The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. 316–331.
doi: 10.1111/ectj.12113

Beyond plausibly exogenous

HANS VAN KIPPERSLUIS[†] AND CORNELIUS A. RIETVELD[†]

[†]*Erasmus School of Economics, Erasmus University Rotterdam, Tinbergen Institute, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, The Netherlands.*
E-mail: hvankippersluis@ese.eur.nl, nrietveld@ese.eur.nl

First version received: February 2018; final version accepted: April 2018

Summary We synthesize two recent advances in the literature on instrumental variable (IV) estimation that test and relax the exclusion restriction. Our approach first estimates the direct effect of the IV on the outcome in a subsample for which the IV does not affect the treatment variable. Subsequently, this estimate for the direct effect is used as input for the plausibly exogenous method developed by Conley, Hansen and Rossi. This two-step procedure provides a novel and informed sensitivity analysis for IV estimation. We illustrate the practical use by estimating the causal effect of (a) attending Catholic high school on schooling outcomes and (b) the number of children on female labour supply.

Keywords: *Exclusion restriction, Instrumental variables, Plausibly exogenous.*

1. INTRODUCTION

Instrumental variable (IV) regression is a powerful tool to establish causal effects of a certain treatment variable on a certain outcome variable. Identification relies on an exclusion restriction: the IV affects only the outcome through the channel of the treatment variable of interest. This assumption is often debatable and cannot be formally tested. Not surprisingly, therefore, researchers dedicate considerable time and effort to convince their readership that the proposed IV satisfies the maintained assumption (Conley et al., 2012).

In recent years, two approaches have become popular to detect and investigate sensitivity to violations of the exclusion restriction. First, starting with Bound and Jaeger (2000), and popularized by Altonji et al. (2005) and Angrist et al. (2010), researchers perform an auxiliary regression as an informal test of the exclusion restriction. The intuition is that in a subsample for which the first stage (that is, the effect of the IV on the treatment variable) is zero, the reduced form (that is, the effect of the IV on the outcome) should be zero too if the exclusion restriction is satisfied. This informal test, henceforth called the zero-first-stage test, can never verify the exclusion restriction, but builds confidence that the exclusion restriction is satisfied. A second development is the work by Conley et al. (2012), who proposed the 'plausibly exogenous' method.[1] Conditional on prior information about the violation of the exclusion restriction, this method allows investigation of the robustness of the IV estimator.

[1] Alternative approaches to dealing with violations of the exclusion restriction include Ashley (2009), Berkowitz et al. (2012), Flores and Flores-Lagunes (2012), Hahn and Hausman (2002), Jones (2015), Kang et al. (2016), Kolesár et al. (2015), Kraay (2012), Nevo and Rosen (2012) and Small (2007).

Both approaches are significant contributions, and have become increasingly popular to make IV estimation more transparent and robust. However, when applied independently, both of these approaches have limitations. The zero-first-stage test is a convincing piece of evidence when one fails to reject a zero reduced-form effect, but forces researchers to drop the IV when one does reject. Quite likely, many IVs that appeared to be promising eventually ended up idle when violations of the exclusion restriction were detected in a zero-first-stage test. At the same time, the plausibly exogenous method is extremely useful if the researcher has prior information on the violation of the exclusion restriction, but provides no guidance on how to obtain a plausible prior. As a result, current applications of the plausibly exogenous approach are exclusively used as a broad-brush sensitivity analysis in the absence of reliable prior information (e.g. Dincecco and Prado, 2012, Ding et al., 2009 and Nunn and Wantchekon, 2011).

In this paper, we argue that a synthesis of the zero-first-stage test and the plausibly exogenous approach is a powerful combination that overcomes the limitations of both approaches. After all, whereas the conventional plausibly exogenous approach does not provide any guidance on how to choose the essential input parameter, the zero-first-stage test gives a direct estimate of the required input parameter. In the other direction, if the zero-first-stage test suggests violations of the exclusion restriction, one can correct for these violations using the plausibly exogenous approach.

Our study relates most closely to the approaches of Angrist and Krueger (1994) and Slichter (2014). Angrist and Krueger (1994) anticipate our approach by using the reduced-form estimate among a zero-first-stage group to fix the direct effect of the IV on the outcome. Although similar in spirit, their approach does not allow for any uncertainty in the direct effect and thereby imposes the rather strong assumption of homogenous direct effects. The main focus of Slichter (2014) is on finding a covariate that induces differential first-stage coefficients. For example, a person's intelligence quotient (IQ; covariate) determines how distance to college (IV) affects the college enrollment decision (treatment). The reduced-form effects of the IV on the outcome among those with very low IQ and very high IQ then provide bounds on the direct effect of the IV on the outcome, under the assumption that the instrument strength is independent of the direct effect (Kolesár et al., 2015). Slichter uses these bounds in a sample selection model with distributional assumptions for set identification of the causal effect of interest.

Our procedure uses the zero-first-stage test as an input for the plausibly exogenous approach. This provides a user-friendly way to gauge violations of the exclusion restriction and correct for those violations while doing justice to parameter uncertainty. It thereby provides a well-informed way to perform sensitivity analyses in IV estimation. In a companion epidemiological paper (van Kippersluis and Rietveld, 2017), we applied this idea in the context of genetic variants as instrumental variables. Here we apply the approach in general IV settings, and illustrate our procedure by estimating the effect of (a) attending Catholic high school on schooling outcomes and (b) the number of children on female labour supply.

