–4.

MARALANI, V. (2008): "The Changing Relationship between Family Size and Educational Attainment Over the Course of Socioeconomic Development: Evidence from Indonesia," *Demography*, 45(3), 693–717.

MOFFITT, R. (2008): "Estimating Marginal Treatment Effects in Heterogeneous Populations," *working paper*.

NEWEY, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–37.

———— (1993): "Efficient Estimation of Models with Conditional Moment Restrictions," in *Handbook of Statistics, Vol. 11*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier.

NEWEY, W. K., AND S. G. DONALD (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–91.

OECD (2002): *Taxing Wages: 2001 Edition.* Organization for Economic Co-operation and Development, Paris.

PRICE, J. (2008): "Parent-Child Quality Time: Does Birth Order Matter?," *Journal of Human Resources*, 43(1), 240–265.

QIAN, N. (2008): "Quantity-Quality and the One Child Policy: The Positive Effect of Family Size on School Enrollment in China," *working paper.*

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

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

STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variable Regression with Weak Instruments," *Econometrica*, 65(3), 557–86.

SVARER, M., AND M. VERNER (2008): "Do Children Stabilize Relationships in Denmark?," *Journal of Population Economics*, 21(2), 395–417.

VANDELL, D. L., AND J. RAMANAN (1992): "Effects of early and recent maternal employment on children from low-income families," *Child Development*, 63(4), 938–949.

VURI, D. (2003): "Propensity Score Estimates of the Effects of Fertility on Marital Dissolution," *EUI Working Paper, 4.*

WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Section and Panel Data.* MIT Press, Cambridge, MA.

YITZHAKI, S. (1996): "On Using Linear Regressions In Welfare Economics," *Journal of Business and Economics Statistics*, 14(4), 478–486.

9 Figures

Figure 1: Linear OLS and IV Estimators (Exogenous Treatment, Homogeneous Effects)
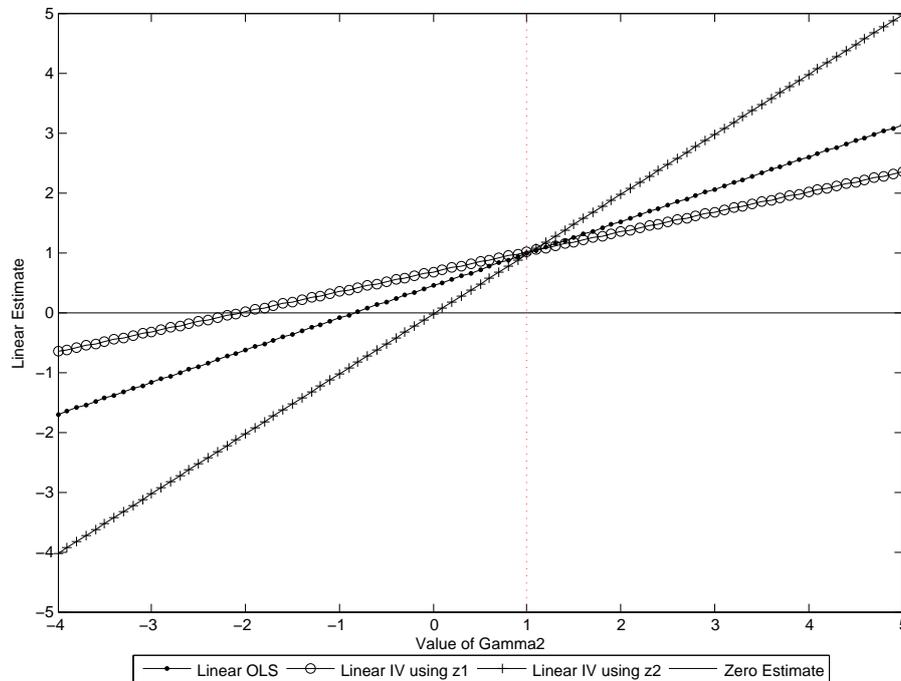
Notes: This figure provides an illustration of how non-linearities in treatment effects differentially affects various linear estimators. The underlying model is $y_i = \gamma_1 d_{1i} + \gamma_2 d_{2i} + \epsilon_i$, where the first marginal effect is fixed at $\gamma_1 = 1$. The vertical axis measures the level of one of three linear estimators: linear OLS $\beta(OLS)$, linear IV using z_1 as an instrument $\beta(z_1)$, and linear IV using an z_2 $\beta(z_2)$. The vertical axis measures γ_2. At $\gamma_2 = \gamma_1 = 1$, the linear model is correct and the marginal effects are constant. This figure is drawn assuming no selection bias and that treatment effects are homogeneous. Therefore, when the linear model is correct, all three estimators are equal: $\beta(OLS) = \beta(z_1) = \beta(z_2)$, modulo sampling error. As γ_2 moves away from $\gamma_2 = 1$, the three estimators diverge from each other due to their different weighting of the marginal effects.

Source: Simulation from data generating process. See Appendix A.

Figure 2: Relative P-Value for Test of No Selection Bias, $\beta(OLS) = \beta(z)$ (Exogenous Treatment, Homogeneous Treatment Effects)

Notes: This figure provides the P-Value from two tests of no selection bias: $H_0 : \beta(OLS) - \beta(z_1) = 0$ using z_1 as the instrument and $H_0 : \beta(OLS) - \beta(z_2) = 0$ using z_2 as the instrument.

