Testing the validity of the sibling sex ratio instrument

Martin Huber

September 1, 2014

University of Fribourg, Dept. of Economics

Abstract: We test the validity of the sibling sex ratio instrument in Angrist and Evans (1998) using the methods proposed by Kitagawa (2008) and Huber and Mellace (2014). The sex ratio of the first two siblings is arguably randomly assigned and influences the probability of having a third child, which makes it a candidate instrument for fertility when estimating the effect of fertility on female labor supply. However, identification hinges on the random assignment of the instrument, an instrumental exclusion restriction, and the monotonicity of fertility in the instrument, see Imbens and Angrist (1994). We find that the instrumental variable tests of Kitagawa (2008) and Huber and Mellace (2014) do not point to a violation of these assumptions in the Angrist and Evans (1998) data (which can, however, not be ruled out even asymptotically as the tests cannot detect all possible violations).

Keywords: instrumental variable, sibling sex ratio, treatment effects, LATE, tests.

JEL classification: C12, C21, C26, J13, J22.

I have benefited from comments by Joe Altonji, Josh Angrist, Clément De Chaisemartin, Xavier D'Haultfoeuille, Michael Lechner, Giovanni Mellace, Herman van Dijk, Anna Sjögren, seminar participants in Munich (July 2012), Uppsala (August 2012), St. Gallen (Oct 2012), Bern (Oct 2012), and Bergen (Nov 2012), and one anonymous referee. Financial support from the Swiss National Science Foundation grant PBSGP1_138770 is gratefully acknowledged. Address for correspondence: Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland, martin.huber@unifr.ch.

1 Introduction

Instrumental variable (IV) estimation is a corner stone of causal inference in empirical economics. The idea of exploiting a variable that is correlated with an endogenous predictor but does itself not have a direct effect on the outcome of interest goes probably back to Wright (1928). In heterogeneous treatment effect models, which is the framework considered in this paper, an instrument is valid if it fulfills two restrictions: Firstly, it must be independent of the joint distribution of potential treatment states and potential outcomes, which implies the random assignment of the instrument and the satisfaction of an exclusion restriction on the outcome. Secondly, the treatment state has to vary with the instrument in a weakly monotonic manner in the population. Under these assumptions, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) show that the local average treatment effect (LATE) on the subpopulation of compliers, whose treatment states react on the instrument in the intended way, is identified.

Up to date, the plausibility of IV validity has been predominantly discussed on theoretical bases, at best supported by reasoning based on empirical findings. We refer to Murray (2006) for an excellent discussion on how to subject candidate instruments to intuitive, empirical, and theoretical scrutiny to assess their validity. In contrast, statistical testing has played little role, at least in set ups where the number of instruments equals the number of endogenous variables.[1] However, many instruments are far from being undisputed because their validity is not unanimously accepted among researchers. Rosenzweig and Wolpin (2000) prominently discuss the potential pitfalls of what they call natural "natural experiments", i.e., instruments that are arguably randomly assigned by nature, such as quarter of birth (see Angrist and Krueger (1991)) or twin births (e.g., Ashenfelter and Krueger (1994)).

One influential instrument challenged in their paper is the sex ratio of the first two siblings proposed by Angrist and Evans (1998) to estimate the female labor supply effect of fertility.[2]

[1] In contrast, tests for IV validity are available for overidentified models where the number of instruments exceeds the number of endogenous regressors. Sargan (1958) was the first to propose such a test for the linear IV model with homogenous effects. However, testing is only consistent if at least one instrument is known to be valid.

[2] The sibling sex ratio instrument has subsequently also been used in Iacovou (2001), Conley and Glauber (2005), Goux and Maurinc (2005), Cruces and Galiani (2007), and Angrist, Lavy, and Schlosser (2010), among others.

Under the presumption that parents have a preference for mixed sex children, the idea is that having two children of the same sex, which is arguably randomly assigned by nature, increases the chances of getting a third child. However, Rosenzweig and Wolpin (2000) argue that having mixed sex siblings may violate the exclusion restriction by directly affecting both the marginal utility of leisure and child rearing costs and, thus, labor supply. Furthermore, based on a data set from rural India, they also provide empirical evidence that expenses for clothing of the third born are significantly lower if the older siblings are of the same sex. It is unclear to which extent this issue carries over to the US data of Angrist and Evans (1998) and whether it is important enough to affect the labor market behavior of females. In contrast to Rosenzweig and Wolpin (2000), Bütikofer (2010) finds no crucial differences in the household economies of scale across families with different sibling sex composition due to cloth- and room-sharing for arguably more comparable countries such as the UK and Switzerland (but also Mexico).

As the second identifying assumption, Angrist and Evans (1998) rely on monotonicity due to the parents' arguable preference for mixed sex siblings. Indeed, parental preferences for a balanced sex composition are well documented for the US, see for instance Ben-Porath and Welch (1976). However, monotonicity fails if a subgroup of parents has a preference for at least two children of the same sex and chooses to have a third child if the first two are of mixed sex. That the latter case may be empirically relevant is for instance corroborated by Lee (2008), who finds that South Korean parents with one son and one daughter are more likely to continue childbearing than parents with two sons. Rosenzweig and Wolpin (2000) observe a similar pattern in their Indian data set, a country well known for its sex bias among parents. While preferences arguably differ strongly across countries, defiers may even exist in the US. Dahl and Moretti (2008) find that the number of children is significantly higher in US families with a first-born girl suggesting a preference for having boys. Yet, this does not directly test the violation of monotonicity. Therefore, even though both the exclusion restriction and monotonicity may be questioned given these theoretical and empirical findings, they can neither be unambiguously refuted nor approved based on the arguments brought forward in the scientific discussion.