2. METHODS

2.1. Instrumental variables

Consider an interest in the causal effect β of an endogenous treatment X on an outcome Y. The idea of IV regression is that there is a vector of instrumental variables Z that is known to be correlated with the treatment X, but is assumed to be uncorrelated with other (unobserved)

determinants of the outcome Y. In terms of equations, where we follow the notation of Conley et al. (2012),

$$Y = X\beta + Z\gamma + \varepsilon, \tag{2.1}$$

$$X = Z\Pi + V. \tag{2.2}$$

Here Y is an $N \times 1$ vector of outcomes, X is an $N \times 1$ vector of treatment variables, Z is an $N \times r$ matrix of $r \geq 1$ instrumental variables, ε and V are $N \times 1$ composite error terms including unobserved confounders, N denotes the sample size, β is the effect of interest, Π is the vector of first-stage coefficients, and γ represents the direct effect of the IV on the outcome (i.e. the possible violation of the exclusion restriction). In these equations, exogenous confounders, including a constant, are assumed to be partialled out. The regular IV assumptions are as follows (e.g. Angrist and Pischke, 2015).

(a) Relevance. The instrumental variables Z have an effect on the treatment X: $\Pi \neq 0$.
(b) Independence. The instrumental variables Z are uncorrelated with any confounders of the exposure–outcome relationship.
(c) Exclusion. The instrumental variables Z affect the outcome Y only through the treatment variable X: $\gamma = 0$.

Instrument relevance can easily be assessed using F-tests with well-known rules of thumb (Bound et al., 1995, Staiger and Stock, 1997 and Stock and Yogo, 2005). The independence assumption can be gauged using balancing or overidentifying restrictions tests (Altonji et al., 2005 and Sargan, 1958, 1998), and is sometimes naturally satisfied when the IV is (as good as) randomly assigned (e.g. the Vietnam War lottery draft in Angrist, 1990, and the Oregon Medicaid lottery in Finkelstein et al., 2012), or when genetic variants are used as IVs (e.g. Smith and Ebrahim, 2003). In contrast, the exclusion restriction is more difficult to assess. Whereas traditional IV assumes that γ is exactly equal to 0, violations of the exclusion restriction imply that $\gamma \neq 0$ in (2.1), which leads to biased estimates of the causal effect of interest β.

2.2. *Assessing the exclusion restriction*

A recent stream of research emphasizes the identification of subgroups for which $\Pi = 0$ to test the exclusion restriction. If the first stage is zero, then the reduced-form effect of the instrument on the outcome should be zero too if the exclusion restriction is satisfied. An early example is Bound and Jaeger (1996, 2000), who question the exclusion restriction of the quarter-of-birth instrument that Angrist and Krueger (1991) use to estimate the effect of educational attainment on earnings. Bound and Jaeger show that men born in the 19th century, who were not affected by compulsory schooling laws that induce the correlation between quarter-of-birth and educational attainment, also display variation in earnings with respect to quarter-of-birth. This suggests that quarter-of-birth also influences earnings through channels other than just educational attainment and that the exclusion restriction is violated.

Similarly, Altonji et al. (2005) investigate the validity of the instrument 'being Catholic' to study the effect of attending a Catholic high school on a wide variety of outcomes. They identify a subsample of public eighth graders among which practically nobody subsequently attends a Catholic high school. Hence, among this subsample the first stage is zero, and any association between the IV (being Catholic) and the outcome reflects a direct effect, indicating a violation of the exclusion restriction. Here too, Altonji et al. find an association between being Catholic and

the relevant outcomes even in the sample of public eight graders, which leads them to conclude that the IV should not be used.

The zero-first-stage test in some cases also provides compelling evidence in favour of the exclusion restriction. For example, Angrist et al. (2010) use Israeli data on twin births and same-sex siblings as IVs for the number of children. They show that Jews of African and Asian origin, as well as mothers who bore their first child at a young age, are less affected by the IVs. In these subsamples, there is no, or a much smaller, effect of the IV on their outcome measures, providing support for their exclusion restriction.

2.3. Beyond plausibly exogenous

In the 'plausibly exogenous' method (Conley et al., 2012), the assumption that $\gamma = 0$ is relaxed, and replaced by a user-specified assumption on a plausible value, range or distribution of γ. Conley et al. propose four different inference approaches, from a frequentist (Uniform) range of values for the parameter γ to a Bayesian approach assuming a specific distribution for the parameter γ. An elegant and user-friendly middle ground, which we focus on here, is obtained when the prior on γ follows a Normal distribution with mean μ_γ and variance Ω_γ, and the uncertainty about γ reduces with the sample size (i.e. 'local-to-zero'). In this case, the plausibly exogenous estimator takes its most convenient form

$$\hat{\beta} \sim N(\beta_{2\text{SLS}} + A\mu_\gamma, W_{2\text{SLS}} + A\Omega_\gamma A'). \tag{2.3}$$

Here $N(\cdot)$ indicates the Normal distribution, $A = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z)$, and $\beta_{2\text{SLS}}$ and $W_{2\text{SLS}}$ denote the traditional two-stage least squares (2SLS) point estimate and variance–covariance matrix, respectively.

