Universität St.Gallen

Statistical verification of a natural "natural experiment": Tests and sensitivity checks for the sibling sex ratio instrument

Martin Huber

August 2012 Discussion Paper no. 2012-19

School of Economics and Political Science,
Department of Economics

University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
School of Economics and Political Science
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25

Electronic Publication: Fax +41 71 224 31 35
http://www.seps.unisg.ch

Statistical verification of a natural "natural experiment": Tests and sensitivity checks for the sibling sex ratio instrument [1]

Martin Huber

Author's address:

Martin Huber, Ph.D.
SEW-HSG
Varnbüelstrasse 14
CH-9000 St. Gallen
Phone +41 71 2299
Fax +41 71 2302
Email martin.huber@unisg.ch
Website www.sew.unisg.ch

[1]I have benefited from comments by Joe Altonji, Josh Angrist, Clément De Chaisemartin, Xavier D'Haultfoeuille, Giovanni Mellace, Herman van Dijk, Anna Sjögren, and seminar participants in Munich (July 2012) and Uppsala (August 2012). Financial support from the Swiss National Science Foundation grant PBSGP1_138770 is gratefully acknowledged.

Abstract

This paper presents statistical evidence about the validity of the sibling sex ratio instrument proposed by Angrist and Evans (1998), a prominent natural "natural experiment" in the sense of Rosenzweig and Wolpin (2000). 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 satisfaction of the instrumental exclusion restriction and the monotonicity of fertility in the instrument, see Imbens and Angrist (1994). Using the methods of Kitagawa (2008), Huber and Mellace (2011a), and Huber and Mellace (2012), we for the first time verify the validity of the sibling sex ratio instrument by statistical hypothesis tests, which suggest that violations are small if not close to nonexistent. We also provide novel sensitivity checks to assess deviations from the exclusion restriction and/or monotonicity in the nonparametric local average treatment effect framework and find the negative labor supply effect of fertility to be robust to a plausible range of violations.

Keywords

Instrumental variable, treatment effects, LATE, tests, sensitivity analysis.

JEL Classification

C12, C21, C26, J13, J22.

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 the context of the heterogeneous treatment effect model, which is the framework considered in this paper, an instrument is valid if it fulfills the following 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 no role in applications with just identified IV models.[1] However, many, if not most 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

[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.

siblings proposed by Angrist and Evans (1998) to estimate the female labor supply effect of fertility.[2] 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 only a small fraction 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 (who for instance prefer two sons over a daughter and a son) 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. While this raises concerns about the monotonicity assumption, the results nevertheless do not

[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.

provide a direct test for its violation. Even though both the exclusion restriction and monotonicity may be brought into question in the light of several theoretical and empirical findings, they can neither be unambiguously rejected nor approved based on the arguments brought forward in the scientific discussion.

Therefore, the main contribution of this paper is to for the first time verify the validity of the sibling sex ratio instrument of Angrist and Evans (1998) based on statistical grounds, i.e., by means of hypothesis tests.[3] To this end, we apply recently developed methods for testing IV validity in heterogenous treatment effect models which have been proposed by Kitagawa (2008), Huber and Mellace (2011a), and Huber and Mellace (2012). We also provide a unified framework for discussing the various tests, which are based on (i) the supremum of violations of IV validity over the potential outcome distribution of compliers, (ii) the sum of violations over the distribution, or (iii) violations in the compliers' mean potential outcomes. While the relative power of the tests relies on the specific data generating process at hand, they all share the common feature that asymptotic non-rejection does not imply IV validity (i.e., the tests do not have power against all deviations from the null 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 any scenario, our results suggest that violations of IV validity are at most very small. The tests never uniformly reject the null in any sample such that large deviations from the null can be plausibly ruled out. However, p-values below 10% of at least one of tests occur in several cases, implying that very moderate violations of the identifying assumptions may nevertheless exist.

[3]Other than the present work and the examples given in Kitagawa (2008), Huber and Mellace (2011a), and Huber and Mellace (2012), testing instruments appears to be virtually non-existent in empirical applications with just identified IV models. The only study we are aware of is Huber and Mellace (2011b), who investigate IV validity in several data sets used for female wage regressions, however, in the conceptually different context of sample selection models.

As a second contribution, we therefore propose a new approach for sensitivity analysis in the presence of an invalid instrument which is specifically tailored to the nonparametric LATE framework. In particular, we assess the robustness of the results under a violation of (i) the exclusion restriction, (ii) the monotonicity assumption, and (iii) both assumptions together. This complements previously suggested methods which focus either exclusively on deviations from the exclusion restriction (by allowing for a direct effect of the instrument on the outcome or a non-zero correlation of the instrument and the unobservables affecting the outcome), see Hirano, Imbens, Rubin, and Zhou (2000), Manski and Pepper (2000), Altonji, Elder, and Taber (2005), Hoogerheide and van Dijk (2006), Nevo and Rosen (2008), Flores and Flores-Lagunes (2010), Choi and Lee (2011), Conley, Hansen, and Rossi (2012), Kraay (2012), and Mealli and Pacini (2012),[4] or on deviations from monotonicity, see Huber and Mellace (2010) and Richardson and Robins (2010). Applying our approach to the sibling sex ratio instrument, we find that the negativity of the effect of fertility on female labor supply found by Angrist and Evans (1998) is robust to a plausible range of IV violations.

The remainder of this paper is organized as follows. Section 2 reviews the IV assumptions allowing for LATE identification and the methods to test their violation in a unified framework. Section 3 applies the tests to the Angrist and Evans (1998) data which come from the 1980 wave of the US Census Public Use Micro Samples. Section 4 presents our novel sensitivity analysis and uses it to investigate the robustness of the results in Angrist and Evans (1998) under violations of the exclusion restriction and monotonicity. Section 5 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

[4]For sensitivity analysis in IV models with overidentifying restrictions, (i.e., more instruments than endogenous treatments), see Small (2007). In this paper, we will focus on the just identified case.

factors that affect both the treatment and the outcome. In this case, identification of treatment effects generally requires an instrument, denoted by Z, 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). In this paper, we will assume that the instrument is binary ($Z \in \{0, 1\}$), e.g. having two mixed sex children ($Z = 0$) vs. having to same sex children ($Z = 1$). However, the results could be easily extended to bounded non-binary instruments. 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 $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$.

Table 1: Types

Type T	D(1)	D(0)	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

As discussed in Angrist, Imbens, and Rubin (1996) and summarized in Table 1, 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 react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types, $D(z) \neq z$ for either $Z = 1$, or $Z = 0$, or both: The always takers are always treated irrespective of the instrument state, the never takers are never treated, and the defiers only take the treatment when $Z = 0$. It is obvious that we cannot directly observe the type any observation belongs to as either $D(1)$ or $D(0)$ remains unknown. This implies that any observation i with a particular combination of the treatment and the instrument may belong to one of two types, see Table 2.

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 that has been inspired by Kitagawa (2009). Let $f(y, D = d|Z = z)$ denote the

Table 2: Observed subgroups and types

Observed values of Z and D	Potential types T
$\{i : Z_i = 1, D_i = 1\}$	observation i belongs either to a or to c
$\{i : Z_i = 1, D_i = 0\}$	observation i belongs either to d or to n
$\{i : Z_i = 0, D_i = 1\}$	observation i belongs either to a or to d
$\{i : Z_i = 0, D_i = 0\}$	observation i belongs either to c or to n

(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 $t \in \{a, c, d, n\}$. From Table 2 follows that the subsequent relationships of 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) make immediate use of the fact that any observed joint density is constituted by the potential outcomes of two different types conditional on Z, unless further assumptions are imposed.

We will now discuss the identifying assumptions. Imbens and Angrist (1994) have shown that the LATE on the compliers is point identified when imposing (i) independence between Z and the joint distribution of the potential treatment states and outcomes and (ii) monotonicity of the treatment in the instrument. Formally, the first assumption can be stated as follows:

Assumption 1:

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

where "$\perp$" denotes independence. Assumption 1 is standard in the LATE literature and implies the randomization of the instrument (such that it is unrelated with factors affecting the treatment

and/or outcome) and the exclusion of direct effects on the outcome. Note that not only the potential outcomes, but also the types, which are defined by the potential treatment states, are independent of the instrument.

Assumption 2:

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

Assumption 2 says that the potential treatment state of any individual does either not decrease (positive monotonicity) or not increase (negative monotonicity) in the instrument. Without loss of generality, we will henceforth only consider the case of positive monotonicity, because the case of negative monotonicity is symmetric. Under $\Pr(D(1) \geq D(0)) = 1$, the existence of defiers (type d) is ruled out because for the latter group, $D(1) < D(0)$. Note that monotonicity is also implicitly assumed in the linear IV model, where the effect of the instrument on the treatment is represented by a homogenous first stage coefficient.