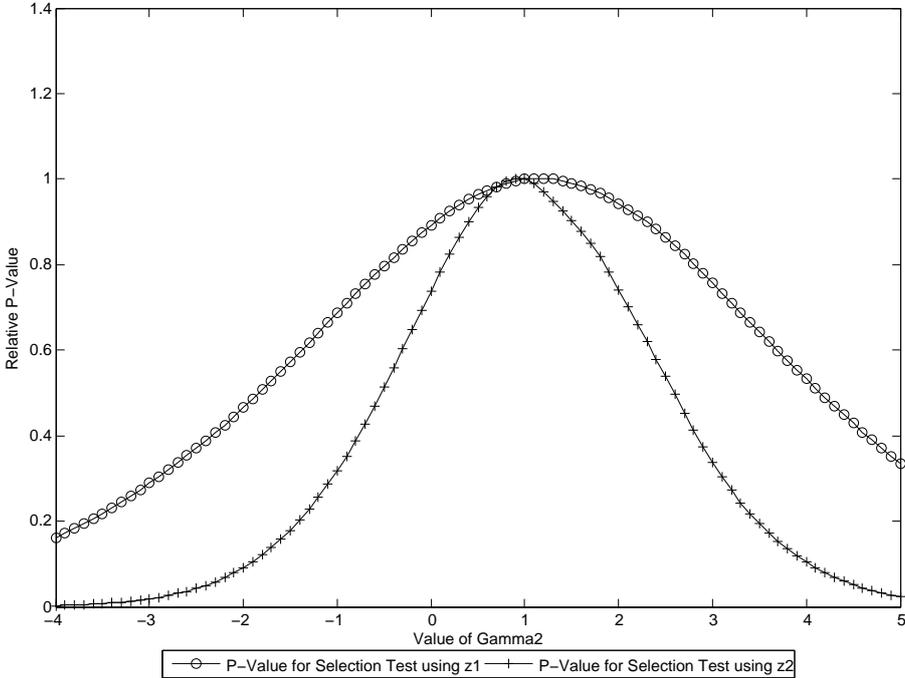Source: Simulation from data generating process. See Appendix A.

Figure 3: Linear OLS and IV Estimators (Endogenous Treatment, Homogeneous Treatment Effects)
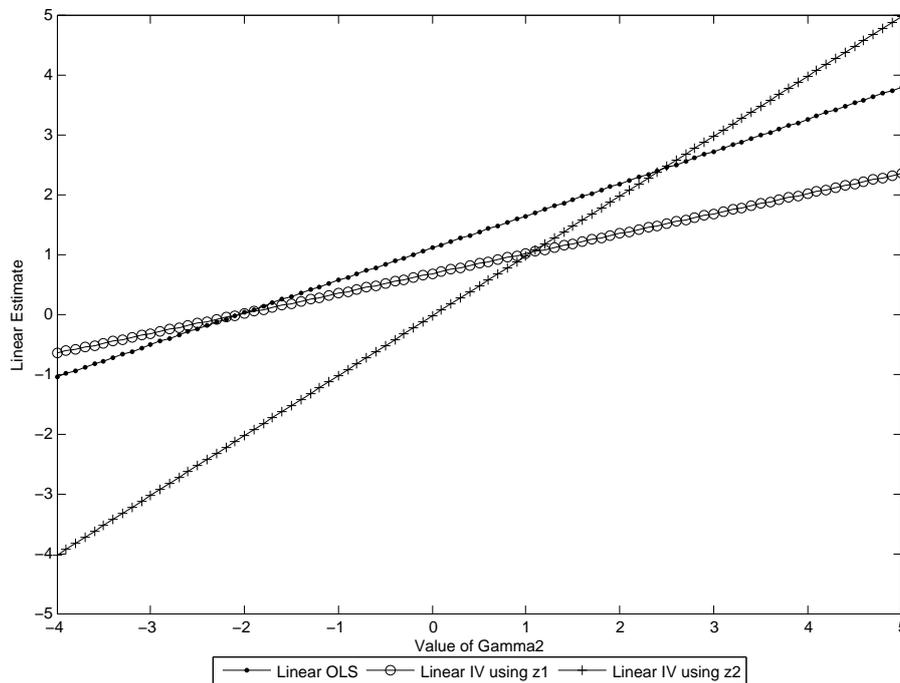
Notes: This figure provides an illustration of how non-linearities in treatment effects differentially affects various linear estimators. The underlying model is $y_i = \gamma_1 d_{1i} + \gamma_2 d_{2i} + \epsilon_i$, where the first marginal effect is fixed at $\gamma_1 = 1$. The vertical axis measures the level of one of three linear estimators: linear OLS $\beta(OLS)$, linear IV using z_1 as an instrument $\beta(z_1)$, and linear IV using an z_2 $\beta(z_2)$. The vertical axis measures γ_2. At $\gamma_2 = \gamma_1 = 1$, the linear model is correct and the marginal effects are constant. In this Ffigure, we impose positive selection bias, therefore $\beta(OLS) \neq \beta(z)$. However, because the treatment effects are constructed to be homogeneous, when the linear model is correct, the linear IV estimators are equal: $\beta(z_1) = \beta(z_2)$, modulo sampling error. As γ_2 moves away from $\gamma_2 = 1$, the linear IV estimators diverge from each other due to their different weighting of the marginal effects.

Source: Simulation from data generating process. See Appendix A.