The contribution of this paper is to for the first time investigate the validity of the sibling sex ratio instrument of Angrist and Evans (1998) based on hypothesis tests. To this end, we apply recently developed methods for testing IV validity in heterogenous treatment effect models proposed by Kitagawa (2008) and Huber and Mellace (2014). The tests share the feature that asymptotic non-rejection does not imply IV validity (i.e., they do not have power against all violations and are therefore conservative), whereas asymptotic rejection points to an invalid instrument. We apply the tests to the full sample as well as to subsamples conditional on covariates such as education, age, marital status, race, and the year of the first birth. Finally, we also split the instrument to separately consider two first born sons and two first born daughters vs. mixed sex siblings in order to tackle potential violations of monotonicity due to a preference for boys. In general, our tests do not point to a violation of IV validity in any scenario, because they are mostly insignificant at any conventional level. Only in few of the many cases considered, p-values below 10% are obtained, but again, this does not necessarily imply that the instrument is valid.

The remainder of this paper is organized as follows. Section 2 reviews the IV assumptions allowing for LATE identification, as well as several methods for testing their violation. Section 3 applies the tests to data from Angrist and Evans (1998). Section 4 concludes.

2 Identifying assumptions and testing

Suppose that we are interested in the average effect of a binary "treatment" $D \in \{1, 0\}$ (e.g., having a third child) on an outcome Y (e.g., labor market participation) evaluated at some point in time after the treatment. Under endogeneity, the effect of D is confounded with unobserved factors that affect both the treatment and the outcome. Assume that there exists a binary instrument, denoted by Z (e.g. having same sex children), that is correlated with the treatment but does not have a direct effect on the outcome (i.e., any impact other than through the treatment). Denote by $D(z)$ the potential treatment state for instrument $Z = z$, and by $Y(d)$ the potential outcome for treatment $D = d$ (see for instance Rubin, 1974, for a discussion of the potential outcome

notation). For each subject, only one of the two potential outcomes and treatment states are observed, because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ and $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$.

As discussed in Angrist, Imbens, and Rubin (1996), the population can be categorized into four types (denoted by $T \in \{a, c, d, n\}$), depending on how the treatment state changes with the instrument. The compliers ($c : D(1) = 1, D(0) = 0$) react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. The always takers ($a : D(1) = 1, D(0) = 1$) are always treated irrespective of the instrument state, the never takers ($n : D(1) = 0, D(0) = 0$) are never treated, and the defiers ($d : D(0) = 1, D(0) = 1$) only take the treatment when $Z = 0$. The types are not identified from the data, because any observed treatment-instrument-combination either reveals $D(1)$ or $D(0)$, so that any observation may belong to one of two candidate types. For instance, a subject with $D = 1$, $Z = 1$ may either be a complier or an always taker, as both types satisfy $D(1) = 1$.

To demonstrate how the IV assumptions of Imbens and Angrist (1994) solve the identification problem of unidentified types and which testable restrictions this implies, we first introduce some notation which heavily borrows from Kitagawa (2009). Let $f(y, D = d|Z = z)$ denote the (observed) joint density of the observed outcome and $D = d$ conditional on $Z = z$ for $d, z \in \{1, 0\}$. Furthermore, denote by $f(y(d), T = t|Z = z)$ the unobserved joint density of the potential outcome and type t conditional on $Z = z$, where $t \in \{a, c, d, n\}$. Note that the subsequent relationships between observed and unobserved joint densities hold for all y in the support of Y:

$$f(y, D = 1|Z = 1) = f(y(1), T = c|Z = 1) + f(y(1), T = a|Z = 1), \quad (1)$$

$$f(y, D = 1|Z = 0) = f(y(1), T = d|Z = 0) + f(y(1), T = a|Z = 0), \quad (2)$$

$$f(y, D = 0|Z = 1) = f(y(0), T = d|Z = 1) + f(y(0), T = n|Z = 1), \quad (3)$$

$$f(y, D = 0|Z = 0) = f(y(0), T = c|Z = 0) + f(y(0), T = n|Z = 0). \quad (4)$$

Equations (1) to (4) follow from the fact that without further assumptions, a particular observed treatment-instrument-combination is consistent with two types, so that any observed joint density

is a function of the potential outcomes of two different types conditional on Z.

Imbens and Angrist (1994) have shown that the LATE on the compliers is point identified when imposing the following identifying assumptions:

Assumption 1:

$Z \perp (D(1), D(0), Y(1), Y(0))$ (joint independence),

where "$\perp$" denotes independence. Assumption 1 implies the randomization of the instrument (so that it is unrelated with factors affecting the treatment and/or outcome) and the exclusion of direct effects on the outcome.

Assumption 2:

$\Pr(D(1) \geq D(0)) = 1$ (monotonicity).

Assumption 2 says that the potential treatment state of any individual does either not decrease (positive monotonicity) in the instrument. Therefore, the existence of defiers (type d) is ruled out because for the latter type, $D(1) < D(0)$.