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:[5]

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

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

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). The latter generally deviates from the average treatment effect in the total population because it does not comprise the effects on the always and never takers, who do not react on the instrument. 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:

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

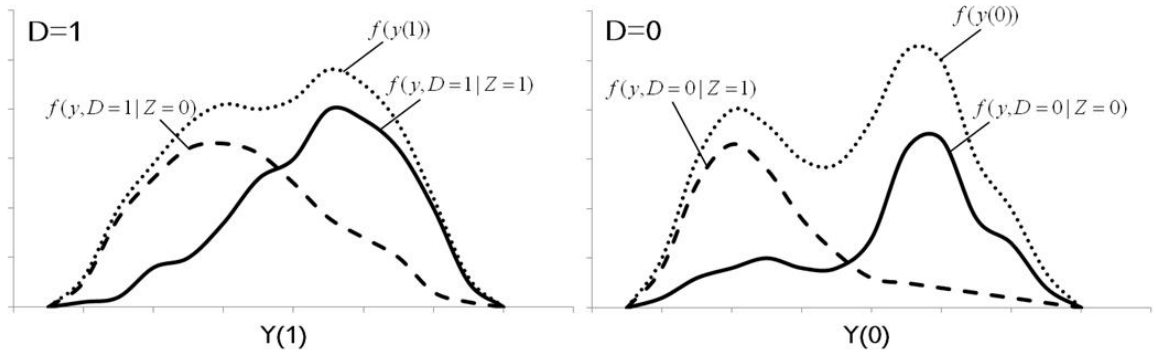
[5]From the subsequent discussion it is easy to see that also the conditional densities $f(y(d)|T = c) = f(y(d), T = c)/\Pr(T = c)$ are identified, which was first acknowledged by Imbens and Rubin (1997).

It is interesting to note that 9 and 10 not only imply the identification of the LATE, but also provide testable implications 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. However, a density can never 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 such that monotonicity is not satisfied, or both. The constraints (13) were first derived by Balke and Pearl (1997) for binary outcomes, while Kitagawa (2008) formulated them in terms of continuous outcomes. For visualization, Figure 1 plots the observed joint densities under treatment and non-treatment of a hypothetical data generating process in which the constraints are violated over a subset of the support of the potential outcomes, i.e., $f(y, D = 1|Z = 1) < f(y, D = 1|Z = 0)$ and $f(y, D = 0|Z = 0) < f(y, D = 0|Z = 1)$. (The unobserved densities of the potential outcomes in the total population, $f(y(1))$ and $f(y(0))$, which consist of the respective sums of type specific densities, are also provided: $f(y(d)) = \sum_{t \in \{a,c,n,d\}} f(y(d), T = t)$.)

Figure 1: Graphical illustration of the violations



For testing, we would ideally like to verify (13) at each value y in the support of Y. However, if the outcome is of rich support (e.g., continuous), it may be preferable in terms of finite sample power to partition the support into a finite number of subsets when applying the tests. In the

latter case, the constraints 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. It follows that (14) must hold for any A. Based on these constraints, Kitagawa (2008) proposes a test that makes use of 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 method for inference.

In contrast, Huber and Mellace (2011a) propose a method which tests constraints on the mean potential outcomes of the compliers rather than the densities or probability measures defined by A. To derive their mean constraints, first multiply (5) by y and integrate over y to obtain

$$\begin{aligned}&\int yf(y, D = 1|Z = 1)dy = \int yf(y(1), T = c)dy + \int yf(y(1), T = a)dy\\ = \;& E(Y, D = 1|Z = 1) = E[Y(1), T = c] + E[Y(1), T = a]\\ = \;& E[Y(1)|T = c] \cdot \pi_c + E[Y(1)|T = a] \cdot \pi_a.\end{aligned} \tag{15}$$

Furthermore, from (5) it is easy to see that

$$\Pr(D = 1|Z = 1) = \int f(y, D = 1|Z = 1)dy = \pi_a + \pi_c, \tag{16}$$

such that dividing (15) by $\Pr(D = 1|Z = 1)$ yields

$$\begin{aligned}&\frac{E(Y, D = 1|Z = 1)}{\Pr(D = 1|Z = 1)} = \frac{E[Y(1)|T = c] \cdot \pi_c + E[Y(1)|T = a] \cdot \pi_a}{\Pr(D = 1|Z = 1)}\\ = \;& E(Y|D = 1, Z = 1) = E[Y(1)|T = c] \cdot \frac{\pi_c}{\pi_a + \pi_c} + E[Y(1)|T = a] \cdot \frac{\pi_a}{\pi_a + \pi_c}.\end{aligned} \tag{17}$$

$\frac{\pi_c}{\pi_a+\pi_c} = \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=1|Z=1)}$ and $\frac{\pi_a}{\pi_a+\pi_c} = 1 - \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=1|Z=1)}$ are the shares of the compliers and always takers among those with $D = 1$ and $Z = 1$, i.e., among the mixed population of compliers and always takers, respectively. For notational ease, we denote the share of always takers in the mixed population by $q = 1 - \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=1|Z=1)}$. It follows that

$$E(Y|D = 1, Z = 1, Y \leq y_q), \qquad E(Y|D = 1, Z = 1, Y \geq y_{1-q}) \tag{18}$$

are the lower and upper bounds on the always takers' mean potential outcome under treatment, where y_q is the qth quantile of Y given $D = 1$ and $Z = 1$. The intuition is that one can bound the mean potential outcome of the always takers by taking averages in the upper and lower proportions of outcomes with $D = 1$ and $Z = 1$ that correspond to the share of the always takers in the mixed population characterized by (17).

Moreover, note that by (6),

$$\begin{aligned} \int yf(y, D = 1|Z = 0)dy &= \int yf(y(1), T = a)dy = E(Y, D = 1|Z = 0) \\ &= E[Y(1), T = a] = E[Y(1)|T = a] \cdot \pi_a, \end{aligned}$$

such that

$$\frac{E[Y, D = 1|Z = 0]}{\Pr(D = 1|Z = 0)} = E(Y|D = 1, Z = 0) = \frac{E[Y(1)|T = a] \cdot \pi_a}{\Pr(D = 1|Z = 0)} = E[Y(1)|T = a], \tag{19}$$

since it also follows from (6) that $\Pr(D = 1|Z = 0) = \pi_a$. I.e., the LATE assumptions imply that $E(Y|D = 1, Z = 0)$ point identifies the mean potential outcome of the always takers under treatment. It is obvious that $E(Y|D = 1, Z = 0)$ must lie within the bounds $E(Y|D = 1, Z = 1, Y \leq y_q)$, $E(Y|D = 1, Z = 1, Y \geq y_{1-q})$, otherwise the identifying assumptions are violated. This implies the testable constraints

$$E(Y|D = 1, Z = 1, Y \leq y_q) \leq E(Y|D = 1, Z = 0) \leq E(Y|D = 1, Z = 1, Y \geq y_{1-q}). \tag{20}$$

An equivalent result applies to the never takers under non-treatment. By analogous arguments as before,

$$E(Y|D=0,Z=0) = E[Y(0)|T=c]\cdot\frac{\pi_c}{\pi_n+\pi_c} + E[Y(0)|T=n]\cdot\frac{\pi_n}{\pi_n+\pi_c}, \quad (21)$$

$$E(Y|D=0,Z=1) = E[Y(0)|T=n], \quad (22)$$

with $\frac{\pi_c}{\pi_n+\pi_c} = \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=0|Z=0)}$ and $\frac{\pi_a}{\pi_a+\pi_c} = 1 - \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=0|Z=0)}$. Denote by r the share of never takers in the mixed population of never takers and compliers, i.e., $r = 1 - \frac{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}{\Pr(D=0|Z=0)}$. It follows that

$$E(Y|D=0,Z=0,Y\leq y_r) \leq E(Y|D=0,Z=1) \leq E(Y|D=0,Z=0,Y\geq y_{1-r}) \quad (23)$$

is a necessary condition for the satisfaction of IV validity, with y_r denoting the rth quantile of Y given $D=0$ and $Z=0$. Huber and Mellace (2011a) therefore propose to jointly test the following mean restrictions:

$$\begin{pmatrix} E(Y|D=1,Z=1,Y\leq y_q) - E(Y|D=1,Z=0) \\ E(Y|D=1,Z=0) - E(Y|D=1,Z=1,Y\geq y_{1-q}) \\ E(Y|D=0,Z=0,Y\leq y_r) - E(Y|D=0,Z=1) \\ E(Y|D=0,Z=1) - E(Y|D=0,Z=0,Y\geq y_{1-r}) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \quad (24)$$

This testing problem fits into the framework of inequality moment constraints, see for instance Chernozhukov, Hong, and Tamer (2007). For this reason, Huber and Mellace (2011a) use the minimum p-value-type test developed by Bennett (2009) for joint testing of inequality moment constraints, even though other methods suggested in the literature (see for instance Andrews and Jia (2008), Andrews and Soares (2010), Chen and Szroeter (2009), Donald and Hsu (2011), and Hansen (2005)) could also be used. The test algorithm is provided in Appendix A.