Figure 4: Relative P-Value for Test of No Selection Bias, $\beta(OLS) = \beta(z)$ (Endogenous Treatment, Homogeneous Treatment Effects)
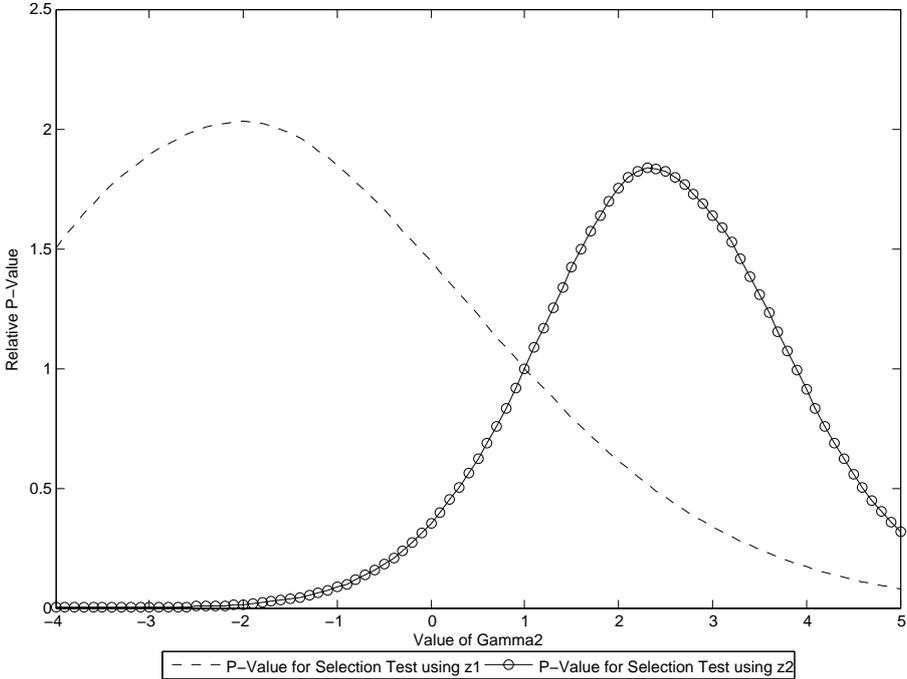
Notes: This figure provides the P-Value from two tests of no selection bias: $H_0 : \beta(OLS) - \beta(z_1) = 0$ using z_1 as the instrument and $H_0 : \beta(OLS) - \beta(z_2) = 0$ using z_2 as the instrument.

Source: Simulation from data generating process. See Appendix A.

Figure 5: Relative P-Value for Test of Homogeneous Treatment Effects, $\beta(z_1) = \beta(z_2)$ (Endogenous Treatment, Homogeneous Treatment Effects)
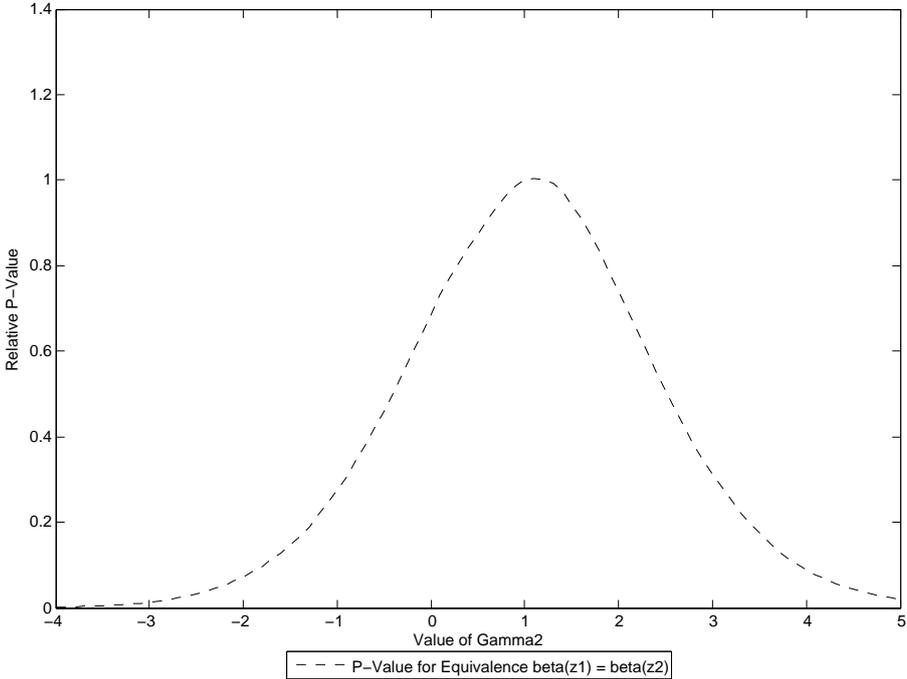
Notes: This figure provides the P-Value from a test of equivalence of the two instrumental variable estimators: $H_0 : \beta(z_1) - \beta(z_2) = 0$.

Source: Simulation from data generating process. See Appendix A.

Figure 6: Average Educational Attainment for First Born Children by Number of Siblings (Relative to Only Children)
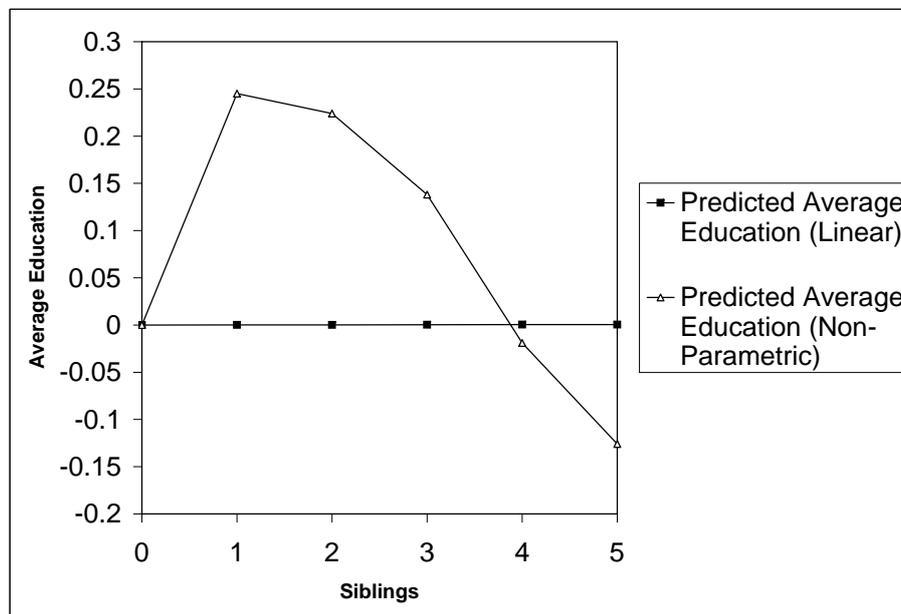
Notes: This figure graphs the linear and non-parametric in family size predictions from OLS regressions. These values are graphed relative to only children (0 siblings), i.e. the education of only children is normalized to 0. The slopes in this figure provide the estimated marginal family size effects at each margin, where the linear model imposes constant slopes whereas the non-parametric model allows non-constant slopes. The linear prediction is $\hat{y} = \hat{\beta} * s$ for $s = 0, 1, \ldots, 5$, where s is number of siblings and $\hat{\beta}$ is the OLS estimate from the first panel of Table 5. Non-parametric prediction is $\hat{y} = \hat{\gamma}_1 * 1\{s \geq 1\} + \cdots + \hat{\gamma}_5 * 1\{s = 6\}$, where $\hat{\gamma}_s$ are the OLS estimates from the second panel of Table 5.