Whereas the plausibly exogenous method provides an elegant way to incorporate a non-zero value of γ, it gives no guidance on how to obtain a plausible value, range or distribution of γ. Our innovation is to use the zero-first-stage test as the necessary input. Consider the reduced-form equation that is obtained by substituting (2.2) into (2.1):

$$Y = Z(\gamma + \beta\Pi) + (\varepsilon + \beta V). \tag{2.4}$$

In a subsample for which the first stage is zero ($\Pi = 0$), the reduced-form coefficient of the IV is an estimator for γ. Hence, by first estimating the reduced form (2.4) in a subsample for which $\Pi = 0$, we obtain the estimator $\hat{\gamma}$, which seems a plausible estimate of the direct effect of the IV on the outcome in the full sample, γ. In practice, we therefore suggest setting $\mu_\gamma = \hat{\gamma}$ in the plausibly exogenous equation (2.3) to observe how the causal effect of interest β changes upon a plausible violation of the exclusion restriction.[2] The estimator is easy to obtain in standard software. For example, the user-written command plausexog is readily available in STATA (Clarke, 2014).

In terms of assumptions, whereas this approach relaxes the exclusion restriction, the relevance and independence assumptions should still be satisfied.[3] Moreover, the selection into

[2] Alternatively, one could estimate all equations (jointly) in a Bayesian framework. However, this compromises on the user-friendliness and Conley et al. (2012) present evidence that their Bayesian approach produces very similar results to the 'local-to-zero' approach we adopt here.

[3] The relevance assumption is clearly important here, as the bias due to a violation of the exclusion restriction is amplified by a weak first stage (e.g. Bound et al., 1995). Small and Rosenbaum (2008) run simulations using IVs with varying strength and validity, and come to the conclusion that a slightly biased but strong instrument may be preferable to a less biased but weak instrument.

the zero-first-stage subgroup should not be driven by the IV and the outcome. Finally, we assume homogenous direct effects γ, defined as an equal direct effect of the IV on the outcome in the zero-first-stage group as in the full sample. Whereas this latter assumption seems weaker in many applications than assuming a direct effect of zero as in regular 2SLS, we acknowledge that it imposes a nonstandard set of restrictions: heterogeneous first-stage effects Π, yet homogenous direct effects γ across the zero-first-stage group and the full sample. The assumption seems most viable in case the first stage is zero by construction rather than by choice for a certain subgroup, but in general we caution against placing too much weight on the point estimates deriving from the analysis.[4] Instead, we encourage researchers to use the method as an informed sensitivity analysis by incorporating uncertainty around $\hat{\gamma}$ through specifying non-zero elements in the variance–covariance matrix Ω_γ.

One possible way to incorporate uncertainty is to borrow Imbens and Rubin's (2015) rule of thumb: They suggest that the normalized difference in a covariate between treatment and control groups in a regression setting should not exceed one-quarter (0.25). Here, one could use the same rule of thumb to fix the variance such that the normalized difference in direct effects $\hat{\gamma}$ between the zero-first-stage group and the full sample does not exceed one-quarter in 95% of the cases. In this case, one sets $\Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$, where S_0 is the standard error of $\hat{\gamma}$ in the zero-first-stage group and S_{-0} is the standard error of $\hat{\gamma}$ in the remainder of the analysis sample.[5]

3. EXAMPLES

Altonji et al. (2005) investigate the instrument 'being Catholic' to study the effect of attending a Catholic high school on several schooling outcomes. In their Table 4 they analyse four schooling outcomes: high school graduation, college attendance, twelfth grade reading score and twelfth grade math score. An association is shown between the instrument and the outcomes, even among public eighth graders among whom practically nobody attended Catholic high school. This indicates a violation of the exclusion restriction. Here we show how the effects estimated among public eight graders can be used in the plausibly exogenous method. Details on the data and empirical model can be found in Appendix A.

Table 1 summarizes the regression effect on schooling outcomes of attending a Catholic high school: those who attend Catholic high school on average have better schooling outcomes (row 1) and this advantage is amplified in the 2SLS estimates (row 2). However, for three of the four considered outcomes, the reduced-form effect in the zero-first-stage group ($N = 5{,}649$–$7{,}343$) is significantly different from zero (see Table A.1 in Appendix A). For the twelfth grade reading score, the reduced-form effect is insignificant among public school eighth graders, but for this outcome the ordinary least squares (OLS) and 2SLS estimators are not significant either

[4] An intuitive example is discussed in van Kippersluis and Rietveld (2017), where certain genetic variants are used as IVs for prostate cancer to study its effect on self-reported health. Since prostate cancer naturally is only a risk factor among males, the first stage among females is zero by construction. Hence, females provide a natural zero-first-stage group in this example.

[5] Solving for γ_{-0} in the equation $0.25 = (\hat{\gamma_0} - \gamma_{-0})/\sqrt{S_0^2 + S_{-0}^2}$ and noting that a 95% confidence interval of the Normal distribution has radius $\sim 2\sigma$, we obtain $\sigma^2 = \Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$.

Table 1. Effect of attending Catholic high school on schooling outcome.

	High school graduation ($N = 8{,}802$)	College attendance ($N = 8{,}724$)	Twelfth grade reading score ($N = 6{,}837$)	Twelfth grade math score ($N = 6{,}839$)
OLS	0.051***	0.133***	0.637	0.882***
	(0.008)	(0.020)	(0.329)	(0.250)
2SLS	0.251***	0.408***	0.160	3.745***
	(0.045)	(0.068)	(1.160)	(0.922)
Plausibly exogenous	0.012	0.059	0.425	0.225
	(0.045)	(0.068)	(1.160)	(0.922)
Plausibly exogenous (with uncertainty)	0.012	0.059	0.425	0.225
	(0.046)	(0.071)	(1.219)	(0.967)

Note: Robust standard errors are reported in parentheses. *** p-value ≤ 0.001; ** p-value ≤ 0.01; * p-value ≤ 0.05 (two-sided). The row 'Plausibly exogenous' assumes $\Omega_\gamma = 0$ and '(with uncertainty)' uses $\Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$.