As a further contribution, Huber and Mellace (2011a) not only consider mean constraints, but also the following probability constraints which they derive using the results of Horowitz and

Manski (1995):

$$\frac{\Pr(Y \in A|D=1,Z=1)-(1-q)}{q} \leq \Pr(Y \in A|D=1,Z=0) \leq \frac{\Pr(Y \in A|D=1,Z=1)}{q}, \tag{25}$$

$$\frac{\Pr(Y \in A|D=0,Z=0)-(1-r)}{r} \leq \Pr(Y \in A|D=0,Z=1) \leq \frac{\Pr(Y \in V|D=0,Z=0)}{r}. \tag{26}$$

It can easily be shown that the second inequalities in (25) and (26) are equivalent to (14), see Appendix A.3 in Huber and Mellace (2011a). In contrast, the first inequalities have not been considered by Kitagawa (2008) and are equivalent to the following constraints formulated in terms of joint densities:

$$\begin{aligned} \Pr(Y \in A, D=1|Z=1)-\Pr(Y \in A, D=1|Z=0) &\leq \Pr(D=1|Z=1)-\Pr(D=1|Z=0), \\ \Pr(Y \in A, D=0|Z=0)-\Pr(Y \in A, D=0|Z=1) &\leq \Pr(D=1|Z=1)-\Pr(D=1|Z=0). \end{aligned} \tag{27}$$

The intuition of (27) is that the joint probability of being a complier and having an outcome that lies in subset A cannot be larger than the (unconditional) complier share in the population. However, it has to be pointed out that these constraints can possibly only improve testing power in setups with overlapping definitions of the subsets A (such that inadmissible negative densities may be averaged out). Otherwise, they do not provide additional information compared to (14), see Huber and Mellace (2011a) for an in depth discussion and an intuitive example. As for the mean restrictions, the Bennett (2009) test can be used to verify (25) and (26) based on the

following inequality constraints:

$$\begin{pmatrix} \frac{\Pr(Y\in A|D=1,Z=1)-(1-q)}{q} - \Pr(Y \in A|D=1,Z=0) \\ \Pr(Y \in A|D=1,Z=0) - \frac{\Pr(Y\in A|D=1,Z=1)}{q} \\ \frac{\Pr(Y\in A|D=0,Z=0)-(1-r)}{r} - \Pr(Y \in A|D=0,Z=1) \\ \Pr(Y \in A|D=0,Z=1) - \frac{\Pr(Y\in A|D=0,Z=0)}{r} \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{28}$$

Similarly to the Kitagawa (2008) test, the total number of constraints tested depends on the number of subsets A. I.e., by (28), the total number of constraints is four times the number of subsets.

The final approach proposed in the literature comes from Huber and Mellace (2012). Instead of basing testing on the supremum of violations of the constraints (13) or coarser functions thereof as in Kitagawa (2008) and Huber and Mellace (2011a), they consider the integral of violations over the support of Y. To be concise, Huber and Mellace (2012) invoke Assumption 1 only and therefore, violations point to the non-satisfaction of Assumption 2 in their framework. For the latter case, they propose an alternative identification strategy based on a local version of monotonicity, i.e., monotonicity given particular values of the potential outcomes, which is to assume that locally either compliers or defiers may exist. However, if the researcher is not even sure about the independence of the instrument, their approach can be used to test Assumptions 1 and 2 jointly. To see this, note that (13) implies that

$$\begin{aligned} & \int [f(y,D=1|Z=0) - \min(f(y,D=1|Z=1), f(y,D=1|Z=0))]dy \\ = \; & \Pr(D=1|Z=0) - \int \min(f(y,D=1|Z=1), f(y,D=1|Z=0))dy = 0, \qquad (29) \\ & \int [f(y,D=0|Z=1) - \min(f(y,D=0|Z=0), f(y,D=0|Z=1))]dy \\ = \; & \Pr(D=0|Z=1) - \int \min(f(y,D=0|Z=0), f(y,D=0|Z=1))dy = 0. \qquad (30) \end{aligned}$$

I.e., if Assumptions 1 and 2 are satisfied, $\min(f(y,D=1|Z=1), f(y,D=1|Z=0)) = f(y,D=1|Z=0)$ because $f(y,D=1|Z=1) \geq f(y,D=1|Z=0)$ and $\min(f(y,D=0|Z=0), f(y,D=$

$0|Z=1)) = f(y, D=0|Z=1)$ because $f(y, D=0|Z=0) \geq f(y, D=0|Z=1)$. Therefore, if the integrals are positive, the constraints in (13) are violated for at least one value of y in the support of Y. If multiple violations occur, this approach is potentially more powerful than tests based on the supremum of violations, while the latter may have better properties if the assumptions are violated only locally at a specific point in the support of Y.

In the application considered in the next section, the outcome is discrete such that $f(y, D = d|Z = d)$ can be estimated at $\sqrt{n}$-rate, which serves as a plug-in estimator for the estimation of (29) and (30) by their respective sample analogs. For inference, we will use subsampling (without replacement), see Politis, Romano, and Wolf (1999), to approximate the distributions of (29) and (30). Then, their positivity can be tested by assessing the rank of the estimates in their respective re-centered subsampling distributions. I.e., we perform a one-sided hypothesis test (as the lower bound of (29) and (30) is zero) based on the share of re-centered subsampling integrals being larger than the original integrals in the data. Even if (29) and (30) are close to zero, subsampling inference is pointwise consistent while the bootstrap is not.[6]

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 such 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

[6]See for instance Andrews (2000) for a formal discussion of why the bootstrap fails close to boundaries of the parameter space.

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 assessment of the sibling sex ratio instrument by means of hypothesis tests to for the first time provide statistical evidence on the validity of this 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 3 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 provides the results for the mean test based on (24). Columns 3 and 4 contain the p-values of the supremum tests of Huber and Mellace (2011a) based on (28) and of Kitagawa (2008) based on

Table 3: P-values when testing the sibling sex ratio instrument

		supremum tests		integral tests	
	Mean test	HM11 test	K08 test	for $D=1$	for $D=0$
full sample (394,840)	0.580	0.638	0.210	0.984	0.999
edu<12 (88,764)	1.000	0.778	0.745	0.869	0.890
edu=12 (189,818)	1.000	0.070	0.034	0.958	0.969
edu>12 (116,258)	1.000	0.961	0.996	0.980	0.999
edu<12, a. 20-25, married, white (11,716)	0.501	0.541	0.567	0.402	0.654
edu=12, a. 20-25, married, white (16,249)	0.520	0.092	0.048	0.892	0.196
edu>12, a. 20-25, married, white (3,232)	0.822	1.000	1.000	0.440	0.474
edu<12, a. 26-30, married, white (19,864)	0.617	0.030	0.022	0.922	0.597
edu=12, a. 26-30, married, white (53,919)	1.000	0.164	0.168	0.812	0.786
edu>12, a. 26-30, married, white (28,266)	1.000	0.304	0.376	0.848	0.812
edu<12, a. 31-35, married, white (23,367)	1.000	0.234	0.106	0.670	0.729
edu=12, a. 31-35, married, white (75,850)	1.000	0.368	0.378	0.947	0.994
edu>12, a. 31-35, married, white (58,684)	1.000	0.674	0.487	0.953	0.980
edu<12, a. 20-25, mar., w., 1st birth 70-72 (2,784)	0.604	0.426	0.559	0.170	0.551
edu=12, a. 20-25, mar., w., 1st birth 70-72 (1,713)	0.477	1.000	0.558	0.279	0.292
edu>12, a. 20-25, mar., w., 1st birth 70-72 (286)	0.153	1.000	0.061	0.331	0.031
edu<12, a. 26-30, mar., w., 1st birth 70-72 (8,592)	0.704	0.105	0.223	0.832	0.508
edu=12, a. 26-30, mar., w., 1st birth 70-72 (24,492)	1.000	0.277	0.228	0.610	0.619
edu>12, a. 26-30, mar., w., 1st birth 70-72 (9,149)	0.703	0.147	0.265	0.797	0.233
edu<12, a. 31-35, mar., w., 1st birth 70-72 (2,389)	0.624	1.000	0.536	0.348	0.224
edu=12, a. 31-35, mar., w., 1st birth 70-72 (15,913)	1.000	0.474	0.239	0.569	0.814
edu>12, a. 31-35, mar., w., 1st birth 70-72 (20,037)	1.000	0.502	0.403	0.868	0.976

Note: The p-values of the mean and supremum tests are based on 999 bootstrap draws. The p-values of the integral tests are based on 999 subsamples where the subsampling size corresponds to the integer closest to $n^{0.9}$.