Source: Administrative registers from Statistics Norway.

Figure 7: Average Educational Attainment for Second Born Children by Number of Siblings (Relative to Children with 1 Sibling)
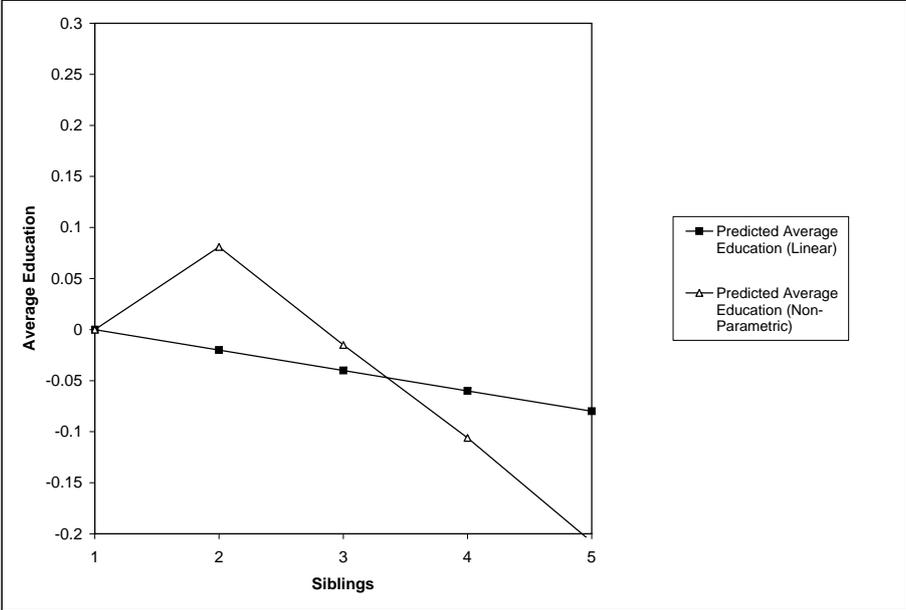
Notes: This figure graphs the linear and non-parametric in family size predictions from OLS regressions. These values are graphed relative to second born children in 2 child families (1 sibling), i.e. the education of second born children in 2 child families is normalized to 0. The slopes in this figure provide the estimated marginal family size effects at each margin, where the linear model imposes constant slopes whereas the non-parametric model allows non-constant slopes. The linear prediction is $\hat{y} = \hat{\beta} * (s - 1)$ for $s = 1, 2, \ldots, 5$, where s is number of siblings and $\hat{\beta}$ is the OLS estimate from the first panel of Table 5. Non-parametric prediction is $\hat{y} = \hat{\gamma}_1 * 1\{s \geq 2\} + \cdots + \hat{\gamma}_5 * 1\{s = 5\}$, where $\hat{\gamma}_s$ are the OLS estimates from the second panel of Table 5.

Source: Administrative registers from Statistics Norway.

10 Tables

Table 1: Descriptive Statistics

	Mean	Std. Dev.
Age in 2000	38.5	8.6
Female	0.48	0.50
Education	12.1	2.6
Mother's education	9.9	1.3
Father's education	10.3	2.2
Mother's age in 2000	65.8	10.6
Father's age in 2000	67.3	10.3
Number of children	2.9	1.2
Twins in family	0.014	0.12

Notes: Descriptive statistics are for 1,429,126 children from 625,068 families with no more than 6 children. (98 % of the full sample). All children are aged at least 25 in 2000. Twins are excluded from the sample. All children and parents are aged between 16 and 74 years at some point between 1986 and 2000.

Source: Administrative registers from Statistics Norway.

Table 2: Distribution of Family Sizes by Children

Family Size	Number	Fraction
1	111,064	0.078
2	477,633	0.334
3	459,831	0.322
4	239,840	0.168
5	99,940	0.070
6	40,818	0.029

Notes: Descriptive statistics are for 1,429,126 children from 625,068 families with no more than 6 children. (98 % of the full sample). All children are aged at least 25 in 2000. Twins are excluded from the sample. All children and parents are aged between 16 and 74 years at some point between 1986 and 2000.

Source: Administrative registers from Statistics Norway.

Table 3: OLS Estimates of Linear and Non-Parametric Models in Family Size

	(1)	(2)	(3)	(4)	(5)	(6)
Linear Family Size	-0.198		-0.112		-0.008	
	(0.003)		(0.002)		(0.003)	
Siblings ≥ 1		0.370		0.042		0.224
		(0.009)		(0.008)		(0.001)
Siblings ≥ 2		-0.148		-0.099		0.020
		(0.007)		(0.006)		(0.006)
Siblings ≥ 3		-0.352		-0.157		-0.073
		(0.009)		(0.007)		(0.008)
Siblings ≥ 4		-0.348		-0.146		-0.089
		(0.014)		(0.012)		(0.012)
Siblings $= 5$		-0.281		-0.131		-0.084
		(0.023)		(0.019)		(0.019)
Birth Order ≥ 2					-0.332	-0.373
					(0.005)	(0.005)
Birth Order ≥ 3					-0.222	-0.219
					(0.006)	(0.006)
Birth Order ≥ 4					-0.157	-0.100
					(0.009)	(0.009)
Birth Order ≥ 5					-0.106	-0.040
					(0.015)	(0.015)
Birth Order $= 6$					-0.117	-0.063
					(0.029)	(0.029)
Control Variables	No	No	Yes	Yes	Yes	Yes
R-Squared	0.008	0.012	0.204	0.204	0.208	0.209

Notes: Each column is a separate regression. Standard errors in parentheses are robust to within family clustering and heteroskedasticity. Control variables include dummy variables for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education.

