TI 2017-096/V
Tinbergen Institute Discussion Paper

Beyond Plausibly Exogenous

Hans (J.L.W.) van Kippersluis[1]
Niels (C.A.) Rietveld[2]

[1] Erasmus School of Economics, The Netherlands; Tinbergen Institute, The Netherlands
[2] Erasmus School of Economics, The Netherlands

Beyond Plausibly Exogenous

Hans van Kippersluis[1,2,*] & Cornelius A. Rietveld[1,2]

[1]Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands, [2]Tinbergen Institute, The Netherlands

*Corresponding author. Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA, Rotterdam, The Netherlands. Phone: +31(0)10 408 88 37. Fax: +31 (0)10 408 91 41. E-mail: hvankippersluis@ese.eur.nl.

Abstract

We synthesize two recent advances in the literature on instrumental variables (IVs) 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 (2012). 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 (i) attending Catholic high school on schooling outcomes, and (ii) the number of children on female labour supply.

Key words: Instrumental variables, plausibly exogenous, exclusion restriction

JEL Codes: C18, C26, J20

Acknowledgments: This work was supported by the Netherlands Organization for Scientific Research [016-145-082 to H.v.K. and 016-165-004 to C.A.R.], and the National Institute on Aging [R01AG037398 to H.v.K.]. We thank Guido Imbens and participants of the Erasmus Statistics Day 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 hoping to use the data found in this article can apply for access to Statistics Netherlands (CBS). The authors of this paper are willing to offer guidance about the application.

1. Introduction

Instrumental Variables (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 only affects 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 in convincing their readership that the proposed IV satisfies the maintained assumption (Conley, Hansen & Rossi, 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, Elder and Taber (2005) and Angrist, Lavy and Schlosser (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, from here 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, Hansen & Rossi (2012), who proposed the "plausibly exogenous" method.[1] Conditional on prior information about the violation of the exclusion restriction, this method allows investigating the robustness of the IV estimator.

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 the test is passed, but forces researchers to drop the IV when the test fails. 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 on itself provides no guidance on how to obtain a plausible prior. As a

[1] Alternative approaches to dealing with violations of the exclusion restriction include Hahn & Hausman (2005); Small (2007); Ashley (2009); Berkowitz, Caner & Fang (2012) ; Kraay (2012); Nevo & Rosen (2012); Flores & Flores-Lagunes (2013); Kolesar et al. (2015); Jones (2015); and Kang et al. (2016).

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., Ding, Lehrer, Rosenquist & Audrain-McGovern, 2009; Nunn & Wantchekon, 2011; Dincecco & Prado, 2012).

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 does not have to dismiss the IV but one can correct for violations using the plausibly exogenous approach. Hence, our procedure provides an informed way of performing sensitivity analyses by using the zero-first-stage test as an input for the plausibly exogenous approach.[2] In a companion epidemiological paper (Van Kippersluis & 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 (i) attending Catholic high school on schooling outcomes, and (ii) the number of children on female labour supply.

2. Methods

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

[2] An independently developed and complementary approach can be found in a working paper by Slichter (2015). Whereas his “placebo-test” closely resembles our zero-first-stage test, his main focus is on finding a covariate that induces differential first-stage coefficients. For example, a person’s 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 (Kolesar et al. 2015). Slichter then uses these bounds in a sample selection model with distributional assumptions for set identification (or bounding) of the causal effect of interest.

uncorrelated with other (unobserved) determinants of the outcome *Y*. In terms of equations, where we follow the notation of Conley, Hansen & Rossi (2012):

$$Y = X\beta + Z\gamma + \varepsilon \tag{1}$$

$$X = Z\Pi + V \tag{2}$$

where *Y* is a ($N \times 1$) vector of outcomes, *X* is a ($N \times 1$) vector of treatment variables, *Z* is a ($N \times r$) 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* represent 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 (e.g., Angrist & Pischke, 2015):

1. Relevance: The instrumental variables *Z* have an effect on the treatment *X*: $\Pi \neq 0$.
2. Independence: The instrumental variables *Z* are uncorrelated with any confounders of the exposure-outcome relationship.
3. 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, Jaeger & Baker, 1995; Staiger & Stock, 1997; Stock & Yogo, 2005). The independence assumption can be gauged using balancing or overidentifying restrictions tests (Sargan, 1958; 1975; Altonji, Elder and Taber, 2005), and is sometimes naturally satisfied if the IV is (as good as) randomly assigned – e.g., the Vietnam War lottery draft (Angrist, 1990); the Oregon Medicaid lottery (Finkelstein et al. 2012); or when using genetic variants as IVs (e.g., Davey-Smith & 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 equation (1), which leads to biased estimates of the causal effect of interest *β*.

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 other channels than just educational attainment and that the exclusion restriction is violated.

Similarly, Altonji, Elder and Taber (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 favor of the exclusion restriction. For example, Angrist, Lavy and Schlosser (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 got 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.