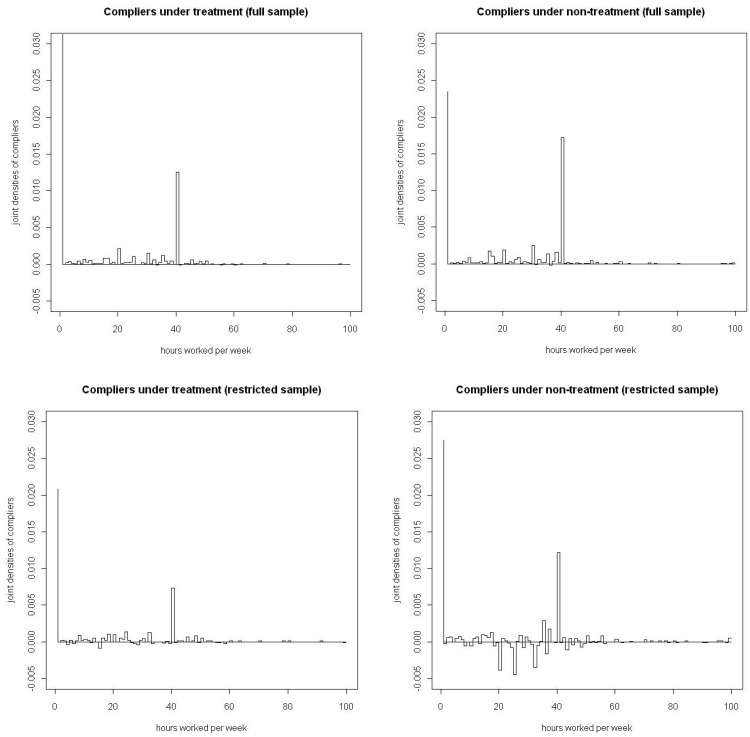
(14). In both tests, the support of the outcome is partitioned into five subsets A, which are defined by an equidistant grid from the minimum to the maximum of the observed outcome. The integral tests (columns 5 and 6) based on the sample analogs of (29) and (30). As the outcome variable is discrete, $f(y, D = d|Z = z)$ can be easily estimated by the empirically observed probability masses at $\sqrt{n}$-rate: $\hat{f}(y, D = d|Z = z) = \frac{1}{\sum I\{Z_i=z\}} \sum (I\{Z_i = z\} \cdot I\{Y_i = y, D_i = d\})$, $z, d \in \{1, 0\}$, where $I\{\cdot\}$ is the indicator function, which is equal to one if its argument holds true and zero otherwise, and i is the index of the observations in the sample. The subsampling inference of the integral tests uses the integer that is closest to $n^{0.9}$ as subsampling size (or block size).[7]

Taking a look at the results in the full sample reveals that no test provides evidence against the violation 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). For 12 years of education, the supremum tests of Huber and Mellace (2011a) and Kitagawa (2008) reject the null at the 10 and 5% level, respectively, while for the other education categories, all test statistics are insignificant. 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 the 22 samples considered at least one test rejects the IV assumptions at the 5% level and there is no single scenario in which all tests reject uniformly. This suggests that violations of instrumental validity are, even though they cannot be ruled out completely, probably quite small.

For illustrative purposes, Figure 2 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).

[7]Therefore, $b \to \infty$ and $b/n \to 0$ as $n \to \infty$, which is required for the consistency of subsampling inference. Smaller choices of the subsampling size (e.g., the integer closest to $n^{0.7}$ or $n^{0.5}$) have also been considered, but are not reported here. They generally entail higher p-values.

Figure 2: Estimates of $f(y(1), T = c)$ and $f(y(0), T = c)$



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 the violations are only minor and not statistically significant in the full sample (possibly due to averaging out negative densities across subpopulations), they are more pronounced in the restricted sample and in particular among the non-treated. For the latter, the Kitagawa (2008) is significant at the 5% level.

In a next step, we split instrument into two separate 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

Table 4: Two boys and two girls instruments

		supremum tests		integral tests	
	Mean test	HM11 test	K08 test	for $D=1$	for $D=0$
two girls vs. mixed					
full sample (290,713)	1.000	0.180	0.190	0.999	0.999
edu<12 (65,597)	1.000	0.888	1.000	0.973	0.829
edu=12 (139,674)	1.000	0.175	0.075	0.776	0.994
edu>12 (85,442)	1.000	0.194	0.264	0.969	0.979
edu<12, age 21-25, married, white (8,662)	0.515	0.057	0.427	0.370	0.699
edu=12, age 21-25, married, white (11,892)	0.709	0.086	0.012	0.410	0.203
edu>12, age 21-25, married, white (2,368)	0.638	1.000	0.903	0.495	0.238
edu<12, age 26-30, married, white (14,664)	1.000	0.032	0.024	0.931	0.713
edu=12, age 26-30, married, white (39,503)	1.000	0.184	0.149	0.784	0.967
edu>12, age 26-30, married, white (20,779)	1.000	0.453	0.505	0.807	0.874
edu<12, age 31-35, married, white (17,142)	1.000	0.471	0.381	0.853	0.828
edu=12, age 31-35, married, white (55,751)	1.000	0.529	0.461	0.882	0.992
edu>12, age 31-35, married, white (42,982)	1.000	0.191	0.358	0.953	0.786
edu<12, a. 21-25, mar., w., 1st birth 70-72 (2,064)	0.544	1.000	0.702	0.292	0.468
edu=12, a. 21-25, mar., w., 1st birth 70-72 (1,267)	0.576	1.000	0.510	0.156	0.133
edu>12, a. 21-25, mar., w., 1st birth 70-72 (204)	0.302	0.721	0.238	0.163	0.146
edu<12, a. 26-30, mar., w., 1st birth 70-72 (6,330)	0.703	0.256	0.205	0.455	0.573
edu=12, a. 26-30, mar., w., 1st birth 70-72 (17,903)	1.000	0.611	0.657	0.477	0.951
edu>12, a. 26-30, mar., w., 1st birth 70-72 (6,736)	1.000	0.492	0.452	0.507	0.238
edu<12, a. 31-35, mar., w., 1st birth 70-72 (1,762)	0.857	1.000	0.632	0.497	0.132
edu=12, a. 31-35, mar., w., 1st birth 70-72 (11,643)	0.920	0.910	1.000	0.711	0.875
edu>12, a. 31-35, mar., w., 1st birth 70-72 (14,690)	1.000	0.430	0.196	0.852	0.907
two boys vs. mixed					
full sample (299,419)	1.000	0.502	0.452	0.963	0.985
edu<12 (67,183)	1.000	0.594	0.485	0.640	0.919
edu=12 (143,863)	1.000	0.094	0.057	0.987	0.812
edu>12 (88,373)	1.000	0.681	0.528	0.918	0.985
edu<12, age 21-25, married, white (8,835)	0.522	0.540	0.493	0.309	0.511
edu=12, age 21-25, married, white (12,283)	0.495	0.513	0.283	0.920	0.247
edu>12, age 21-25, married, white (2,442)	0.677	1.000	0.360	0.211	0.679
edu<12, age 26-30, married, white (15,072)	0.650	0.063	0.327	0.823	0.239
edu=12, age 26-30, married, white (40,946)	1.000	0.338	0.300	0.832	0.423
edu>12, age 26-30, married, white (21,543)	0.511	0.352	0.474	0.621	0.766
edu<12, age 31-35, married, white (17,882)	0.521	0.089	0.096	0.409	0.559
edu=12, age 31-35, married, white (57,592)	1.000	0.536	0.490	0.978	0.991
edu>12, age 31-35, married, white (44,665)	1.000	0.182	0.085	0.828	0.967
edu<12, a. 21-25, mar., w., 1st birth 70-72 (2,068)	0.704	0.595	0.519	0.083	0.638
edu=12, a. 21-25, mar., w., 1st birth 70-72 (1,294)	0.577	1.000	0.756	0.371	0.306
edu>12, a. 21-25, mar., w., 1st birth 70-72 (226)	0.122	1.000	0.056	0.151	0.014
edu<12, a. 26-30, mar., w., 1st birth 70-72 (6,491)	0.653	0.186	0.589	0.575	0.239
edu=12, a. 26-30, mar., w., 1st birth 70-72 (18,685)	1.000	0.348	0.323	0.880	0.576
edu>12, a. 26-30, mar., w., 1st birth 70-72 (7,022)	0.687	0.148	0.225	0.310	0.420
edu<12, a. 31-35, mar., w., 1st birth 70-72 (1,817)	0.580	1.000	0.300	0.071	0.499
edu=12, a. 31-35, mar., w., 1st birth 70-72 (12,076)	1.000	0.232	0.132	0.380	0.723
edu>12, a. 31-35, mar., w., 1st birth 70-72 (15,265)	0.617	0.103	0.322	0.679	0.989

Note: The p-values of the mean and supremum tests are based on 999 bootstrap draws. The p-values of the integral tests are based on 999 subsamples where the subsampling size corresponds to the integer closest to $n^{0.9}$.

parents prefer two boys over mixed sex siblings) as one might suspect from the empirical findings in Lee (2008) and Dahl and Moretti (2008). If the non-monotonicity in the sample with two boys exclusively causes the IV violation, the tests should pass when using the instrument defined by having two girls and reject when using two boys. In this context, 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). Given 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. At the very least, this confirms that a preference for boys over girls is evident in the PUMS, too.

Table 4 presents the results when running the tests separately for the two girls and two boys instruments in the respective full samples and conditional on covariates. We find only limited heterogeneity in the rejection rates that would otherwise point to one instrument being more valid than the other. For either instrument, the null is rejected just once at the 5% level by at least one test across the 22 samples. At the 10% level, rejections of at least one test occur four times for the two girls instrument and seven times for two boys instrument, which weakly supports the larger concerns about the boys instrument. Nevertheless, the tests do not provide evidence for severe violations of Assumptions 1 and 2 under either instrument.