Under Assumption 1, $f(y(d), T = t|Z = 1) = f(y(d), T = t|Z = 0) = f(y(d), T = t)$ for any type and treatment state, otherwise the potential treatment states and/or potential outcomes were not independent of the instrument (e.g., due to a direct effect of the instrument on the outcome). Under Assumption 2, $f(y(1), T = d)$, and $f(y(0), T = d)$ are equal to zero. Therefore, equations (1) to (4) simplify to

$$f(y, D = 1|Z = 1) = f(y(1), T = c) + f(y(1), T = a), \quad (5)$$

$$f(y, D = 1|Z = 0) = f(y(1), T = a), \quad (6)$$

$$f(y, D = 0|Z = 1) = f(y(0), T = n), \quad (7)$$

$$f(y, D = 0|Z = 0) = f(y(0), T = c) + f(y(0), T = n). \quad (8)$$

By subtracting (6) from (5) and (7) from (8), the joint densities of the compliers under treatment

and non-treatment are identified:[3]

$$f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0) = f(y(1), T = c), \tag{9}$$

$$f(y, D = 0|Z = 0) - f(y, D = 0|Z = 1) = f(y(0), T = c). \tag{10}$$

To see how this permits the identification of the LATE on the compliers, first note that $\int f(y(d), T = c)dy = \pi_c$, where $\pi_c = \Pr(T = c)$ denotes the share of compliers in the population (and more generally, $\pi_t = \Pr(T = t)$ will henceforth denote the share of type t). It follows that π_c is identified by

$$\begin{aligned} \pi_c &= \int [f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0)]dy \\ &= \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0) = E(D|Z = 1) - E(D|Z = 0). \end{aligned} \tag{11}$$

Furthermore, $\int y[f(y(d), T = c)]dy = \int y[f(y(d)|T = c)]\pi_c dy = E[Y(d)|T = c] \cdot \pi_c$ implies that

$$\begin{aligned} & E[Y(1) - Y(0)|T = c] \cdot \pi_c \\ = & \int y\{[f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0)] - [f(y, D = 0|Z = 0) - f(y, D = 0|Z = 1)]\}dy \\ = & \int y[f(y|Z = 1) - f(y|Z = 0)]dy = E(Y|Z = 1) - E(Y|Z = 0), \end{aligned}$$

which is the intention to treat effect (ITT). By scaling the ITT by the share of compliers we obtain the standard identification result for the LATE (denoted by Δ_c), which corresponds to the probability limit of the Wald estimator:

$$E[Y(1) - Y(0)|T = c] = \Delta_c = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}. \tag{12}$$

(9) and (10) not only entail the identification of the LATE, but also provide testable implica-

[3]From the subsequent discussion it is easy to see that also the conditional densities $f(y(d)|T = c) = f(y(d), T = c)/\Pr(T = c)$ are identified, which was first acknowledged by Imbens and Rubin (1997).

tions of the identifying assumptions. For all y in the support of Y, it must hold that

$$f(y, D = 1|Z = 1) \geq f(y, D = 1|Z = 0), \quad f(y, D = 0|Z = 0) \geq f(y, D = 0|Z = 1), \tag{13}$$

otherwise the joint densities of the compliers would be negative, even though a density cannot be smaller than zero. Therefore, if one or both of the weak inequalities in (13) are violated, either IV independence as postulated in Assumption 1 does not hold, or the defiers stochastically dominate the compliers so that monotonicity is not satisfied, or both. The constraints in (13) were first derived by Balke and Pearl (1997) (at that time, for binary outcomes).

For testing, (13) should be verified at each value y in the support of Y. However, if the outcome is of rich support (e.g., continuous), finite sample power may be higher when partitioning the support into a finite number of subsets. The constraints then become

$$\begin{aligned} &\Pr(Y \in A, D = 1|Z = 1) \geq \Pr(Y \in A, D = 1|Z = 0), \\ &\Pr(Y \in A, D = 0|Z = 0) \geq \Pr(Y \in A, D = 0|Z = 1), \end{aligned} \tag{14}$$

where A denotes a subset of the support of Y, and (14) must hold for any A. Kitagawa (2008) proposes a test of these constraints that is based a two sample Kolmogorov-Smirnov-type statistic on the supremum of $\Pr(Y \in A, D = 1|Z = 0) - \Pr(Y \in A, D = 1|Z = 1)$ and $\Pr(Y \in A, D = 0|Z = 1) - \Pr(Y \in A, D = 0|Z = 0)$, respectively, across subsets A. As the statistic does not converge to any known distribution, he suggests a bootstrap (or more concisely, a permutation) method for inference which is similar to Abadie (2002).

Huber and Mellace (2014) suggest an alternative set of testable constraints on mean potential outcomes rather than densities or probability measures defined by A. They for instance show that the LATE assumptions imply the following restrictions related to the mean outcome under treatment of the always takers:

$$E(Y|D = 1, Z = 1, Y \leq y_q) \leq E(Y|D = 1, Z = 0) \leq E(Y|D = 1, Z = 1, Y \geq y_{1-q}), \tag{15}$$

where $q = \frac{\pi_a}{\pi_a+\pi_c} = \frac{\Pr(D=1|Z=0)}{\Pr(D=1|Z=1)}$ denotes the share of always takers among those with $D = 1$ and $Z = 1$, i.e., in the mixed population of compliers and always takers, and y_q is the qth quantile of Y given $D = 1$ and $Z = 1$. To see this, first note that $E(Y|D = 1, Z = 0)$ point identifies the mean potential outcome of the always takers under treatment, as any observation with $D = 1$, $Z = 0$ must be an always taker in the absence of defiers. Secondly, the mean potential outcomes of the always takers are bounded by the averages in the upper and lower outcome proportions with $D = 1$ and $Z = 1$ that correspond to the share of the always takers in the mixed population: $E(Y|D = 1, Z = 1, Y \leq y_q)$, $E(Y|D = 1, Z = 1, Y \geq y_{1-q})$. Thirdly, $E(Y|D = 1, Z = 0)$ must lie within the latter bounds, otherwise the identifying assumptions are necessarily violated. An equivalent result applies to the never takers under non-treatment:

$$E(Y|D = 0, Z = 0, Y \leq y_r) \leq E(Y|D = 0, Z = 1) \leq E(Y|D = 0, Z = 0, Y \geq y_{1-r}), \tag{16}$$

with $r = \frac{\pi_n}{\pi_n+\pi_c} = \frac{\Pr(D=0|Z=1)}{\Pr(D=0|Z=0)}$ being the share of never takers those with $D = 0$ and $Z = 0$, the mixed population of never takers and compliers, and y_r denoting the rth quantile of Y given $D = 0$ and $Z = 0$. Huber and Mellace (2014) apply the minimum p-value-based bootstrap method of Bennett (2009) for testing the multiple inequality constraints in (15) and (16), see Appendix A.1 for the test algorithm.