Source: Administrative registers from Statistics Norway.

Table 4: Linear OLS Weights on Marginal Effects for Full Sample

Sibling Margin:	0-1	1-2	2-3	3-4	4-5
	$w_1(OLS)$	$w_2(OLS)$	$w_3(OLS)$	$w_4(OLS)$	$w_5(OLS)$
Sample: All Families and all Children:					
OLS Weight	0.110	0.336	0.313	0.175	0.066

Notes: This table reports the weights for the linear OLS estimator with no additional covariates. These weights applied to the OLS marginal effects in Column (2) of Table 3 produce the linear OLS estimate in Column (1) of Table 3, modulo rounding: $\beta(OLS) = \sum_{s=1}^{5} w_s(OLS)\gamma_s(OLS) = 0.110 * 0.370 + 0.336 * (-0.148) + 0.313 * (-0.352) + 0.175 * (-0.348) + 0.066 * (-0.281) = -0.1986$.

Source: Administrative registers from Statistics Norway.

Table 5: OLS Estimates by Birth Order of Linear and Non-Parametric Models in Family Size

| | Birth Order | | | | |
	1	2	3	4	5
Panel I: Linear					
Numb. of Child.	0.0001	-0.020	-0.037	-0.037	-0.006
	(0.003)	(0.004)	(0.007)	(0.013)	(0.033)
Panel II: Non-Parametric					
Siblings ≥ 1	0.245				
	(0.009)				
Siblings ≥ 2	-0.021	0.081			
	(0.007)	(0.008)			
Siblings ≥ 3	-0.086	-0.096	-0.010		
	(0.011)	(0.010)	(0.012)		
Siblings ≥ 4	-0.157	-0.091	-0.055	-0.010	
	(0.019)	(0.019)	(0.018)	(0.020)	
Siblings $= 5$	-0.107	-0.072	-0.102	-0.091	-0.006
	(0.033)	(0.032)	(0.0301)	(0.031)	(0.033)

Notes: Each column of each panel is a separate regression. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are heteroskedastic robust but clustering is not necessary given that regression includes only 1 child from each family.

Source: Administrative registers from Statistics Norway.

Table 6: IV Estimates of Linear Model in Family Size using Twin Instruments

Instrument(s):	OLS	2SLS Twin2	2SLS Twin3	2SLS Twin4	2SLS Twin5	2SLS All
	(1)	(2)	(3)	(4)	(5)	(6)
Sample: Fam. Size ≥ 2, 1st Birth						
Numb. of Child.	-0.063	0.053	-0.035	0.017	-0.044	0.013
	(0.004)	(0.050)	(0.059)	(0.098)	(0.188)	(0.035)
		[-0.044 0.151]	[-0.150, 0.079]	[-0.174, 0.208]	[-0.412, 0.324]	[-0.055, 0.081]
Sample: Fam. Size ≥ 3, 2nd Birth						
Numb. of Child.	-0.078		-0.051	-0.026	0.123	-0.031
	(0.006)		(0.058)	(0.096)	(0.238)	(0.047)
			[-0.165, 0.062]	[-0.213, 0.162]	[-0.344, 0.589]	[-0.124 0.061]
Sample: Fam. Size ≥ 4, 3rd Birth						
Numb. of Child.	-0.051			-0.107	-0.168	-0.122
	(0.013)			(0.089)	(0.160)	(0.077)
				[-0.282, 0.067]	[-0.490, 0.138]	[-0.274, -0.029]

Notes: Each column in each panel is separate estimation. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. TwinC is an indicator for a twin at the Cth birth, e.g. Twin2 is an indicator for twin at the second birth (second and third children are twins). Standard errors in parentheses are heteroskedastic robust but clustering is not necessary given that regression includes only 1 child from each family. 95 percent confidence intervals in brackets.

Source: Administrative registers from Statistics Norway.

Table 7: OLS and IV Weights on Marginal Family Size Effects

Sibling Margin:	0-1	1-2	2-3	3-4	4-5
Sample: Fam. Size ≥ 2, 1st Birth					
OLS Weight	–	0.444	0.344	0.154	0.053
IV (Instr: Twin2) Weight	–	0.763	0.170	0.050	0.016
Sample: Fam. Size ≥ 3, 2nd Birth					
OLS Weight	–	–	0.547	0.333	0.119
IV (Instr: Twin3) Weight	–	–	0.851	0.122	0.027
Sample: Fam. Size ≥ 4, 3rd Birth					
OLS Weight	–	–	–	0.678	0.322
IV (Instr: Twin4) Weight	–	–	–	0.882	0.117

Notes: Weights are for the models without additional covariates. The formula for the OLS and IV weights is given in the text.

Source: Administrative registers from Statistics Norway.

Table 8: IV Estimates for Linear Model in Family Size using Predicted Fertility Instruments

Instrument(s):	2SLS $pr(s_i \geq 2)$ (1)	2SLS $pr(s_i \geq 3)$ (2)	2SLS $pr(s_i \geq 4)$ (3)	2SLS $pr(s_i \geq 5)$ (4)	2SLS All (5)
Sample: Fam. Size $\geq$ 2, 1st Birth					
Numb. of Child.	-0.0036 (0.0483)	-0.4615 (0.0465)	-0.6015 (0.0516)	-0.6392 (0.0634)	-0.4051 (0.027)
Sample: Fam. Size $\geq$ 3, 2nd Birth					
Numb. of Child.		-0.1708 (0.0541)	-0.4570 (0.0612)	-0.4283 (0.0830)	-0.3313 (0.0383)
Sample: Fam. Size $\geq$ 4, 3rd Birth					
Numb. of Child.			-0.1914 (0.0875)	-0.4119 (0.1058)	-0.2756 (0.0685)

Notes: Each column in each panel is separate estimation. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. $pr(s_i \geq s)$ is the predicted fertility instrument that the number of siblings exceeds s, e.g. $pr(s_i \geq 2)$ is the predicated probability that a child has 2 or more siblings. Standard errors in parentheses are heteroskedastic robust but clustering is not necessary given that regression includes only 1 child from each family. 95 percent confidence intervals in brackets.