4 Sensitivity analysis

Even though the results presented in the last section do not point to strong violations of the validity of the sibling sex ratio instrument, there may nevertheless exist moderate deviations from Assumptions 1 and 2 that are hard to detect for the tests. It therefore appears reasonable to verify the robustness of the LATE estimate w.r.t. to a plausible range of violations. To this end, we present new sensitivity checks which are specifically tailored to the nonparametric LATE framework and allow for both the non-satisfaction the exclusion restriction and the monotonicity assumption. To outline our approach, we first reconsider (17) and (21), which give the relationships

of observed conditional means and the mean potential outcomes of the latent types if IV validity is satisfied. Solving both equations for the mean potential outcome of compliers yields

$$\begin{aligned} E[Y(1)|T=c] &= \frac{(\pi_a+\pi_c)\cdot E(Y|D=1,Z=1)-\pi_a\cdot E[Y(1)|T=a]}{\pi_c} \\ &= \frac{\Pr(D=1|Z=1)\cdot E(Y|D=1,Z=1)-\Pr(D=1|Z=0)\cdot E(Y|D=1,Z=0)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)} \end{aligned} \tag{31}$$

and

$$\begin{aligned} E[Y(0)|T=c] &= \frac{(\pi_n+\pi_c)\cdot E(Y|D=0,Z=0)-\pi_n\cdot E[Y(1)|T=n]}{\pi_c} \\ &= \frac{\Pr(D=0|Z=0)\cdot E(Y|D=0,Z=0)-\Pr(D=0|Z=1)\cdot E(Y|D=0,Z=1)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}, \end{aligned} \tag{32}$$

where we make use of the fact that $E(Y|D=1,Z=0)=E[Y(1)|T=a]$ and $E(Y|D=0,Z=1)=E[Y(1)|T=n]$ (see (19) and (22)) if the instrument is valid. Note that by subtracting one from another, we obtain

$$\begin{aligned} \Delta_c &= \frac{\Pr(D=1|Z=1)\cdot E(Y|D=1,Z=1)+\Pr(D=0|Z=1)\cdot E(Y|D=0,Z=1)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)} \\ &- \frac{\Pr(D=1|Z=0)\cdot E(Y|D=1,Z=0)+\Pr(D=0|Z=0)\cdot E(Y|D=0,Z=0)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)} \\ &= \frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}, \end{aligned} \tag{33}$$

which is an alternative derivation of the probability limit of the Wald estimator to that offered in Section 2. In the subsequent discussion, we will consider modifications of (31), (32), (33) that permit accommodating the non-satisfaction of Assumptions 1 and/or 2.

To allow for a violation of the exclusion restriction, we assume that the instrument has a direct effect on the mean potential outcomes of the various types. This implies that the conditional mean outcomes given $Z=1$, i.e., $E(Y|D=1,Z=1)$ and $E(Y|D=0,Z=1)$, not only consist of the mean potential outcomes of the types, but also of additional components reflecting the direct effect, which is absent under $Z=0$. To accommodate this in our framework, we add potentially

heterogenous direct effects of Z to the mean potential outcomes of the compliers, always takers, and never takers in (17) and (19), respectively, which are denoted by γ_c, γ_a, and γ_n:

$$E(Y|D=1,Z=1) = (E[Y(1)|T=c]+\gamma_c)\cdot\frac{\pi_c}{\pi_a+\pi_c}+(E[Y(1)|T=a]+\gamma_a)\cdot\frac{\pi_a}{\pi_a+\pi_c} \quad (34)$$

$$E(Y|D=0,Z=1) = E[Y(0)|T=n]+\gamma_n \quad (35)$$

E.g., considering the compliers, γ_c reflects the fact that $E[Y(1)|T=c,Z=1] \neq E[Y(1)|T=c,Z=0]$ if Assumption 1 does not hold and captures the difference in both quantities: $\gamma_c = E[Y(1)|T=c,Z=1]-E[Y(1)|T=c,Z=0]$. Rearranging terms gives

$$E(Y|D=1,Z=1)-\gamma_c\cdot\frac{\pi_c}{\pi_a+\pi_c}-\gamma_a\cdot\frac{\pi_a}{\pi_a+\pi_c} = E(Y(1)|T=c)\cdot\frac{\pi_c}{\pi_a+\pi_c}+E[Y(1)|T=a]\cdot\frac{\pi_a}{\pi_a+\pi_c}, \quad (36)$$

$$E(Y|D=0,Z=1)-\gamma_n = E[Y(0)|T=n], \quad (37)$$

implying that we can correct for biases due to the non-satisfaction of Assumption 1 by subtracting (a weighted average of) the type-specific direct effects from the observed conditional mean outcomes. Note that modeling violations of the exclusion restriction based on mean potential outcomes differs from the conventional approach of considering violations in (parametric) structural models found in several recent studies.[8] E.g., Conley, Hansen, and Rossi (2012) assume a direct effect of Z on the observed (rather than potential) Y, while Altonji, Elder, and Taber (2005), Nevo and Rosen (2008), and Choi and Lee (2011) investigate the implications of a non-zero correlation of Z and the error term in the structural equation.[9] In contrast to these studies,

[8]Examples not following this path of tight parametric modeling are Manski and Pepper (2000), who assume the potential outcomes to be weakly monotonic in the instrument, Flores and Flores-Lagunes (2010), who impose various sets of weak dominance assumptions of mean potential outcomes across or within subpopulations defined by the values of the potential treatment status under each value of the instrument, and Mealli and Pacini (2012), who exploit restrictions on the joint distribution of the outcome and auxiliary variables (such as secondary outcomes and covariates) that are implied by the randomization of the instrument.

[9]The three latter papers differ w.r.t. their assumptions about the correlation of the instrument and the error term. Altonji, Elder, and Taber (2005) assume that the latter can be approximated based on the correlation of the instrument with the (index of the) observed covariates other than the endogenous treatment that also determine the outcome. This requires the observed covariates be randomly drawn from the total of characteristics that affect the outcome. Nevo and Rosen (2008) assume that the instrument is in absolute terms less correlated with the error than the treatment. Choi and Lee (2011) assume the direction of the correlation between the instrument and the

our method has the advantage that one can in principle allow for heterogeneity in the violation of the (mean) exclusion restriction across always takers, never takers, and compliers, if the researcher has information on the severity of the violation across types.

Of course, if we have no prior believe on whether such a heterogeneity exists, we may consider the same range of supposed direct effects for the mean potential outcomes of all three types. In this case, $\gamma_a = \gamma_n = \gamma_c = \gamma$, such that (36) and (37) simplify to

$$E(Y|D=1,Z=1) - \gamma = E[Y(1)|T=c] \cdot \frac{\pi_c}{\pi_a + \pi_c} + E[Y(1)|T=a] \cdot \frac{\pi_a}{\pi_a + \pi_c}, \quad (38)$$

$$E(Y|D=0,Z=1) - \gamma = E[Y(0)|T=n]. \quad (39)$$

Replacing $E(Y|D=1,Z=1)$ in (31) by $E(Y|D=1,Z=1) - \gamma$ and $E(Y|D=0,Z=1)$ in (32) by $E(Y|D=0,Z=1) - \gamma_n$ yields the potential outcomes of the compliers when accounting for the direct effect:

$$E[Y(1)|T=c] = \frac{\Pr(D=1|Z=1) \cdot (E(Y|D=1,Z=1) - \gamma) - \Pr(D=1|Z=0) \cdot E(Y|D=1,Z=0)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}, \quad (40)$$

$$E[Y(0)|T=c] = \frac{\Pr(D=0|Z=0) \cdot E(Y|D=0,Z=0) - \Pr(D=0|Z=1) \cdot (E(Y|D=0,Z=1) - \gamma)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}. \quad (41)$$

Similarly to (33), subtracting one from another gives

$$\Delta_c = \frac{E(Y|Z=1) - \gamma - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}. \quad (42)$$

This corresponds to the probability limit of the slope coefficient in TSLS when regressing D on one and Z in the first stage and $Y - Z\gamma$ on one and the first stage prediction of D in the second stage. The latter approach is discussed in Conley, Hansen, and Rossi (2012) along with suitable (but conservative) inference procedures under alternative assumptions about γ.

error to be known.

Therefore, if γ is homogeneous across types, our identification result is numerically equivalent to the approach of Conley, Hansen, and Rossi (2012), however, the interpretation differs. In our nonparametric framework, we only claim to identify the LATE on the defiers, whereas the parametric assumptions in Conley, Hansen, and Rossi (2012) imply the identification of the average treatment effect in the entire population. Furthermore, if we considered heterogenous $\gamma_a, \gamma_n, \gamma_c$, then $E(Y|D=1,Z=1)$ and $E(Y|D=0,Z=1)$ in (31) and (32) would have to be replaced by the left hand side expressions of (36) and (37), such that the LATE does no longer simplify to (42) or the probability limit of TSLS.