(Table A.1). Consistent with the implied bias computed by Altonji et al., the row 'plausibly exogenous' shows that the effect of attending Catholic high school on schooling outcomes disappears completely when correcting for the estimated direct effect of the IV on the outcome. This illustrates that even without incorporating uncertainty around the direct effect, the size of the direct effect at face value is sufficiently large to substantively alter the conclusions. The result implies that we cannot reject a zero effect of attending a Catholic high school on schooling outcomes, and that the positive OLS coefficients seem to be the result of the selection of comparatively better-performing individuals into Catholic high schools. The analyses that incorporate uncertainty about the direct effect $\hat{\gamma}$ following Imbens and Rubin's rule of thumb (row 4) are in line with these conclusions, and move the p-values even further away from statistical significance.

Table 2. Effect of number of children on female labour supply.

	Working ($N = 2{,}008{,}896$)	Log hours of work ($N = 2{,}008{,}896$)
OLS	−0.047***	−0.352***
	(0.000)	(0.003)
2SLS	−0.029*	−0.235**
	(0.013)	(0.093)
Plausibly exogenous	−0.049***	−0.057
	(0.013)	(0.093)
Plausibly exogenous (with uncertainty)	−0.049***	−0.057
	(0.021)	(0.143)

Note: Robust standard errors are reported in parentheses. *** p-value ≤ 0.001; ** p-value ≤ 0.01; * p-value ≤ 0.05 (two-sided). The row 'Plausibly exogenous' assumes $\Omega_\gamma = 0$ and '(with uncertainty)' uses $\Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$.

Inspired by Angrist et al. (2010), our second example, summarized in Table 2, uses the entire 2014 Dutch population of mothers aged 25–65 who have at least two children ($N = 2{,}008{,}896$) to study the effect of number of children on mother's employment status and hours of work (see Appendix B for more information). The OLS coefficients are negative (Table 2, row 1), suggesting that an additional child reduces the probability of working by 4.7 percentage points (7%) and reduces hours of work by 35%. The IV we consider is whether the first two children were both boys, and Table B.1 in Appendix B indicates that women in the Netherlands have on average 0.065 (0.002, $F = 1814$) more children in case the first two children were boys compared with the case in which the first two children were of mixed sex.[6] The zero-first-stage group ($N = 22{,}548$) comprises women born in countries that, according to the OECD Gender, Institutions, and Development Database (Organisation for Economic Co-operation and Development, 2014), have a strong preference for sons. Indeed, whereas the first-stage effect when using 'two girls' as the IV is strongly significant at 0.241 (0.018) in this group, the first-stage effect using two boys as the IV equals −0.017 (0.017) and is not significant.

The 2SLS estimates (row 2) show that having one more child decreases employment by 2.9 percentage points (4%) and decreases hours of work by 24%. Consistent with the validity of the exclusion restriction, the direct effect of having two boys on employment and hours of work is statistically insignificant for mothers born in countries with preferences for sons (see Table B.1 in Appendix B). The plausibly exogenous approaches (rows 3 and 4) return a significant estimate that is larger in absolute value compared with the 2SLS estimate for the binary indicator of working. The causal effect estimate for hours of work turns insignificant due to a reduction in the coefficient as well as a larger standard error.

4. CONCLUSION

In this paper we synthesized the zero-first-stage test and the plausibly exogenous method. Under the assumptions that (a) the selection into the zero-first-stage subsample is not a consequence of both the instrumental variable and the outcome, and (b) the direct effect of the IV on the outcome is homogenous across the zero-first-stage subsample and the remaining sample, our approach provides an informed way to deal with violations of the exclusion restriction. We acknowledge that these assumptions are strong and impossible to test, but argue that at the very least, the zero-first-stage test provides a natural starting point for the plausibly exogenous approach. Therefore, we feel most comfortable with presenting our two-step procedure as a better-informed sensitivity analysis of IV estimators.

We illustrated our approach with two examples, where in one case, the direct effect of the IV on the outcome was large enough to render the causal effect indistinguishable from zero; in the other case, the direct effect of the IV on the outcome was nonsignificant, leaving our correction arguably superfluous. These examples constitute extreme cases, and we believe there will be many intermediate cases in which this procedure can give a second life to IVs that appeared to be promising but eventually ended up idle when violations of the exclusion restriction were suspected or detected.

[6] In the full sample, the first-stage effect when using 'two girls' as the IV is similar in size 0.071 (0.002). The 2SLS results obtained with this IV are very similar to the results presented in Table 2.

ACKNOWLEDGEMENTS

This work was supported by the Netherlands Organization for Scientific Research (016-145-082 to H. van Kippersluis and 016-165-004 to C. A. Rietveld), the National Institute on Aging (R01AG037398 to H. van Kippersluis), and the Hong Kong Research Grants Council General Research Fund (No.14500815 to H. van Kippersluis). We thank Guido Imbens, participants of the Erasmus Statistics Day, the Editor and three anonymous referees for helpful comments; Taehyun Kim for excellent research assistance; and Statistics Netherlands (CBS) for access to linked data resources (GBAPERSOONSTAB, GBABURGERLIJKESTAATBUS, KINDOUDERTAB, PARTNERBUS, SECMBUS, SPOLISBUS). Researchers who wish to use the data found in this article can apply for access to Statistics Netherlands (CBS). The authors are happy to offer guidance with regard to the application.