Source: Administrative registers from Statistics Norway.

Table 9: Correlation of Predicted Fertility Instruments with Sibling Treatments for First Born Children

Sibling Treatment:	$1\{s_i \geq 2\}$	$1\{s_i \geq 3\}$	$1\{s_i \geq 4\}$	$1\{s_i \geq 5\}$
Instrument:				
$pr(s_i \geq 2)$	0.3096	0.2643	0.1769	0.0978
$pr(s_i \geq 3)$	0.2562	0.3172	0.2049	0.1167
$pr(s_i \geq 4)$	0.2168	0.2627	0.2670	0.1330
$pr(s_i \geq 5)$	0.1599	0.2069	0.2000	0.2274

Notes: This table provides the correlation between the predicted fertility instrument and each marginal treatment, e.g. the first entry is $Corr(pr(s_i \geq 2), d_{2i}) = 0.3096$, where $d_{2i} = 1\{s_i \geq 2\}$. The sample used is for first born children with at least 1 sibling.

Source: Administrative registers from Statistics Norway.

Table 10: IV Estimates of Non-Parametric Model in Family Size using Twin Instruments

Birth:	1st		2nd		3rd		4th	
Sample:	Fam. Size ≥ 2		Fam. Size ≥ 3		Fam. Size ≥ 4		Fam. Size ≥ 5	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Siblings ≥ 2	-0.0127 (0.008)	0.0793 (0.0677) [-0.054, 0.212]						
Siblings ≥ 3	-0.079 (0.011)	-0.044 (0.069) [-0.179, 0.092]	-0.081 (0.011)	-0.059 (0.070) [-0.196, 0.078]				
Siblings ≥ 4	-0.151 (0.0192)	0.023 (0.111) [-0.195, [0.242]	-0.078 (0.019)	-0.044 (0.106) [-0.251, 0.163]	-0.034 (0.019)	-0.096 (0.104) [-0.300, 0.107]		
Siblings $= 5$	-0.102 (0.033)	-0.051 (0.188) [-0.419, 0.318]	-0.063 (0.032)	0.133 (0.205) [-0.269, 0.535]	-0.088 (0.031)	-0.178 (0.159) [-0.492 0.107]	-0.082 (0.032)	-0.100 (0.173) [-0.439, 0.238]

Notes: Each column is a separate regression. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are robust to heteroskedasticity but clustering is not necessary given that regression includes only 1 child from each family. 95 percent confidence intervals in brackets.

Source: Administrative registers from Statistics Norway.

Table 11: IV Estimates of Non-Parametric Model in Family Size using Predicted Fertility as Instruments

Birth:	1st	2nd	3rd	4th
Sample:	Fam. Size ≥ 2	Fam. Size ≥ 3	Fam. Size ≥ 4	Fam. Size ≥ 5
	(1)	(2)	(3)	(4)
Siblings	0.153			
≥ 2	(0.066)			
Siblings	-0.474	-0.084		
≥ 3	(0.075)	(0.070)		
Siblings	-0.800	-0.572	-0.145	
≥ 4	(0.131)	(0.116)	(0.108)	
Siblings	-0.787	-0.504	-0.520	-0.175
$= 5$	(0.219)	(0.218)	(0.162)	(0.172)

Notes: Each column is separate regression. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses. Standard errors in parentheses are robust to heteroskedasticity but clustering is not necessary given that each regression includes only 1 child from each family.

Source: Administrative registers from Statistics Norway.

APPENDIX

A Simulation Details

This appendix provides the details for the data simulation experiment in Section 3. The observed outcome is

$$y_i = \gamma_{1i} d_{1i} + \gamma_{2i} d_{2i} + \epsilon_i,$$

where $\gamma_{1i} = \gamma_1 + \sigma_{\phi 1} \phi_{1i}$ and $\gamma_{2i} = \gamma_2 + \sigma_{\phi 2} \phi_{2i}$ with $\phi_{si} \sim N(0,1)$ for $s = 1, 2$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.

The level of treatment is given by

$$d_{1i} = 1\{\pi_{11} z_{1i} + \pi_{12} z_{2i} + \alpha_{1,\epsilon} \epsilon_i + \phi_{1i} + \psi_{1i} \geq 0\},$$

$$d_{2i} = d_{1i} 1\{\pi_{21} z_{1i} + \pi_{22} z_{2i} + \alpha_{2,\epsilon} \epsilon_i + \phi_{2i} + \psi_{2i} \geq 0\},$$

where $\psi_{si} \sim N(0, \sigma_{\psi 1}^2)$, and z_{1i} and z_{2i} are binary instruments distributed as $z_{si} = 1$ with probability 0.5 and $z_{si} = 0$ otherwise, for $s = 1, 2$. They are constructed to be uncorrelated: $z_{1i} \perp z_{2i}$.

We simulate the model for two combinations of parameters.

Case 1: Exogenous Treatment and Homogeneous Treatment Effects: $\gamma_1 = 1$, $\sigma_{\phi 1} = 0$, $\sigma_{\phi 2} = 0$, $\sigma_\epsilon = 10$, $\sigma_{\psi 1} = 1$, $\sigma_{\psi 2} = 1$, $\pi_{11} = 3$, $\pi_{12} = 0$, $\pi_{21} = 0$, $\pi_{22} = 3$, $\alpha_{1,\epsilon} = 0$, $\alpha_{2,\epsilon} = 0$.
Figure 1 plots the average value of the linear estimates across the replications of the data generating process, and Table A-1 provides the corresponding weights on the marginal effects.