As a second source of IV invalidity, we consider the consequences of a violation of the monotonicity assumption for LATE identification, which has to the best of our knowledge previously only been discussed in Huber and Mellace (2010) (for general bounded outcomes) and Richardson and Robins (2010) (for binary outcomes). The latter studies provide nonparametric bounds for the case that merely monotonicity does not hold, while the IV independence assumption is maintained. After assessing the implications of non-monotonicity for (31) and (32), we will finally allow for the violation of both assumptions at the same time. Two issues arise when monotonicity is not satisfied. Firstly, the shares of the compliers or any other type in (31) and (32) are no longer point identified, because without assuming monotonicity, (6) and (7) become

$$f(y, D=1|Z=0) = f(y(1), T=a) + f(y(1), T=d), \tag{43}$$

$$f(y, D=0|Z=1) = f(y(0), T=n) + f(y(0), T=d), \tag{44}$$

such that

$$\begin{aligned} \int f(y, D=1|Z=0)dy &= \Pr(D=1|Z=0) = \int f(y(1), T=a) + f(y(1), T=d)dy = \pi_a + \pi_d, \\ &\Leftrightarrow \Pr(D=1|Z=0) - \pi_d = \pi_a, \end{aligned} \tag{45}$$

$$\begin{aligned} \int f(y, D=0|Z=1)dy &= \Pr(D=0|Z=1) = \int f(y(0), T=n) + f(y(0), T=d)dy = \pi_n + \pi_d, \\ &\Leftrightarrow \Pr(D=0|Z=1) - \pi_d = \pi_n. \end{aligned} \tag{46}$$

Therefore, in contrast to (11),

$$\Pr(D=1|Z=1)-\Pr(D=1|Z=0)=\pi_c-\pi_d \quad \Leftrightarrow \quad \Pr(D=1|Z=1)-\Pr(D=1|Z=0)+\pi_d=\pi_c, \tag{47}$$

which implies that the share of compliers is not point identified.

Secondly, the mean potential outcomes of the always takers in (31) and the never takers in (32) are no longer identified, which is required for the point identification of the compliers' mean potential outcomes under treatment and non-treatment. The reason is that (19) and (22) do not hold because $E(Y|D=1,Z=0)$ and $E(Y|D=0,Z=1)$ are now mixtures of the mean potential outcomes of defiers and always or never takers, respectively. To see this, note that analogously to (15) and (17),

$$\frac{\int yf(y,D=1|Z=0)dy}{\Pr(D=1|Z=0)} = E(Y|D=1,Z=0)=E[Y(1)|T=d]\cdot\frac{\pi_d}{\pi_a+\pi_d}+E[Y(1)|T=a]\cdot\frac{\pi_a}{\pi_a+\pi_d}, \tag{48}$$

$$\frac{\int yf(y,D=0|Z=1)dy}{\Pr(D=0|Z=1)} = E(Y|D=0,Z=1)=E[Y(0)|T=d]\cdot\frac{\pi_d}{\pi_n+\pi_d}+E[Y(0)|T=n]\cdot\frac{\pi_n}{\pi_n+\pi_d}. \tag{49}$$

I.e., to perform sensitivity analysis in the absence of monotonicity, the parameters to be manipulated are the suspected share of defiers (which deterministically sets the share of any other type) as well as the mean potential outcomes of the always takers and compliers under treatment and of the never takers and compliers under non-treatment. Concerning the latter problem, one may express the potential outcome of one population as proportion of that of the other population in the respective mixture. I.e., define $E[Y(1)|T=a]=\rho_a E[Y(1)|T=c]$ with the ratio $\rho_a=E[Y(1)|T=a]/E[Y(1)|T=c]$ and $E[Y(0)|T=n]=\rho_n E[Y(0)|T=c]$ with the ratio $\rho_n=E[Y(0)|T=n]/E[Y(0)|T=c]$. It follows that (31) and (32) can be rewritten in terms

of the mean potential outcomes of the compliers only:

$$E(Y|D=1,Z=1) = E[Y(1)|T=c]\cdot\frac{\pi_c}{\pi_a+\pi_c}+\rho_a E[Y(1)|T=c]\cdot\frac{\pi_a}{\pi_a+\pi_c}, \quad (50)$$

$$E(Y|D=0,Z=0) = E[Y(0)|T=c]\cdot\frac{\pi_c}{\pi_n+\pi_c}+\rho_n E[Y(0)|T=c]\cdot\frac{\pi_n}{\pi_n+\pi_c}. \quad (51)$$

Solving for the potential outcomes of the compliers yields

$$E[Y(1)|T=c] = \frac{(\pi_a+\pi_c)\cdot E(Y|D=1,Z=1)}{\rho_a\pi_a+\pi_c}, \quad (52)$$

$$E[Y(0)|T=c] = \frac{(\pi_n+\pi_c)\cdot E(Y|D=0,Z=0)}{\rho_n\pi_n+\pi_c}. \quad (53)$$

By varying ρ_a, ρ_n over a plausible range, the robustness of Δ_c can be assessed. E.g., it is easy to see that $E(Y|D=1,Z=1)=E[Y(1)|T=c]$ if $\rho_a=1$, implying that the mean potential outcomes of always takers and compliers are equal. As a natural starting point for the choice of ρ_a and ρ_n, one may consider

$$\phi_a = \frac{E(Y|D=1,Z=0)}{\frac{\Pr(D=1|Z=1)\cdot E(Y|D=1,Z=1)-\Pr(D=1|Z=0)\cdot E(Y|D=1,Z=0)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}}, \quad (54)$$

$$\phi_n = \frac{E(Y|D=0,Z=1)}{\frac{\Pr(D=0|Z=0)\cdot E(Y|D=0,Z=0)-\Pr(D=0|Z=1)\cdot E(Y|D=0,Z=1)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}}. \quad (55)$$

ϕ_a and ϕ_n are the ratios of the observed quantities in (31) and (19) as well as in (32) and (22), respectively, which correspond to ρ_a and ρ_n if Assumption 1 and 2 are satisfied. Therefore, under moderate deviations from monotonicity, ϕ_a and ϕ_n are most likely not too different from the actual ratios ρ_a and ρ_n. By defining

$$\rho_a = \beta_a\phi_a, \qquad \rho_n = \beta_n\phi_n, \quad (56)$$

we can use any $\beta_a > 0$, $\beta_n > 0$ different to unity to gauge the supposed violation as a function of the empirically advised outcome ratios. E.g., $0 < \beta_a < 1$ presumes that the mean potential outcome of the always takers under treatment is in reality relatively smaller compared to that of

the compliers than the empirical data makes us believe when considering ϕ_a and incorrectly assuming monotonicity. Likewise, $\beta_n = 1.2$ presumes that the ratio of the mean potential outcomes of never takers and compliers under non-treatment is 20% higher than suggested by the data.

On top of β_a, β_n (or ρ_a, ρ_n directly), we also need to incorporate uncertainty about the defier share π_d. By considering (45), (46), (47), and the fact that $\Pr(D = 1|Z = 1) = \pi_a + \pi_c$ (see equation (16)) and $\Pr(D = 0|Z = 0) = \pi_n + \pi_c$ in the definitions of the potential outcomes of the compliers given in (52) and (53), we obtain

$$E[Y(1)|T = c] = \frac{\Pr(D = 1|Z = 1) \cdot E(Y|D = 1, Z = 1)}{\beta_a \phi_a (\Pr(D = 1|Z = 0) - \pi_d) + \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0) + \pi_d}, \tag{57}$$

$$E[Y(0)|T = c] = \frac{\Pr(D = 0|Z = 0) \cdot E(Y|D = 0, Z = 0)}{\beta_n \phi_n (\Pr(D = 0|Z = 1) - \pi_d) + \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0) + \pi_d}, \tag{58}$$

where we have also made use of (56). Conducting sensitivity analysis w.r.t. to non-monotonicity therefore requires varying the three parameters β_a, β_n, (or ρ_a, ρ_n) and π_d over a plausible range. In this context, it is important to note that the results of Huber and Mellace (2010) and Richardson and Robins (2010) can be used to obtain "worst case" upper and lower bounds for each parameter value under the violation of Assumption 2 and the satisfaction of Assumption 1. Clearly, the range of values considered must lie within these bounds to be consistent with the data at hand.

Finally, the sensitivity checks under (separate) violations of Assumptions 1 and 2 can be combined in order to assess deviations from both assumptions jointly. Recalling that the prevalence of a direct effect requires substituting $E(Y|D = 1, Z = 1)$ and $E(Y|D = 0, Z = 1)$ by $E(Y|D = 1, Z = 1) - \gamma$ and $E(Y|D = 0, Z = 1) - \gamma$, respectively, we modify (57), (58), (54), and (55) accordingly to accommodate violations of the exclusion restriction and monotonicity at the same

time. This gives

$$E[Y(1)|T=c] = \frac{\Pr(D=1|Z=1)\cdot(E(Y|D=1,Z=1)-\gamma)}{\beta_a\phi_a^\gamma(\Pr(D=1|Z=0)-\pi_d)+\Pr(D=1|Z=1)-\Pr(D=1|Z=0)+\pi_d} \quad (59)$$

$$\text{with } \phi_a^\gamma = \frac{E(Y|D=1,Z=0)}{\frac{\Pr(D=1|Z=1)\cdot(E(Y|D=1,Z=1)-\gamma)-\Pr(D=1|Z=0)\cdot E(Y|D=1,Z=0)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}},$$

$$E[Y(0)|T=c] = \frac{\Pr(D=0|Z=0)\cdot E(Y|D=0,Z=0)}{\beta_n\phi_n^\gamma(\Pr(D=0|Z=1)-\pi_d)+\Pr(D=1|Z=1)-\Pr(D=1|Z=0)+\pi_d} \quad (60)$$