The authors not only consider mean constraints, but also the following probability constraints which they derive using the results of Horowitz and Manski (1995):

$$\frac{\Pr(Y \in A|D = 1, Z = 1) - (1 - q)}{q} \leq \Pr(Y \in A|D = 1, Z = 0) \leq \frac{\Pr(Y \in A|D = 1, Z = 1)}{q}, \tag{17}$$

$$\frac{\Pr(Y \in A|D = 0, Z = 0) - (1 - r)}{r} \leq \Pr(Y \in A|D = 0, Z = 1) \leq \frac{\Pr(Y \in V|D = 0, Z = 0)}{r}. \tag{18}$$

It can be shown that the second inequalities in (17) and (18) are equivalent to (14). Furthermore, if testing is based on non-overlapping subsets A that jointly cover the entire support of Y, the

first inequalities in (17) are redundant. In this case, testing based on (17), (18) and (14) is asymptotically equivalent. In the application, we consider the Huber and Mellace (2014) tests based on both mean and probability constraints, as well as the Kitagawa (2008) method.

3 Empirical application

In this section, we apply the various tests to the 1980 wave of the Angrist and Evans (1998) data coming from the US Census Public Use Micro Samples (PUMS). Angrist and Evans (1998) aim at assessing the effect of fertility on female labor supply, but face the problem that fertility and labor supply decisions are most likely endogenous. For this reason, they suggest to use the sex ratio of the first two children as instrument for fertility, i.e., to get (at least) one more child. They justify the plausibility of Assumptions 1 and 2 by arguing that the sex of children should not have any direct effects on labor supply and that parents in the US have a preference for mixed sex siblings so that fertility is monotonic in the instrument.

As discussed in the introduction, the validity of the exclusion restriction has been challenged in Rosenzweig and Wolpin (2000) based on both theoretical arguments concerning fertility and labor supply choices and empirical data from rural India. They argue that mixed sex siblings may increase child rearing costs compared to same sex siblings for instance due to the absence of sex-specific hand-me-downs and can therefore affect labor market behavior through other channels than fertility. Secondly, also the monotonicity assumption, which has to hold for every unit in the population, appears suspicious given the empirical evidence on the preference for boys not only outside, see Lee (2008), but also within the US, see Dahl and Moretti (2008). These concerns motivate the application of the Kitagawa (2008) and Huber and Mellace (2014) tests to the sibling sex ratio instrument.

The 1980 PUMS evaluation sample of Angrist and Evans (1998) contains 394,840 observations of mothers who were aged between 21 and 35 years, had been 15 or older when giving birth for the first time, and had at least two children (with the second being at least one year old). The

treatment dummy D indicates whether a mother has three or more children (158,751 observations or 40.21%) or just two (236,089 observations or 59.79%). The instrument Z is one if the first two children have the same sex (199,548 or 50.54%) and zero otherwise (195,292 observations or 49.46%). The first stage regression of D on Z is highly significant and suggests that if Assumptions 1 and 2 hold, same sex increases the probability of a third child by roughly 6 percentage points. The outcome Y is defined as hours worked per week. Two stage least squares (TSLS), which is equivalent to the Wald estimator in the binary treatment and instrument case, predicts a reduction of -5.187 hours among compliers which is again highly significant (robust standard error: 1.006). When conditioning on the covariates age, education, marital status, race, and year of the first birth in the TSLS regression, the effect is somewhat less negative (-4.456) but still quite similar to and not significantly different from the unconditional estimate.

Table 1: P-values when testing the sibling sex ratio instrument

	mean test	HM11 (5)	K08 (5)	HM11 (12)	K08 (12)
full sample (394,840)	0.580	0.638	0.984	0.257	0.755
edu<12 (88,764)	1.000	0.778	0.997	0.258	0.766
edu=12 (189,818)	1.000	0.070	0.971	0.235	1.000
edu>12 (116,258)	1.000	0.961	0.996	0.954	1.000
edu<12, a. 21-25, married, white (11,716)	0.501	0.541	0.993	0.944	1.000
edu=12, a. 21-25, married, white (16,249)	0.520	0.092	0.405	0.182	0.745
edu>12, a. 21-25, married, white (3,232)	0.822	1.000	1.000	0.521	0.879
edu<12, a. 26-30, married, white (19,864)	0.617	0.030	0.941	0.355	0.979
edu=12, a. 26-30, married, white (53,919)	1.000	0.164	0.369	0.973	1.000
edu>12, a. 26-30, married, white (28,266)	1.000	0.304	0.992	0.741	1.000
edu<12, a. 31-35, married, white (23,367)	1.000	0.234	0.960	0.754	1.000
edu=12, a. 31-35, married, white (75,850)	1.000	0.368	0.982	0.394	1.000
edu>12, a. 31-35, married, white (58,684)	1.000	0.674	0.989	0.767	1.000
edu<12, a. 21-25, mar., w., 1st birth 70-72 (2,784)	0.604	0.426	0.958	0.359	0.913
edu=12, a. 21-25, mar., w., 1st birth 70-72 (1,713)	0.477	1.000	0.931	0.774	0.556
edu>12, a. 21-25, mar., w., 1st birth 70-72 (286)	0.207	0.097	0.303	1.000	0.237
edu<12, a. 26-30, mar., w., 1st birth 70-72 (8,592)	0.704	0.105	0.955	0.406	0.926
edu=12, a. 26-30, mar., w., 1st birth 70-72 (24,492)	1.000	0.277	0.992	0.429	0.997
edu>12, a. 26-30, mar., w., 1st birth 70-72 (9,149)	0.703	0.147	0.956	0.590	0.927
edu<12, a. 31-35, mar., w., 1st birth 70-72 (2,389)	0.624	1.000	0.965	0.853	0.981
edu=12, a. 31-35, mar., w., 1st birth 70-72 (15,913)	1.000	0.474	0.980	0.631	0.992
edu>12, a. 31-35, mar., w., 1st birth 70-72 (20,037)	1.000	0.502	0.988	0.786	0.999