Case 2: Endogenous Treatment and Homogeneous Treatment Effects: Same parameters as Case 1 except $\alpha_{1,\epsilon} = 0.01$, $\alpha_{2,\epsilon} = 0.01$.
The degree of endogeneity in this simulation creates on average an upward bias in the linear OLS estimate by 63 percent over the true coefficient: $\beta(OLS) - 1 = 1.63$, when $\gamma_1 = \gamma_2 = 1$.

Table A-1: Weights on Marginal Effects by Estimator

	Margin 1 (γ_1)	Margin 2 (γ_2)
Linear OLS Weight	0.4621	0.5379
Linear IV Weight using z_1	0.6666	0.3334
Linear IV Weight using z_2	0	1

Source: Simulation from data generating process.

B Predicted Fertility Instrument Simulation

In this appendix, we use a simulation exercise to examine the small sample properties of the IV estimators using the predicted fertility instruments. We focus on first born children with 1 to 3 siblings (2 to 4 total children). For each first born child, the data consists of a number of

siblings $s_i \in \{1, 2, 3\}$, one scalar exogenous covariate x_i (e.g. mother's education), two twin birth instruments $twin_{2i}$ (twin on second birth) and $twin_{3i}$ (twin on third birth), and an observed outcome for the first born child y_i.

We specify the following data generating process. In the absence of twin births, the choice of family size takes an ordered choice form with latent utility from children given by $u_i = \alpha x_i + \epsilon_i$. The number of siblings is selected as $s_i = 1$ if $u_i < \pi_2$, $s_i = 2$ if $\pi_2 \leq u_i < \pi_3$, and $s_i = 3$ if $u_i \geq \pi_3$. The twin birth instruments exogenously increase siblings by one child: $s_i = 2$ if $twin_{2i} = 1$ and $s_i = 3$ if $twin_{2i} = 1$. The observed outcome is then $y_i = \gamma_2 d_{2i} + \gamma_3 d_{3i} + \rho x_i + \epsilon_i$, where $d_{2i} = 1\{s_i \geq 2\}$ and $d_{3i} = 1\{s_i \geq 3\}$. Random variables are distributed $x_i \sim N(1, 1)$ and $twin_{ki} = 1$ with probability 0.05, for $k = 2, 3$. The remaining parameters are set at $\pi_2 = 1$, $\pi_3 = 1.5$, $\alpha = 1$, $\gamma_2 = 1$, $\gamma_3 = -1$, and $\rho = 1$. In this data generating process the marginal effects of family size are homogeneous across families but non-constant across margins ($\gamma_2 \neq \gamma_3$).

Table B-1 presents the simulation results for 500 replications. For each replication, we draw a sample of 10,000 observations from the data generating process. We conduct two simulations. The first simulation assumes ϵ_i is distributed standard normal: $\epsilon_i \sim N(0, 1)$. The second simulation assumes ϵ_i is distributed according to the Gamma distribution with shape parameter of 2 and scale parameter of 1: $\epsilon_i \sim G(1, 2)$. This parametrization implies that distribution of ϵ_i has skewness $2/\sqrt{2}$ and kurtosis 3. By contrast, the Normal distribution has skewness and kurtosis of 0. For each simulated sample, we compute three estimators of the γ_2 and γ_3 parameters: i) OLS, ii) IV using the twin birth instruments directly, and iii) IV using the predicted fertility as instruments, where the predicted fertility instruments are constructed as discussed above.

The results in Table B-1 display several finite sample characteristics for each estimator. Across the $R = 500$ replications of the data generating process, we calculate the mean of the absolute bias for each parameter: $\frac{1}{R} \sum_{r=1}^{R} |\hat{\gamma}_{sr} - \gamma_s|$ for $s = 1, 2$, where γ_s is the true parameter and $\hat{\gamma}_{sr}$ is the rth simulation estimate. We also calculate the standard deviation of the estimates across the simulations: $\sqrt{\frac{1}{R} \sum_{r=1}^{R} (\hat{\gamma}_{sr} - \bar{\hat{\gamma}}_s)^2}$, where $\bar{\hat{\gamma}}_s$ is the mean of the estimates across the simulations. Finally, we calculate mean squared error as the variance in the estimators across the replications plus the mean squared bias.

For each parameter and error distribution assumption, the OLS estimator is severely biased with the mean absolute value of bias around 1 or higher. All the IV estimators have substantially lower levels of bias than the OLS estimators. However, the IV estimators using the twin births directly have higher levels of bias, higher variance, and higher mean squared error than the IV estimators using the predicted fertility instruments. This is true across parameters and assumptions about the distribution of the error. When the ϵ_i follows a Gamma distribution that is highly Non-Normal, the finite sample bias is larger than when the ϵ_i distribution is Normal. However, the finite sample bias increases for the IV estimator using the twin birth instruments directly as well, and the finite sample bias is still smaller for the predicted fertility instruments IV than the twin births instruments IV. This simulation indicates that using a misspecified probit model to generate the instruments does not introduce any larger degree of finite sample bias relative to the more standard IV estimation using linear functions of the instruments.

Table B-1: Simulation Results

Distributional Assumption:	$\epsilon_i \sim N(0,1)$		$\epsilon_i \sim G(1,2)$	
True Parameters:	$\gamma_2 = 1$	$\gamma_3 = -1$	$\gamma_2 = 1$	$\gamma_3 = -1$
i) OLS				
Mean Absolute Value of Bias	1.193132	1.317498	.9635493	1.92922
ii) IV using Twin Instr. Directly				
Mean Absolute Value of Bias	.0719994	.1219983	.2543027	.3152533
Standard Deviation of Estimates	.0902498	.152864	.3163337	.3943571
Mean Squared Error	0.0163	0.0467	0.2000	0.3108
iii) IV using Pred. Fertility Instr.				
Mean Absolute Value of Bias	.0455519	.0567577	.086395	.102743
Standard Deviation of Estimates	.0632768	.0632768	.1078351	.1279812
Mean Squared Error	0.0064	0.0089	0.0233	0.0329

Notes: Simulation results from 500 replications of the data generating process described above.