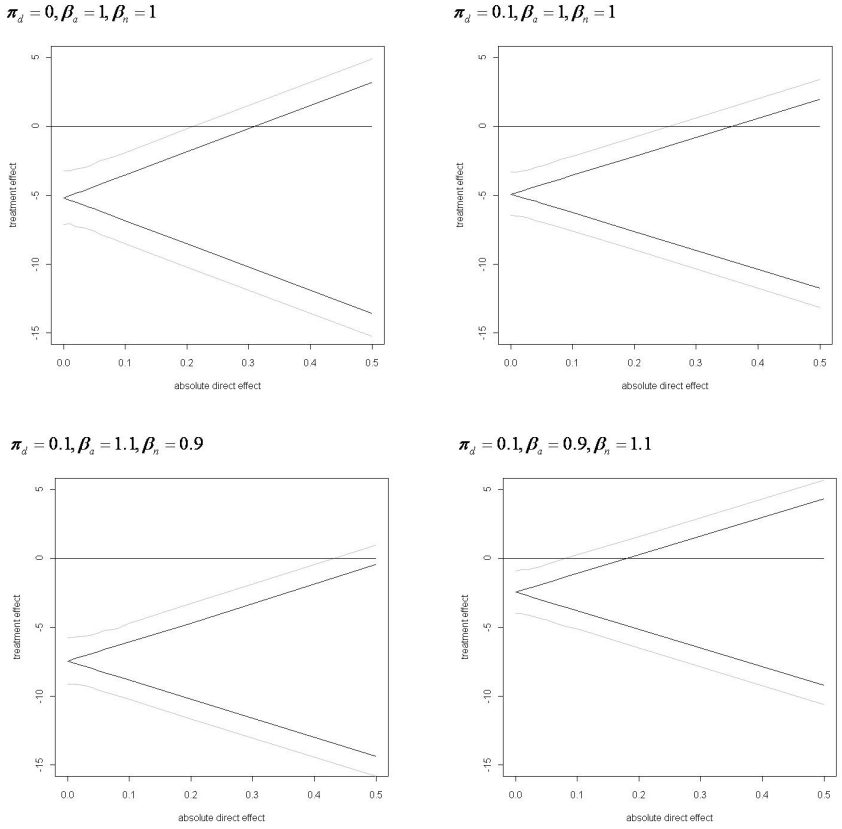
$$\text{with } \phi_n^\gamma = \frac{E(Y|D=0,Z=1)}{\frac{\Pr(D=0|Z=0)\cdot E(Y|D=0,Z=0)-\Pr(D=0|Z=1)\cdot(E(Y|D=0,Z=1)-\gamma)}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}}.$$

Expressions (59) and (60) for the first time allow investigating the robustness of LATE estimation to the non-satisfaction of both identifying assumptions. Therefore, they imply (40), (41), when only Assumption 1 is violated, (57), (58), when only Assumption 2 is violated, and (31), (32), when both assumptions hold, as special cases.

We use the sensitivity checks to investigate moderate violations of the validity of the sibling sex ratio instrument in the full sample of the 1980 wave of the PUMS data. Figure 3 shows the results for various choices of the direct effect γ and the parameters gauging non-monotonicity, π_d, β_a, and β_n. All graphs plot the upper and lower bounds of the LATE on the y-axis as a function of the presumed absolute value of the direct effect on the x-axis to allow for both positive and negative violations of the exclusion restriction, but differ w.r.t. π_d, β_a, and β_n. The bounds are given by the black lines, while the grey lines represent the 95% confidence intervals of the LATE, which are weakly tighter than those of the identified set (i.e., the entire region between the lower and upper bound). To construct these confidence intervals, we use the method proposed by Imbens and Manski (2004) in their equations (6) and (7) and estimate the standard errors of the bounds required therein by bootstrapping 1999 times. As discussed in Imbens and Manski (2004) and Stoye (2009), to obtain asymptotically correct coverage, the estimated upper and lower bounds of the LATE need to (i) satisfy asymptotic normality, (ii) have bounded second moments,

and (iii) be almost surely ordered such that the estimated upper bound always weakly dominates the lower one. Given that the Wald estimator is asymptotically normal and that the difference of (59) and (60) corresponds to the probability limit of a Wald estimator that is perturbed by some constants γ, π_d, β_a, and β_n, (i) and (ii) appear innocuous. Furthermore, (iii) holds by construction, because for π_d, β_a, and β_n fixed, the spread between the estimated upper and the lower bounds is a deterministic function of γ.

Figure 3: Bounds on the LATE (black lines) and 95% confidence intervals (grey lines)



The upper left graph of Figure 3 shows the sensitivity of the LATE under violations of the exclusion restriction, while monotonicity is assumed to hold. It is obvious from (42) that the LATE increases as the direct effect decreases. For γ's larger than -0.21 hours per week, which corresponds to roughly 4% of the LATE estimate if Assumptions 1 and 2 are satisfied, the LATE remains significantly negative. In the upper right graph, monotonicity is supposedly violated by setting the defier share π_d to 0.1. Note that by the definition (47), the share of compliers increases

in π_d. This implies that the ITT is now scaled by a larger complier population than before such that the absolute magnitude of the LATE ceteris paribus becomes smaller for any value of γ. Therefore, the LATE is slightly shifted towards zero for $\gamma = 0$ and the bounds are narrower than before for any $\gamma \neq 0$, because β_a and β_n are kept at unity just as in the initial set up. The latter implies that the identification of the mean potential outcomes of the compliers, always takers, and never takers is not compromised by the violation of monotonicity. From (48) and (49) we see that this is to assume that the mean potential outcomes of the defiers and always takers under treatment and of the defiers and the never takers under non-treatment are the same. Under these assumptions, the LATE is significantly negative for γ's larger than -0.25.

In the lower two graphs the potential outcomes of the compliers, always takers, and never takers are assumed to be affected by non-monotonicity by setting β_a, and β_n to values different than unity. In the left graph, the ratio of the mean potential outcomes of the always takers and compliers under treatment is supposedly 110% of that suggested by the data if monotonicity was satisfied. Under non-treatment, the mean potential outcome of the never takers is assumed to be in reality relatively smaller compared to the compliers than one might suspect from the data. As these assumptions reduce the mean potential outcome of the compliers under treatment and increase that under non-treatment, the LATE becomes more negative and remains significant even if γ is as low as -0.4. Finally, the graph on the right presents the reversed picture. There, the mean potential outcome is increased under treatment and reduced under non-treatment by setting $\beta_a = 0.9$ and $\beta_n = 1.1$. This shifts the LATE upwards and makes it insignificantly different from zero for any γ smaller than -0.09.

The sensitivity of the results in particular to γ, β_a, β_n raises the important question of how to choose values that appear realistic in practice. Concerning γ, we would suspect that increased costs of mixed sex siblings, for instance due to the lacking possibility of room-sharing and hand-me-downs, should if anything induce women to provide more labor to increase household income. This suggests the direct effect of the instrument to be non-negative, implying that any LATE

estimate with $\gamma = 0$ represents an upper bound for the true effect.[10] This argument is strongly in favor of a negative effect, as it requires β_a and β_n to be even smaller and larger, respectively, to overthrow negativity. We therefore also discuss which choices of the latter parameters appear reasonable. In this context, note that the estimated ϕ_a=1.11, suggesting that if IV validity holds, always takers work more hours than compliers under treatment. Furthermore, the estimated ϕ_n=1.02, implying that never taker and compliers are fairly comparable under non-treatment.

We will now take a look at labor market relevant covariates to judge whether this ordering of potential outcomes seems plausible. In fact, one can estimate the mean values of the covariates across subpopulations by treating them as "outcomes" in equations (31), (32), (19), and (22). This of course hinges on the satisfaction of Assumptions 1 and 2 w.r.t. to the covariates. If the instrument is indeed randomly assigned, we would typically think that Assumption 1 is more likely to hold when considering covariates rather than (post-treatment) outcomes, because the instrument cannot have a direct effect on covariates if the latter are predetermined. However, the satisfaction of Assumption 2 is not a function of the outcome considered. If it does not hold, it will affect our analysis of the covariates just as much as it jeopardizes LATE estimation. Therefore, we are in a sense trying to fit a square peg into a round hole: We would like to base the sensitivity analysis of the LATE on the means of the covariates across subspopulations, but identification of the latter may be complicated (at least) by non-monotonicity concerns. In this context, it is interesting to note that testing whether the "LATE" on the (supposedly predetermined) covariates of the compliers is zero provides a check for violations, because predetermined covariates must not be influenced by treatments occurring later in time. Table 5 presents the mean values of education, age, and marital status across subpopulations. Notably, the LATE given in the last column is never significant, suggesting that the covariate means of the compliers across treatments, see columns 3 and 6, are not statistically different. While this does not completely rule out violations given the non-negligible magnitude of the standard errors, it suggests that they are not too large

[10] A counter-argument is that mixed sex siblings are ceteris paribus more time consuming due to sex-specific activities such that females may be induced to substitute time at work by time spent with children. This would result in a negative direct effect on labor supply. In this case, assuming $\gamma \geq 0$ only holds if the cost argument on average outweighs the time substitution argument.

and hopefully small enough to justify our subsequent reasoning.