REFERENCES

Altonji, J. G., T. E. Elder and C. R. Taber (2000). Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. NBER Working Paper 7831.

Altonji, J. G., T. E. Elder and C. R. Taber (2005). An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling. *Journal of Human Resources 40*, 791–821.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review 80*, 313–36.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics 106*, 979–1014.

Angrist, J. D. and A. B. Krueger (1994). Why do World War II veterans earn more than nonveterans? *Journal of Labor Economics 12*, 74–97.

Angrist, J. D. and J. S. Pischke (2015). *Mastering Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.

Angrist, J. D., 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.

Ashley, R. (2009). Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics 24*, 325–37.

Berkowitz, D., M. Caner and Y. Fang (2012). The validity of instruments revisited. *Journal of Econometrics 166*, 255–66.

Bound, J. and D. A. Jaeger (1996). On the validity of season of birth as an instrument in wage equations: A comment on Angrist and Krueger's 'Does compulsory school attendance affect schooling and earnings?'. NBER Working Paper 5835.

Bound, J. and D. A. Jaeger (2000). Do compulsory school attendance laws alone explain the association between quarter of birth and earnings? In *Research in Labor Economics*, *Volume 19*, 83–108. Bingley, West-Yorkshire: Emerald Group Publishing Limited.

Bound, J., D. A. Jager and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association 90*, 443–50.

Clarke, D. (2014). PLAUSEXOG: Stata module to implement Conley et al.'s plausibly exogenous bounds. *Statistical Software Components S457832*, https://ideas.repec.org/c/boc/bocode/s457832.html.

Conley, T. G., C. B. Hansen and P. E. Rossi (2012). Plausibly exogenous. *Review of Economics and Statistics 94*, 260–72.

Dincecco, M. and M. Prado (2012). Warfare, fiscal capacity, and performance. *Journal of Economic Growth 17*, 171–203.

Ding, W., S. F. Lehrer, J. N. Rosenquist and J. Audrain-McGovern (2009). The impact of poor health on academic performance: new evidence using genetic markers. *Journal of Health Economics 28*, 578–97.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker and Oregon Health Study Group (2012). The Oregon health insurance experiment: evidence from the first year. *Quarterly Journal of Economics 127*, 1057–106.

Flores, C. A. and A. Flores-Lagunes (2012). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business and Economic Statistics 31*, 534–45.

Hahn, J. and J. Hausman (2002). A new specification test for the validity of instrumental variables. *Econometrica 70*, 163–89.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Jones, D. (2015). The economics of exclusion restrictions in IV models. NBER Working Paper 21391.

Kang, H., A. Zhang, T. T. Cai and D. S. Small (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association 111*, 132–44.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser and G. W. Imbens (2015). Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics 33*, 474–84.

Kraay, A. (2012). Instrumental variables regressions with uncertain exclusion restrictions: a Bayesian approach. *Journal of Applied Econometrics 27*, 108–28.

Nevo, A. and A. M. Rosen (2012). Identification with imperfect instruments. *Review of Economics and Statistics 94*, 659–71.

Nunn, N. and L. Wantchekon (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review 101*, 3221–52.

Organisation for Economic Co-operation and Development (2014). Gender, Institutions and Development Database 2014, https://stats.oecd.org/Index.aspx?DataSetCode=GIDDB2014.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica 26*, 393–415.

Sargan, J. D. (1998). Testing for misspecification after estimating using instrumental variables. In *Contributions to Econometrics: John Denis Sargan*. New York, NY: Cambridge University Press.

Slichter, D. (2014). Testing instrument validity and identification with invalid instruments. Department of Economics, University of Rochester, http://www.sole-jole.org/14436.pdf.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association 102*, 1049–58.

Small, D. S. and P. R. Rosenbaum (2008). War and wages. *Journal of the American Statistical Association 103*, 924–33.

Smith, G. D. and S. Ebrahim (2003). Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology 32*, 1–22.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*, 557–86.

Stock, D. and J. H. Yogo (2005). Testing for weak instruments in linear IV regression. In *Andrews DWK Identification and Inference for Econometric Models*, 80–108. New York, NY: Cambridge University Press.

van Kippersluis, H. and C. A. Rietveld (2017). Pleiotropy-robust Mendelian randomization. *International Journal of Epidemiology*, https://doi.org/10.1093/ije/dyx002.

APPENDIX A: ANALYSIS DETAILS FOR EXAMPLE 1

In the first example, we replicate the analysis results presented in Table 4 of Altonji et al. (2005). For this purpose, we analyse data from the National Education Longitudinal Study of 1988 (NELS:88). These data are publicly available (after registration) via the website https://nces.ed.gov/surveys/nels88/. The NELS:88 is a nationally representative sample of eighth graders who were interviewed for the first time in 1988. Follow-up interviews were conducted in 1990, 1992, 1994 and 2000. Altonji et al. (2005) use data from the first three waves, 1988–1994. In 1994, most sample members had completed high school. This data set is referred to as NELS:88/94. We used the descriptions in Altonji et al. (2005) and Appendix B of Altonji et al. (2000) to reproduce the following variables.