Note: The p-values of the tests are based on 999 bootstrap draws.

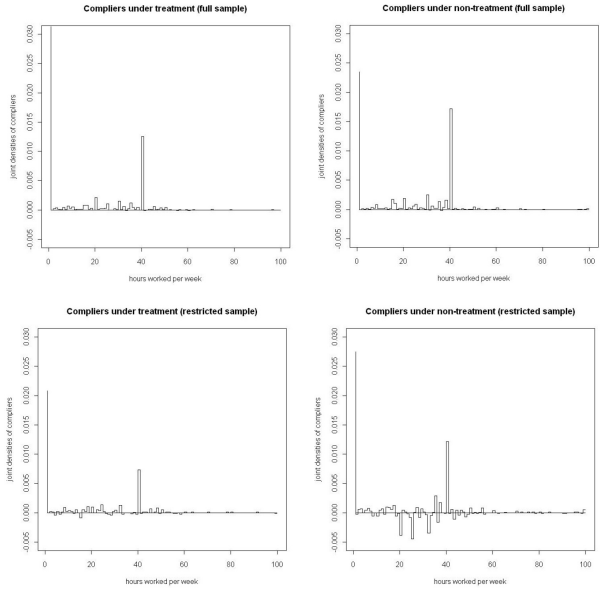
Table 1 presents the p-values of the tests for various sample definitions, i.e., for the full sample as well as for subsamples defined upon the values of the covariates. The second column (mean

test) provides the results for the mean test of Huber and Mellace (2014) based on (15) and (16). Columns 3 and 4 (HM11 (5) and K08 (5)) contain the p-values of the probability-based tests of Huber and Mellace (2014) using (17) and (18) and of Kitagawa (2008) using (14). In both tests, the support of the outcome is partitioned into five subsets A, which are defined by an equidistant grid from the minimum to the maximum of the observed outcome. Columns 5 and 6 (HM11 (12) and K08 (12)) present the results for the same tests, however, based on 12 subsets A set up in the following way: $A_1 = [0], A_2 = (0, 5], A_3 = (5, 10], A_4 = (10, 15], ..., A_{11} = (45, 50], A_{12} = (50, 100]$. As the probability mass of y has a spike at zero (as many women do not work at all), zero is now treated as an own category. Furthermore, only few women state to work more than 50 hours per week, which motivates the large range of the last subset. Several other numbers and definitions of A were also considered, but did not change the conclusions to be drawn and are therefore not reported.

In the full sample, no test provides evidence against the validity of the IV assumptions. As mentioned in Section 1, this does not imply that Assumption 1 and 2 hold, because the tests are inconsistent in the sense that they cannot detect every possible violation. It therefore seems advisable to consider testing conditional on covariates, too, as one potential source of non-detection is that deviations from the null average out in the full sample, while they might be detected in subsamples. We start by splitting the sample into three levels of education: less than 12 years, 12s year of education (which corresponds to completed high school education), and more than 12 years (higher education). No test rejects the null at the 5% level and only for 12 years of education, HM11 (5) is significant at the 10% level.[4] Partitioning the sample further by education, age brackets (21-25, 26-30, and 31-35), marital status, race, and year of the first birth does not crucially change the picture. In only four out of a total of 22 samples, HM11 (5) rejects the IV assumptions at the 10% level (thereof once at 5%), while the other tests are never significant.

[4]In general, we would expect the probability-based test of Huber and Mellace (2014), which pre-estimates the constraints that are (close to being) violated, to be more powerful than the Kitagawa (2008) test, which depends on the least favorable null, and we see this pattern in our application.

Figure 1: Estimates of $f(y(1), T = c)$ and $f(y(0), T = c)$



For illustrative purposes, Figure 1 graphically shows the violations in the full sample (upper graphs) and conditional on age 21-25, 12 years of education, white, and married (lower graphs). For each value of the outcome, the estimates of the compliers' densities $f(y(1), T = c)$, see (9), and $f(y(0), T = c)$, see (10), are displayed. As discussed in Section 2, a necessary condition for IV validity is that the estimated densities of the compliers are asymptotically non-negative. While violations are only minor in the full sample, they are more pronounced in the restricted sample (in particular among the non-treated), so that the statistic of HM11 (5) is significant at the 10% level.