C First Stage Results

Table C-1: First Stage for Table 6

	Twin2	Twin3	Twin4	Twin5
Sample: Fam. Size $\geq$ 2, 1st Birth				
Column (2)	0.684 (0.012)			
Column (3)		0.760 (0.0162)		
Column (4)			0.787 (0.027)	
Column (5)				0.753 (0.0501)
Column (6)	0.688 (0.012)	0.781 (0.016)	0.803 (0.027)	0.759 (0.050)
Sample: Fam. Size $\geq$ 3, 2nd Birth				
Column (3)		0.763 (0.014)		
Column (4)			0.791 (0.023)	
Column (5)				0.753 (0.043)
Column (6)		0.771 (0.014)	0.823 (0.023)	0.763 (0.042)
Sample: Fam. Size $\geq$ 4, 3rd Birth				
Column (4)			0.786 (0.019)	
Column (5)				0.754 (0.035)
Column (6)			0.795 (0.019)	0.783 (0.035)

Notes: Each row reports the first stage estimate of number of children on twin birth instrument for the indicated column from Table 6. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are robust to heteroskedasticity but clustering is not necessary given that each regression includes only 1 child from each family.

Source: See Table 6.

Table C-2: First Stage for Table 8

	$pr(s_i \geq 2)$	$pr(s_i \geq 3)$	$pr(s_i \geq 4)$	$pr(s_i \geq 5)$
Sample: Fam. Size ≥ 2, 1st Birth				
Column (2)	1.2995942 (.01599981)			
Column (3)		1.4151114 (.01984729)		
Column (4)			2.2627584 (.0357567)	
Column (5)				3.7423315 (.05568987)
Column (6)	1.2822086 (.01553299)	1.1723997 (.02292579)	1.8199337 (.04494965)	2.915029 (.08532851)
Sample: Fam. Size ≥ 3, 2nd Birth				
Column (3)		1.2290325 (.0138201)		
Column (4)			1.6251123 (.02673695)	
Column (5)				2.7159697 (.04455262)
Column (6)		1.159176 (.01349736)	1.4149385 (.03103561)	2.2388828 (.06047967)
Sample: Fam. Size ≥ 4, 3rd Birth				
Column (4)			1.1570436 (.01603174)	
Column (5)				1.5215385 (.0348386)
Column (6)			1.1275344 (.01596862)	1.4585248 (.03798706)

Notes: Each row reports the first stage estimate of number of children on twin birth instrument for the indicated column from Table 8. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are robust to heteroskedasticity but clustering is not necessary given that each regression includes only 1 child from each family.

Source: See Table 8.

Table C-3: First Stage for Table 10

	First Stage				
	Twin2	Twin3	Twin4	Twin5	F-Stat
Column (2):					
Siblings ≥ 2	0.518	0.010	-0.013	-0.005	344.39
	(0.007)	(0.009)	(0.015)	(0.028)	
Siblings ≥ 3	0.127	0.661	0.018	-0.017	358.21
	(0.005)	(0.018)	(0.011)	(0.021)	
Siblings ≥ 4	0.033	0.092	0.708	0.024	225.58
	(0.003)	(0.004)	(0.007)	(0.013)	
Siblings $= 5$	0.010	0.017	0.090	0.756	141.78
	(0.002)	(0.002)	(0.003)	(0.007)	
Column (4):					
Siblings ≥ 3		0.648	0.020	-0.016	200.43
		(0.009)	(0.014)	(0.027)	
Siblings ≥ 4		0.103	0.703	0.023	154.65
		(0.006)	(0.009)	(0.178)	
Siblings $= 5$		0.020	0.100	0.756	90.56
		(0.003)	(0.005)	(0.010)	
Column (6):					
Siblings ≥ 4			0.693	0.026	71.56
			(0.014)	(0.026)	
Siblings $= 5$			0.101	0.757	47.60
			(0.008)	(0.015)	
Column (8):					
Siblings $= 5$				0.752	18.87
				(0.025)	

Notes: Each panel reports the first stage estimate of the family size indicators on twin birth instruments for the indicated column from Table 10. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are robust to heteroskedasticity but clustering is not necessary given that each regression includes only 1 child from each family.

Source: See Table 10.

Table C-4: First Stage for Table 11

| | First Stage | | | | |
	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	F-Stat
Column (1):					
Siblings ≥ 2	1.012	0.134	0.173	0.191	350.03
	(0.013)	(0.013)	(0.021)	(0.040)	
Siblings ≥ 3	0.209	0.960	0.459	0.550	377.88
	(0.010)	(0.010)	(0.016)	(0.031)	
Siblings ≥ 4	0.048	0.079	1.08	0.874	269.54
	(0.006)	(0.006)	(0.009)	(0.018)	
Siblings $= 5$	0.014	-0.001	0.112	1.30	184.93
	(0.003)	(0.003)	(0.005)	(0.009)	
Column (2):					
Siblings ≥ 3		1.01	0.289	0.361	209.60
		(0.0133)	(0.019)	(0.036)	
Siblings ≥ 4		0.131	1.01	0.681	173.58
		(0.009)	(0.013)	(0.024)	
Siblings $= 5$		0.018	0.114	1.20	111.53
		(0.005)	(0.007)	(0.013)	
Column (3):					
Siblings ≥ 4			1.016	0.313	74.26
			(0.013)	(0.019)	
Siblings $= 5$			0.140	1.06	52.16
			(0.009)	(0.013)	
Column (4):					
Siblings $= 5$				1.00	19.18
				(0.033)	

Notes: Each panel reports the first stage estimate of the family size indicators on predicted fertility instruments for the indicated column from Table 11. All models include covariates for gender, child's age (in 2000), mother's age (in 2000), father's age (in 2000), mother's education, and father's education. Standard errors in parentheses are heteroskedastic robust but clustering is not necessary given that each regression includes only 1 child from each family.

Source: See Table 11.