Outcomes

- *Twelfth grade reading score*. The twelfth grade reading score is based on variable F22XRTH.
- *Twelfth grade math score*. The twelfth grade math score is based on variable F22XMTH.
- *Enrolled in college in 1994*. The dummy variables for whether students enrolled in a 4-year college as of April 1994 are based on variable ENRL0494.
- *High school graduation*. The dummy variable that indicates whether students received a high school diploma as of 1994 are based on variable HSSTAT.

Main explanatory variable

- *Attending Catholic high school*. Use 1 if yes; use 0 otherwise. Based on variable G10CTRL1.

Instrumental variable

- *Catholic background*. Use 1 if Catholic; use 0 otherwise. Based on variable BYP29.

Control variables

- *Male*. Use 1 if true; use 0 otherwise. Based on variable SEX.
- *Race*. Dummy variables for Black, Asian and Hispanic. Based on variable RACE.
- *Father's education*. Father's years of education. Based on variable BYS34A.
- *Mother's education*. Mother's years of education. Based on variable BYS34B.
- *Family income*. Family income in dollars. Based on variable BYFAMINC.
- *Household composition*. Dummy variable for whether the student lives only with his/her mother. Based on variable BYFCOMP.
- *Parent's marital status*. Dummy variable for whether the parents are married or in a marriage-like relationship; use 0 otherwise. Based on variable BYPARMAR.
- *Urbanicity*. Dummy variables for eighth-grade school in urban, suburban or rural area. Based on variable G8URBAN.

Table A.1. Summary of the regression results.

	High school graduation ($N = 8{,}802$)	College attendance ($N = 8{,}724$)	Twelfth grade reading score ($N = 6{,}837$)	Twelfth grade math score ($N = 6{,}839$)
	Effect of attending Catholic high school on schooling outcomes			
OLS	0.051***	0.133***	0.637	0.882***
	(0.008)	(0.020)	(0.329)	(0.250)
2SLS	0.251***	0.408***	0.160	3.745***
	(0.045)	(0.068)	(1.160)	(0.922)
Plausibly exogenous	0.012	0.059	0.425	0.225
	(0.045)	(0.068)	(1.160)	(0.922)
Plausibly exogenous	0.012	0.059	0.425	0.225
(with uncertainty)	(0.046)	(0.071)	(1.219)	(0.967)
	Effect of Catholic background on schooling outcomes			
Reduced form	0.037***	0.061***	0.025	0.589***
(full sample)	(0.007)	(0.010)	(0.182)	(0.143)
Direct effect	0.036***	0.052***	−0.042	0.554***
(zero-first-stage group)	(0.008)	(0.011)	(0.216)	(0.168)
	$N = 7{,}343$	$N = 7{,}280$	$N = 5{,}649$	$N = 5{,}651$
Direct effect	0.031**	−0.018	−0.312	−0.341
(remaining sample)	(0.012)	(0.025)	(0.422)	(0.327)
	$N = 1{,}459$	$N = 1{,}444$	$N = 1{,}188$	$N = 1{,}188$
	Effect of Catholic background on attending Catholic high school			
First stage	0.149***	0.150***	0.158***	0.157***
(full sample)	(0.007)	(0.007)	(0.008)	(0.008)
First stage	0.009***	0.009***	0.009***	0.009***
(zero-first-stage group)	(0.002)	(0.002)	(0.003)	(0.003)
	$N = 7{,}343$	$N = 7{,}280$	$N = 5{,}649$	$N = 5{,}651$
First stage	0.437***	0.440***	0.434***	0.433***
(remaining sample)	(0.024)	(0.024)	(0.026)	(0.026)
	$N = 1{,}459$	$N = 1{,}444$	$N = 1{,}188$	$N = 1{,}188$

Note: Robust standard errors are reported in parentheses. *** p-value ≤ 0.001; ** p-value ≤ 0.01; * p-value ≤ 0.05 (two-sided). The row 'Plausibly exogenous' assumes $\Omega_\gamma = 0$ and '(with uncertainty)' uses $\Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$; 'remaining sample' indicates the full sample bar the zero-first-stage group.

- *Fighting*. Student got in a fight in eighth grade in the past semester: never (0); once or twice (1); more than twice (2). Based on variable BYS55F.
- *Student rarely completes homework*. Dummy for whether the student rarely completes homework. Based on variable BYT1_3 and BYT4_3.
- *Student frequently disruptive in class*. Dummy for whether the student is frequently disruptive in class. Based on variables BYT1_8 and BYT4_8.