In a next step, we split instrument into two variables for boys and girls. I.e., we separately analyze two overlapping samples, the first one consisting of mothers with either two sons or mixed sex siblings (299,419 observations, of which 34.8% have two boys), the second one of mothers with either two daughters or mixed sex siblings (290,713 observations, of which 32.8% have two girls). To see the usefulness of this approach, assume that monotonicity only holds for girls (i.e., all parents prefer mixed sex siblings over two girls) but not for boys (i.e., some parents prefer two boys over mixed sex siblings) as one might suspect from the findings in Lee (2008) and Dahl and

Table 2: Two boys and two girls instruments

	Mean test	HM11 (5)	K08 (5)	HM11 (12)	K08 (12)
two girls vs. mixed					
full sample (290,713)	1.000	0.180	0.980	0.803	1.000
edu<12 (65,597)	1.000	0.888	0.988	0.601	1.000
edu=12 (139,674)	1.000	0.175	0.976	0.258	0.998
edu>12 (85,442)	1.000	0.194	0.990	0.614	1.000
edu<12, age 21-25, married, white (8,662)	0.515	0.057	0.983	0.956	0.997
edu=12, age 21-25, married, white (11,892)	0.709	0.086	0.200	0.001	0.437
edu>12, age 21-25, married, white (2,368)	0.638	0.957	0.989	1.000	0.909
edu<12, age 26-30, married, white (14,664)	1.000	0.032	0.958	0.934	1.000
edu=12, age 26-30, married, white (39,503)	1.000	0.184	0.984	0.627	0.999
edu>12, age 26-30, married, white (20,779)	1.000	0.453	0.992	0.977	1.000
edu<12, age 31-35, married, white (17,142)	1.000	0.471	0.996	0.987	1.000
edu=12, age 31-35, married, white (55,751)	1.000	0.529	0.987	0.440	1.000
edu>12, age 31-35, married, white (42,982)	1.000	0.191	0.983	0.430	0.986
edu<12, a. 21-25, mar., w., 1st birth 70-72 (2,064)	0.544	1.000	0.869	0.548	0.899
edu=12, a. 21-25, mar., w., 1st birth 70-72 (1,267)	0.576	1.000	0.905	0.981	0.773
edu>12, a. 21-25, mar., w., 1st birth 70-72 (204)	0.302	0.721	0.874	1.000	0.417
edu<12, a. 26-30, mar., w., 1st birth 70-72 (6,330)	0.703	0.256	0.970	0.952	0.997
edu=12, a. 26-30, mar., w., 1st birth 70-72 (17,903)	1.000	0.611	0.993	0.119	0.958
edu>12, a. 26-30, mar., w., 1st birth 70-72 (6,736)	1.000	0.492	0.990	0.207	0.767
edu<12, a. 31-35, mar., w., 1st birth 70-72 (1,762)	0.857	1.000	0.956	0.300	0.807
edu=12, a. 31-35, mar., w., 1st birth 70-72 (11,643)	0.920	0.910	1.000	0.402	0.983
edu>12, a. 31-35, mar., w., 1st birth 70-72 (14,690)	1.000	0.430	0.966	0.722	0.949
two boys vs. mixed					
full sample (299,419)	1.000	0.502	0.992	0.454	1.000
edu<12 (67,183)	1.000	0.594	0.989	0.961	1.000
edu=12 (143,863)	1.000	0.094	0.976	0.773	1.000
edu>12 (88,373)	1.000	0.681	0.996	0.867	1.000
edu<12, age 21-25, married, white (8,835)	0.522	0.540	0.989	0.688	0.999
edu=12, age 21-25, married, white (12,283)	0.495	0.513	0.852	0.612	0.803
edu>12, age 21-25, married, white (2,442)	0.677	1.000	0.910	0.530	0.931
edu<12, age 26-30, married, white (15,072)	0.650	0.063	0.963	0.328	0.957
edu=12, age 26-30, married, white (40,946)	1.000	0.338	0.994	0.590	0.962
edu>12, age 26-30, married, white (21,543)	0.511	0.352	0.981	0.376	1.000
edu<12, age 31-35, married, white (17,882)	0.521	0.089	0.931	0.648	0.999
edu=12, age 31-35, married, white (57,592)	1.000	0.536	0.987	0.948	1.000
edu>12, age 31-35, married, white (44,665)	1.000	0.182	0.972	0.374	0.985
edu<12, a. 21-25, mar., w., 1st birth 70-72 (2,068)	0.704	0.595	0.947	0.030	0.923
edu=12, a. 21-25, mar., w., 1st birth 70-72 (1,294)	0.577	0.142	0.990	1.000	0.594
edu>12, a. 21-25, mar., w., 1st birth 70-72 (226)	0.122	0.103	0.178	1.000	0.357
edu<12, a. 26-30, mar., w., 1st birth 70-72 (6,491)	0.653	0.186	0.944	0.444	0.925
edu=12, a. 26-30, mar., w., 1st birth 70-72 (18,685)	1.000	0.348	0.997	0.570	0.987
edu>12, a. 26-30, mar., w., 1st birth 70-72 (7,022)	0.687	0.148	0.844	0.822	0.994
edu<12, a. 31-35, mar., w., 1st birth 70-72 (1,817)	0.580	1.000	0.861	0.800	0.824
edu=12, a. 31-35, mar., w., 1st birth 70-72 (12,076)	1.000	0.232	0.960	0.840	0.970
edu>12, a. 31-35, mar., w., 1st birth 70-72 (15,265)	0.617	0.103	0.955	0.495	0.999

Note: The p-values of the tests are based on 999 bootstrap draws.

Moretti (2008). If the non-monotonicity in the sample with two boys caused IV violation, the tests should pass when using the instrument defined by having two girls and reject when using two boys. It is worth noting that the first stage is smaller for two boys (5.1 percentage points) than for two girls (6.9 percentage points). Assuming that the exclusion restriction holds, the smaller first stage of the former could either be due to a lower share of compliers (if monotonicity is satisfied), or to a larger share of defiers (if monotonicity is not satisfied), or both.