Table 5: Differences in observed covariates

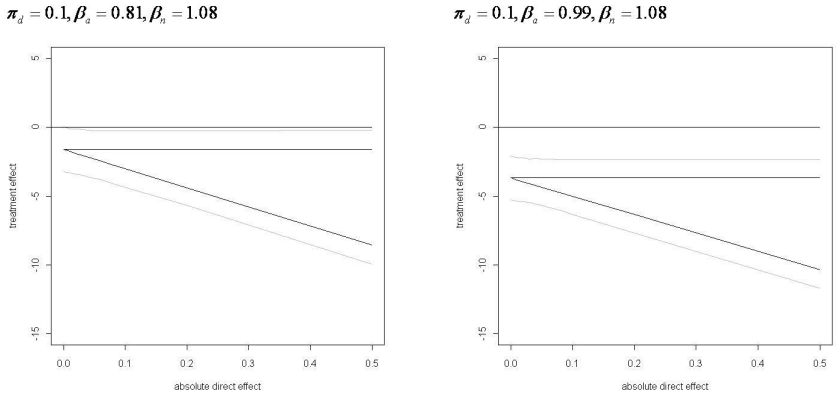
	D=1			D=0			
Variable	always takers	compliers	diff	never takers	compliers	diff	LATE
Education	11.646	12.041	-0.395**	12.446	12.093	0.353**	-0.052
Age	30.559	30.346	0.213**	29.804	30.618	-0.815**	-0.272
Marital status	0.848	0.897	-0.048**	0.850	0.887	-0.037**	0.010

Note: **: Significant at the 1% level. *: Significant at the 5% level. +: Significant at the 10% level.

When taking the results in Table 5 at face value, we see that always takers are somewhat older, less educated, and less likely to be married than compliers. Albeit these differences are not huge, the are statistically significant. The empirical literature finds that female labor supply in the 20th century is positively related to schooling, but negatively to being married, see for instance Killingsworth and Heckman (1986). As it appears difficult to argue which relationship is relatively more important, we cannot be conclusive about whether the mean potential outcome of the always takers should dominate that of compliers or vice versa. Also, since the differences in the labor market relevant characteristics are not striking, we would not expect the two groups to be too different in terms of potential hours worked. This motivates choices of β_a within a rather narrow range that entails ρ_a's that are not too different from 1. The never takers are found to have a somewhat higher level of education and a lower probability to be married than the compliers, even though the differences with the compliers are again not thrillingly large. We would therefore suspect the never takers to have a somewhat higher labor market attachment given the slightly more favorable characteristics and their choice not to have further children irrespective of the sex ratio, which may point to a higher valuation of work life. Therefore, setting β_n to a value such that ρ_n is moderately larger than 1 appears plausible.

Figure 4 presents our preferred choices for conducting sensitivity analysis. In either graph, γ is (in line with our previous argumentation) assumed to be non-negative and $\beta_n = 1.08$. The latter implies that $\rho_n = 1.1$, because of the supposedly higher labor market attachment of the never takers. In the left graph, $\beta_a = 0.81$ such that $\rho_a = 0.9$, assuming that the mean potential outcome of the always takers is at least 90% of that of the compliers. Even if so, the negative effect

Figure 4: Bounds on the LATE (black lines) and 95% confidence intervals (grey lines)



of fertility on female labor supply remains marginally significant. In the right graph, $\beta_a = 0.99$ and $\rho_a = 1.1$, assuming that the mean potential outcome of the always takers is at least 10% larger than that of the compliers. This entails somewhat more negative effects for any value of γ. Finally, it is worth mentioning that the results are qualitatively robust to the choice of π_d because of the upper bound on γ being zero. Therefore, π_d never affects the upper bound, but increases the lower bound as it grows.[11] We conclude from our sensitivity analysis that the negativity of the effect of fertility on female labor supply in Angrist and Evans (1998) is robust to a plausible range of violations of IV validity.

5 Conclusion

In this paper, we have revisited the sibling sex ratio instrument of Angrist and Evans (1998) and for the first time applied statistical hypothesis tests to verify IV validity in the nonparametric LATE framework directly rather than through theoretical reasoning or empirically motivated back of the envelope calculations. Using the methods proposed in Kitagawa (2008), Huber and Mellace (2011a), and Huber and Mellace (2012), the tests do not provide strong evidence against the validity of the instrument. This arguably rules out severe violations which would have given

[11] Given that the instrument is randomly assigned w.r.t. treatment, the upper bound on the share of defiers that is consistent with the data is $\pi_d = 0.372$, see Huber and Mellace (2010), which is thus the admissible value that minimizes the length of the identified set.

more power to the tests. However, small deviations from the exclusion restriction and/or the monotonicity of the treatment in the instrument may nevertheless exist given the small p-values of a subset of tests in some subsamples of the 1980 US Census data considered. For this reason, we have proposed new and easily implementable sensitivity checks for LATE estimation which for the first time allow (i) assessing violations of both the exclusion restriction and monotonicity and (ii) incorporating heterogeneity in the violations across different types (e.g., compliers and always takers). The results suggest that the negativity of the effect of fertility on female labor supply found in Angrist and Evans (1998) is robust to moderate deviations from IV validity and therefore corroborate the findings in their study.

While the application considered in this paper is an example of the binary treatment and binary instrument case, testing and the sensitivity checks could be easily extended to settings with a binary treatment and a multivalued instrument with bounded support. As a future research agenda, the methods could therefore be used to statistically assess a broad range of arguably debatable instruments. Potential applications include electoral cycles instruments as suggested in Levitt (1997), instruments based on changes in public policies such as compulsory schooling laws, see for instance Oreopoulos (2006), geographical distance instruments as used in Card (1995), or further natural "natural experiments" such as the date of birth, see Angrist (1990) and Angrist and Krueger (1991),[12] to give only a small, non-exhaustive list.

[12]Testing in Huber and Mellace (2012) suggests that the quarter of birth instrument in Angrist and Krueger (1991), which has been prominently contested by Bound, Jaeger, and Baker (1995) on theoretical and empirical grounds, is indeed likely to be invalid. Conley, Hansen, and Rossi (2012) and Hoogerheide and van Dijk (2006) provide sensitivity checks under a violation of the exclusion restriction.

A The test algorithm based on Bennett (2009)

We denote by θ the vector of inequality moment constraints. E.g., when testing the mean constraints (24), θ consists of four elements. 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 $n \to \infty$.[13]
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. 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:

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

[13] 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).

References

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.

Andrews, D. W. K. (2000): "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica*, 68.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Cowles Foundation Discussion Paper 1676*.

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Angrist, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," *American Economic Review*, 80, 313–336.

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*.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

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

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

Chen, L.-Y., and J. Szroeter (2009): "Hypothesis testing of multiple inequalities: the method of constraint chaining," *CeMMAP working paper 13/09*.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Choi, J.-y., and M.-j. Lee (2011): "Bounding endogenous regressor coefficients using moment inequalities and generalized instruments," *forthcoming in Statistica Neerlandica*.

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 (2010): "Partial Identification of Local Average Treatment Effects with an Invalid Instrument," *mimeo, University of Florida*.

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

Hansen, P. R. (2005): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23, 365–380.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000): "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

Hoogerheide, L., and H. K. van Dijk (2006): "A reconsideration of the Angrist-Krueger analysis on returns to education," *Report 2006-15 of the Econometric Institute (Erasmus University Rotterdam)*.

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

Huber, M., and G. Mellace (2010): "Sharp IV bounds on average treatment effects under endogeneity and noncompliance," *University of St Gallen, Dept. of Economics Discussion Paper no. 2010-31*.

——— (2011a): "Testing instrument validity for LATE identification based on inequality moment constraints," *University of St Gallen, Dept. of Economics Discussion Paper no. 2011-43*.

——— (2011b): "Testing instrument validity in sample selection models," *University of St Gallen, Dept. of Economics Discussion Paper no. 2011-45*.

——— (2012): "Relaxing monotonicity in the identification of local average treatment effects," *University of St Gallen, Department of Economics Discussion Paper No. 2012-12*.

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 C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

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

Killingsworth, M. R., and J. J. Heckman (1986): "Female labor supply: A survey," vol. 1 of *Handbook of Labor Economics*, pp. 103–204. Elsevier.

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*.

Kraay, A. (2012): "Instrumental variables regressions with uncertain exclusion restrictions: a Bayesian approach," *Journal of Applied Econometrics*, 27, 108–128.

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

Levitt, S. D. (1997): "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," *The American Economic Review*, 87, 270–290.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68, 997–1010.

Mealli, F., and B. Pacini (2012): "Using secondary outcomes and covariates to sharpen inference in instrumental variable settings," *mimeo*.

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 (2008): "Identification with Imperfect Instruments," *NBER Working Paper No. 14434*.

Oreopoulos, P. (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *The American Economic Review*, 96, 152–175.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer-Verlag, New York.

Richardson, T. S., and J. M. Robins (2010): "Analysis of the Binary Instrumental Variable Model," *mimeo*.

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.

Small, D. (2007): "Sensitivity Analysis for Instrumental Variables Regression with Overidentifying Restrictions," *Journal of the American Statistical Association*, 102, 1049–1058.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77, 1299–1315.

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