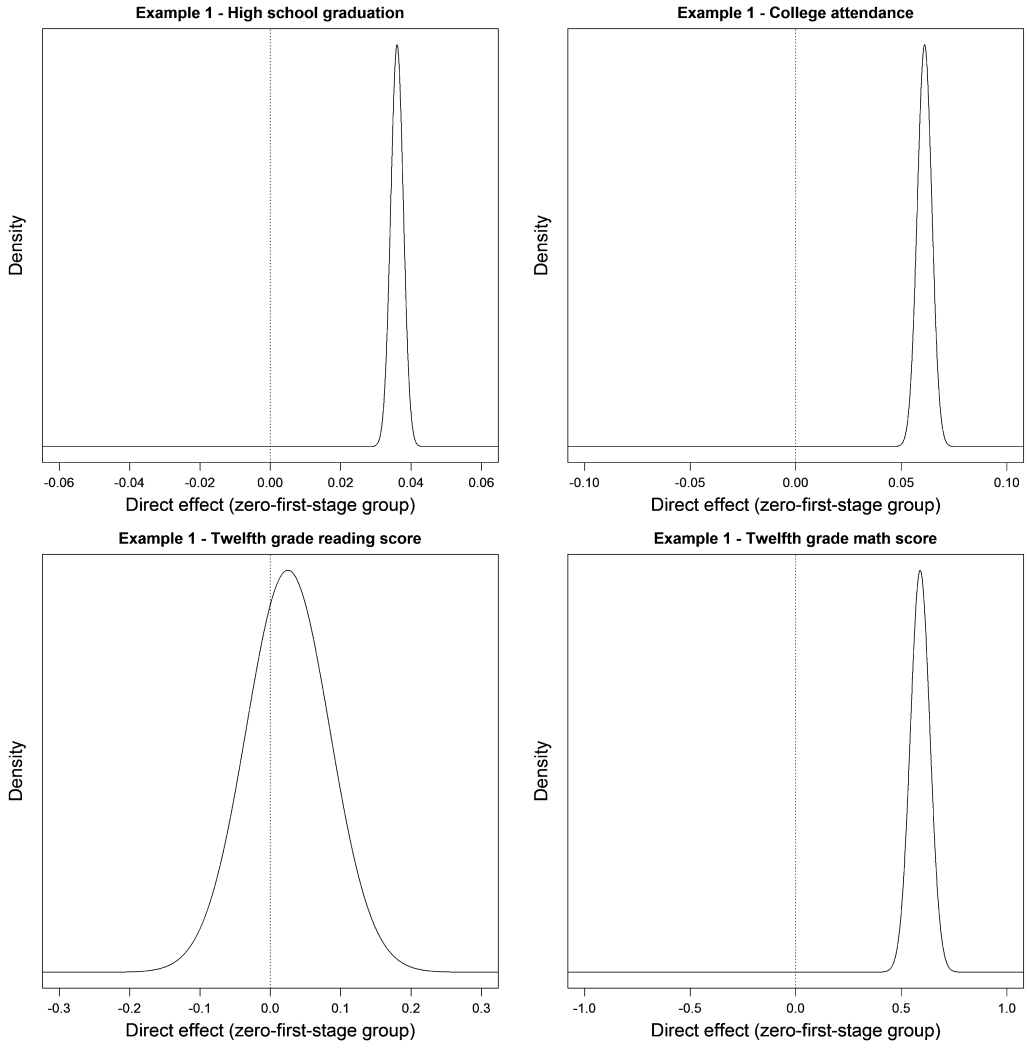

Figure A.1. Distribution of direct effects in Example 1.

- *Delinquency index*. Variable ranging from 0 to 4 that indicates whether the student misbehaved or whether the parents were contacted because of a behaviour problem. Based on variables BYS55A and BYS55E.
- *Repeated grade 4–8*. Dummy variable for whether a student repeated any of grades 4–8. Based on variables BYP46E–BYP46I and BYS74E–BYS74I.
- *Risk index*. Variable ranging from 0 to 6 that indicates the risk of dropping out of school. Based on variable BYRISK.
- *Unpreparedness index*. Variable ranging from 3 to 12 that indicates whether the student comes unprepared to class. Based on variables BYS78A, BYS78B and BYS78C.
- *Grade index*. Variable ranging from 0 to 4 that indicates a composite score for English, mathematics, science and social studies. Based on variable BYGRADS.

– *Eighth-grade reading score*. Eighth-grade reading score. Based on variable BY2XRTH.
– *Eighth-grade math score*. Eighth-grade math score. Based on variable BY2XMTH.

Zero-first-stage group

– Students attending a public eighth grade. Based on variable G8CTRL1.

Table A.1 is an extended version of Table 1 in the main text and additionally includes the first-stage effect (the effect of Catholic background on attending Catholic high school), the reduced-form effect (the effect of Catholic background on the outcome in the full sample), the direct effect (the effect of Catholic background on the outcome in the zero-first-stage group) and the plausibly exogenous results. Although we were not able to replicate the results of Altonji et al. (2005) exactly, our results are generally similar in sign, magnitude and significance. STATA code to reproduce the results is available in the replication package. Figure A.1 displays the distribution of the direct effect estimates used in the plausibly exogenous procedure with uncertainty.

APPENDIX B: ANALYSIS DETAILS EXAMPLE 2

Data for the second example originate from 2014 register data from Statistics Netherlands on the entire Dutch population. In this illustration we use (a) the municipality register for demographic information on gender, country of birth and month of birth (GBAPERSOONSTAB); (b) the intergenerational linkage register to link parents to their children (KINDOUDERTAB); (c) marital status and partner registers (GBABURGERLIJKESTAATBUS; PARTNERBUS); (d) tax register on sources of income (SECMBUS), and hours of work from the so-called SPOLISBUS files. These registers can be linked to each other using a unique personal identifier. These data are proprietary and can only be accessed upon registration.

We restrict the sample to women between 25 and 65 in 2014 with at least 2 and at most 15 children, and who were at least 15 when the first child was born. The mean age of these women is 48, they have on average 2.5 children, the first child was born when the women were on average 27.0 and they were on average 29.9 when the second child was born. 68% of the women were working for an average of 782 hours per year.

Outcomes

– *Working*. Binary indicator of employment status. Use 1 if main source of income throughout the year was work; use 0 otherwise.
– *Log hours of work*. Natural logarithm of the hours of work, where hours of work denotes contractual hours plus paid overtime hours. Hours of work are set to 0 when women are not working. We add 1 to hours of work before taking logarithms.

Main explanatory variable

– *Number of children*. Integer value representing the number of children in 2014.