Table 2 presents the results of separate tests for the two girls and two boys instruments in the respective full samples and conditional on covariates. We find no heterogeneity in rejection rates that would point to one instrument being more valid than the other. For either instrument, the null is rejected just three times by HM11 (5) at the 10% level and once by HM11 (12) at the 5% level across the 22 samples. All other tests are insignificant throughout. We therefore conclude that all in all, testing does not point to a violation of Assumptions 1 and 2 in the data considered when using the sibling sex ratio instrument.

We once more stress that this does not necessarily imply that the instrument is valid. In fact, a non-rejection of the testable constraints may occur due to three reasons: (i) instrument validity (Assumptions 1 and 2) is indeed satisfied, (ii) instrument validity is violated, but the finite sample power of the tests is too low to reject the null even though the constraints do not hold, or (iii) instrument validity is violated, but the violation is small enough that the constraints hold, so that the tests fail to detect the problem even asymptotically. Considering the Huber and Mellace (2014) mean test, for instance, the latter occurs if the violation is so little that the point identified mean outcomes of the always takers under treatment and never takers under non-treatment still lie within their respective upper and lower bounds. Therefore, the asymptotic power of the tests is stronger, the smaller the complier population is, which implies narrower bounds for the always and never takers in the mixed populations.

As the complier share in our application is rather limited when taken at face value, one might conclude that the asymptotic power of the tests is high. However, Table 3, which for the full sample reports the estimates of the bounds and point identified mean outcomes among always and

Table 3: Estimates of the bounds and point identified means in (15) and (16) for the full sample

type	upper bound	lower bound	point identified mean
always takers	18.828	11.850	16.448
(standard errors)	(0.105)	(0.129)	(0.068)
never takers	22.579	17.796	20.478
(standard errors)	(0.081)	(0.082)	(0.055)

Note: The standard errors of the parameters (in parentheses) are estimated based on 999 bootstrap draws.

never takers in (15) and (16), respectively, provides a different picture. In fact, the widths of the bounds are not negligible, amounting to roughly 7 hours (18.828-11.850) for the always takers and 5 hours (22.579-17.796) for the never takers. In either case, the point identified mean is within its respective bounds. Therefore, while too little finite sample power is seemingly no issue given the rather small standard errors of the estimates (provided in parentheses in Table 3), the widths of the bounds imply that the Huber and Mellace (2014) mean test can (asymptotically) not detect violations which are not comparably large. For this reason, it appears advisable to investigate the robustness of the estimated labor supply effects of fertility to (at least) mild violations of the IV assumptions. To this end, we refer to Huber (2014), who proposes sensitivity checks for the LATE under the non-satisfaction of the IV exclusion restriction or monotonicity and applies them to the same Angrist and Evans (1998) data.[5] His results suggest that the negativity of the impact of having a third child on females' working hours is robust to moderate deviations from either assumption.

4 Conclusion

In this paper, we have applied statistical tests to the sibling sex ratio instrument of Angrist and Evans (1998) to assess the identifying assumptions underlying the LATE framework. This novel approach complements earlier discussions arguing in favor or against the validity of the instrument which were mainly based on theoretical reasoning or empirically motivated back of the

[5] Alternative approaches for modelling violations of the IV exclusion restriction include Altonji, Elder, and Taber (2005), Choi and Lee (2012), Conley, Hansen, and Rossi (2012), Nevo and Rosen (2012), Flores and Flores-Lagunes (2013), and Mealli and Pacini (2013), while Richardson and Robins (2010) consider deviations from monotonicity.

envelope calculations rather than formal tests. Using the methods proposed in Kitagawa (2008) and Huber and Mellace (2014), testing did generally not refute the validity of the instrument. As a word of caution, this does not imply that the sibling sex ratio can be safely assumed to be a good instrument, because even asymptotically, the tests cannot find all possible violations of instrument validity if they are not large enough. Furthermore, it remains to be seen whether our findings can be replicated in other data sets.

A Appendix

A.1 The test algorithm based on Bennett (2009)

We denote by θ the vector of inequality moment constraints to be tested. E.g., when testing the mean constraints (15) and (16), the vector consists of four elements: $\theta = \begin{pmatrix} E(Y|D=1,Z=1,Y\leq y_q) - E(Y|D=1,Z=0) \\ E(Y|D=1,Z=0) - E(Y|D=1,Z=1,Y\geq y_{1-q}) \\ E(Y|D=0,Z=0,Y\leq y_r) - E(Y|D=0,Z=1) \\ E(Y|D=0,Z=1) - E(Y|D=0,Z=0,Y\geq y_{1-r}) \end{pmatrix}$. The null hypothesis is that all elements in θ are smaller or equal to zero. The algorithm of the minimum p-value-type test proposed by Bennett (2009) is as follows:

1. Estimate the vector of parameters $\hat{\theta}$ in the original sample.
2. Draw B_1 bootstrap samples of size n from the original sample.
3. In each bootstrap sample, compute the fully recentered vector $\tilde{\theta}_b^f \equiv \hat{\theta}_b - \hat{\theta}$ as well as the partially recentered vector $\tilde{\theta}_b^p \equiv \hat{\theta}_b - \max(\hat{\theta}, -\delta_n)$ for the mP.p test, where δ_n is a sequence such that $\delta_n \to 0$ and $\sqrt{n}\cdot\delta_n \to \infty$ as $n \to \infty$.[6]
4. Estimate the vector of p-values under full recentering, denoted by $P_{\tilde{\theta}^f}$:

$$P_{\tilde{\theta}^f} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n}\cdot\tilde{\theta}_b^f > \sqrt{n}\cdot\hat{\theta}\}. \tag{A.1}$$

5. Compute the minimum p-value under full recentering:

$$\hat{p}_f = \min(P_{\tilde{\theta}^f}). \tag{A.2}$$

[6] In the application, we choose $\delta_n = \sqrt{\frac{2\cdot\ln(\ln(n))}{n}}\cdot\hat{\sigma}_{\theta_i}, \quad i \in \{1,2,3,4\}$, where $\hat{\sigma}_{\theta_i}$ is the estimated (in the B_1 first stage bootstrap samples) standard deviation of the i-th inequality constraint, as suggested by Bennett (2009). It is, however, not guaranteed that this choice is optimal, see for instance the discussion in Donald and Hsu (2011).

6. Draw B_2 values from the distribution of $\tilde{\theta}_b^p$. We denote by $\tilde{\theta}_{b_2}^p$ the resampled observations in the second bootstrap.

7. In each bootstrap sample, compute the minimum p-value, denoted by $\hat{p}_{p,b_2}$:

$$\hat{p}_{p,b_2} = \min(P_{\tilde{\theta}^p,b_2}), \tag{A.3}$$

where

$$P_{\tilde{\theta}^p,b_2} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \tilde{\theta}_{b_2}^p\}. \tag{A.4}$$

8. Compute the p-value of the test as the share of bootstrapped minimum p-values that are smaller than or equal to the minimum p-value of the original sample:

$$\hat{p}_{\text{test}} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{\hat{p}_{p,b_2} \leq \hat{p}_f\}. \tag{A.5}$$

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," *The Journal of Human Resources*, 40, 791–821.

Angrist, J., and W. Evans (1998): "Children and their parents labor supply: Evidence from exogeneous variation in family size," *American Economic Review*, 88, 450–477.

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

Angrist, J., and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 979–1014.

Angrist, J., V. Lavy, and A. Schlosser (2010): "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," *Journal of Labor Economics*, 28, 773–824.

Ashenfelter, O., and A. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 84, 1157–74.

Balke, A., and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Ben-Porath, Y., and F. Welch (1976): "Do Sex Preferences Really Matter?," *The Quarterly Journal of Economics*, 90, 285–307.

Bennett, C. J. (2009): "Consistent and Asymptotically Unbiased MinP Tests of Multiple Inequality Moment Restrictions," *Working Paper 09-W08, Department of Economics, Vanderbilt University.*

Bütikofer, A. (2010): "Sibling Sex Composition and Cost of Children," *mimeo.*

Choi, J.-y., and M.-j. Lee (2012): "Bounding endogenous regressor coefficients using moment inequalities and generalized instruments," *Statistica Neerlandica*, 66, 161–182.

Conley, D., and R. Glauber (2005): "Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variations in Fertility," *NBER Working Paper No. 11302*.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly Exogenous," *Review of Economics and Statistics*, 94, 260–272.

Cruces, G., and S. Galiani (2007): "Fertility and female labor supply in Latin America: New causal evidence," *Labour Economics*, 14, 565–573.

Dahl, G. B., and E. Moretti (2008): "The Demand for Sons," *Review of Economic Studies*, 75, 1085–1120.

Donald, S. G., and Y.-C. Hsu (2011): "A New Test for Linear Inequality Constraints When the Variance Covariance Matrix Depends on the Unknown Parameters," *Economics Letters*, 113, 241–243.

Flores, C. A., and A. Flores-Lagunes (2013): "Partial Identification of Local Average Treatment Effects with an Invalid Instrument," *Journal of Business & Economic Statistics*, 31, 534–545.

Goux, D., and E. Maurinc (2005): "The effect of overcrowded housing on children's performance at school," *Journal of Public Economics*, 89, 797–819.

Horowitz, J. L., and C. F. Manski (1995): "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.

Huber, M. (2014): "Sensitivity checks for the local average treatment effect," *Economics Letters*, 123, 220–223.

Huber, M., and G. Mellace (2014): "Testing instrument validity for LATE identification based on inequality moment constraints," *forthcoming in the Review of Economics and Statistics*.

Iacovou, M. (2001): "Fertility and female labour supply," *ISER Working Paper Number 2001-19*.

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

Imbens, G. W., and D. Rubin (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64, 555–574.

Kitagawa, T. (2008): "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model," *mimeo*.

——— (2009): "Identification Region of the Potential Outcome Distribution under Instrument Independence," *CeMMAP working paper 30/09*.

Lee, J. (2008): "Sibling size and investment in childrens education: an Asian instrument," *Journal of Population Economics*, 21, 855–875.

Mealli, F., and B. Pacini (2013): "Using secondary outcomes and covariates to sharpen inference in instrumental variable settings," *Journal of the American Statistical Association*, 108, 1120–1131.

Murray, M. P. (2006): "Avoiding Invalid Instruments and Coping with Weak Instruments," *The Journal of Economic Perspectives*, 20, 111–132.

Nevo, A., and A. M. Rosen (2012): "Identification with Imperfect Instruments," *Review of Economics and Statistics*, 94, 659–671.

Richardson, T. S., and J. M. Robins (2010): "Analysis of the Binary Instrumental Variable Model," in *Heuristics, probability and causality: a tribute to Judea Pearl*, ed. by R. Dechter, H. Geffner, and J. Y. Halpern, pp. 415–440, London, UK. College Publications.

Rosenzweig, M. R., and K. I. Wolpin (2000): "Natural "Natural Experiments" in Economics," *Journal of Economic Literature*, 38, 827–874.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Sargan, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393–415.

Wright, P. (1928): *The Tariff on Animal and Vegetable Oils*. US Tariff Commission.