Table B.1. Summary of the regression results for the effect of number of children on female labour supply.

	Working ($N = 2{,}008{,}896$)	Log hours of work ($N = 2{,}008{,}896$)
	Effect of number of children on labour supply	
OLS	−0.047***	−0.352***
	(0.000)	(0.003)
2SLS	−0.029*	−0.235**
	(0.013)	(0.093)
Plausibly exogenous	−0.049***	−0.057
	(0.013)	(0.093)
Plausibly exogenous	−0.049***	−0.057
(with uncertainty)	(0.021)	(0.143)
	Effect of having two boys on labour supply	
Reduced form	−0.002*	−0.015*
(full sample)	(0.001)	(0.006)
Direct effect	0.001	−0.012
(zero-first-stage group)	(0.008)	(0.056)
	$N = 22{,}548$	$N = 22{,}548$
Direct effect	−0.002*	−0.014*
(remaining sample)	(0.001)	(0.006)
	$N = 1{,}986{,}348$	$N = 1{,}986{,}348$
	Effect of having two boys on number of children	
First stage	0.065***	0.065***
(full sample)	(0.002)	(0.002)
First stage	−0.017	−0.017
(zero-first-stage group)	(0.017)	(0.017)
	$N = 22{,}548$	$N = 22{,}548$
First stage	0.066***	0.066***
(remaining sample)	(0.002)	(0.002)
	$N = 1{,}986{,}348$	$N = 1{,}986{,}348$

Note: Robust standard errors are reported in parentheses. $^{***}p$-value ≤ 0.001; $^{**}p$-value ≤ 0.01; $^{*}p$-value ≤ 0.05 (two-sided). The row 'Plausibly exogenous' assumes $\Omega_\gamma = 0$ and '(with uncertainty)' uses $\Omega_\gamma = (0.125\sqrt{S_0^2 + S_{-0}^2})^2$. The remaining sample indicates the full sample bar the zero-first-stage group.

Instrumental variable

– *Two boys*. Use 1 if the first two children were both boys; use 0 otherwise.

Control variables

– Whereas control variables are not required in this example since the gender of the child is as good as randomly distributed, we follow Angrist et al. (2010) to include the control

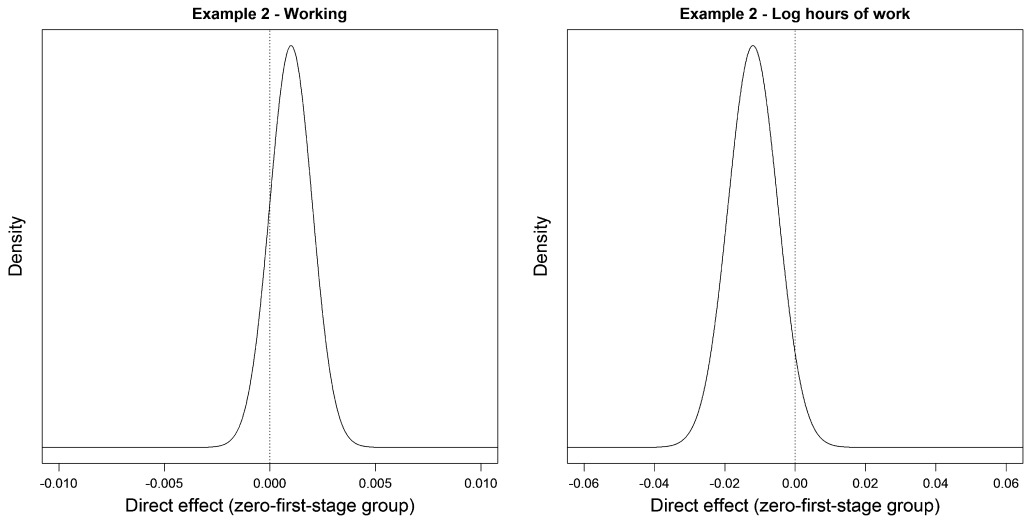

Figure B.1. Distribution of direct effects in Example 2.

variables *year of birth*, *age at birth of first child*, *age at birth of second child*, and *whether the first child was a boy*.

Zero-first-stage group

– The zero-first-stage group is defined as mothers whose country of birth has strong preferences for sons. The Organization for Economic Cooperation and Development (OECD) Gender, Institutions and Development Database 2014 (http://stats.oecd.org/index.aspx?datasetcode=GIDDB2014) ranks countries according to 'fertility preference', which is defined as 'the share of males as the last child from women currently not desiring additional children or sterilised'. We use the top quintile of countries from this list, which comprises the countries Albania, Armenia, Azerbaijan, Bangladesh, Burkina Faso, China, Egypt, Georgia, Guatemala, India, Iraq, Jordan, Kenya, Kyrgyzstan, Macedonia, Nepal, Pakistan, Palestine, Syria, Tajikistan, Tunisia and Uzbekistan. Initially we also included Benin, Ghana, Indonesia, Turkey and Vietnam, but the first-stage estimates on the effect of having two boys on the number of children turned out to be significant among women born in these countries, so we dropped them from the list.

Table B.1 is an extended version of Table 2 in the main text and additionally includes the first-stage effect (the effect of having two boys on number of children), the reduce-form effect (the effect of having two boys on labour supply in the full sample) and the direct effect (the effect of having two boys on labour supply in the zero-first-stage group). STATA code to reproduce the results is available in the replication package. Figure B.1 displays the distribution of the direct effect estimates used in the plausibly exogenous procedure with uncertainty.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Replication files