Beyond plausibly exogenous

In the "plausibly exogenous" method (Conley, Hansen & Rossi 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:

$$\hat{\beta} \sim N\left(\beta_{2SLS} + A\mu_\gamma, \mathrm{W}_{2\mathrm{SLS}} + A\Omega_\gamma A'\right) \qquad (3)$$

where *N( )* indicates the Normal distribution, $A = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}(X'Z)$, and β_{2SLS} and W_{2SLS} are the traditional Two-Stage Least Squares (2SLS) point estimate and variance-covariance matrix, respectively.

Whereas the plausibly exogenous method provides an elegant way of incorporating 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) into (1):

$$Y = Z(\gamma + \beta\Pi) + (\varepsilon + \beta V) \qquad (4)$$

In a subsample for which the first stage is zero (Π = 0), the reduced form coefficient of the IV is an estimator for γ. Hence, by first estimating the reduced-form (4) in a subsample for which Π = 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 (3) to observe how the causal effect of interest β changes upon a plausible violation of the exclusion restriction.[3] The estimator is easy to obtain in standard software. For example, the user-written command "plausexog" is readily available in STATA (Clarke, 2014).

[3] Whereas this procedure provides a convenient way of obtaining a plausible value for *γ*, it goes somewhat against the frequentist paradigm. More logically consistent, one could estimate all equations (jointly) in a Bayesian framework. However, a Bayesian approach compromises on the user-friendliness, and Conley, Hansen & Rossi (2012) present evidence that their Bayesian approach produces very similar results to the "local-to-zero" approach we adopt here.

In terms of assumptions, whereas this approach relaxes the exclusion restriction, the relevance and independence assumption should still be satisfied. Moreover, the selection into 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. This latter assumption is rather strong and impossible to test. However, the assumption seems weaker in many applications than assuming a direct effect of zero as in 2SLS. Moreover, it is straightforward to incorporate uncertainty around $\hat{\gamma}$ by specifying non-zero elements in the variance-covariance matrix Ω_{γ}.

One possible way of incorporating uncertainty is to borrow Imbens and Rubin (2015)'s 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 Ω_{γ} equal to $\left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$, 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.[4]

3. Examples

Altonji, Elder & Taber (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 analyze 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 which 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 the Appendix.

[4] Solving for γ_F in the equation $0.25 = \frac{\widehat{\gamma_0} - \gamma_F}{\sqrt{S_0^2 + S_{-0}^2}}$, and noting that a 95% confidence interval of the Normal distribution has radius ~2σ, we obtain $\sigma^2 = \Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$.

Table 1 shows that those attending 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 is significantly different from zero (see Table A1 in the Appendix). For the Twelfth grade reading score the reduced form effect is insignificant among public eight graders, but for this outcome the OLS and 2SLS estimators are not significant either (Table A1). 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 direct effect of the IV on the outcome. This 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 & Rubin's rule of thumb (row 4) are in line with these conclusions.

Table 1. Summary of the regression results for the effect of attending a Catholic high school on schooling outcomes. Robust standard errors are reported between parentheses.

	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*** (0.008)	0.133*** (0.020)	0.637 (0.329)	0.882*** (0.250)
2SLS	0.251*** (0.045)	0.408*** (0.068)	0.160 (1.160)	3.745*** (0.922)
Plausibly exogenous	0.012 (0.045)	0.059 (0.068)	0.425 (1.160)	0.225 (0.922)
Plausibly exogenous (with uncertainty)	0.012 (0.046)	0.059 (0.071)	0.425 (1.219)	0.225 (0.967)

*** 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 $\Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$

Inspired by Angrist, Lavy & Schlosser (2010), our second example uses the entire 2014 Dutch population of mothers aged 25-65 with 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 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 percent) and hours of work by 35 percent. The IV we consider is whether the first two children were both boys, and Table A2 in the Appendix indicates that women in the Netherlands have on average 0.065 (0.002, F = 1814) more children in case the first two were boys compared with the case in which the first two were of mixed sex.[5] The zero-first-stage group comprises women born in countries that – according to the OECD Gender, Institutions, and Development Database (OECD, 2014) – have a strong preference for sons. Indeed, whereas the first stage effect when using "two girls" as IV is strongly significant at 0.241 (0.018) in this group, the first stage effect using two boys as IV equals -0.017 (0.017) and is not significant.

Table 2. Summary of the regression results for the effect of number of children on female labour supply. Robust standard errors are reported between parentheses.

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

*** 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 $\Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$

The 2SLS estimates (row 2) show that having one more child decreases employment by 2.9 percentage points (4 percent) and hours of work by 24 percent. 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 son preferences (see Table A2 in the Appendix). 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, but becomes smaller and turns insignificant for hours of work.

[5] In the full sample, the first stage effect when using "two girls" as 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.

4. Conclusion

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

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 non-significant, 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.

References

Altonji, J. G., Elder, T. E., & Taber, C. R. (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*, 313-336.

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

Angrist, J., Lavy, V., & Schlosser, A. (2010). Multiple experiments for the causal link between the quantity and quality of children. *Journal of Labor Economics, 28*, 773-824.

Angrist, J. D., & Pischke, J. S. (2015). *Mastering metrics: The path from cause to effect*. Princeton University Press.

Ashley, R. (2009). Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics, 24*, 325-337.

Berkowitz, D., Caner, M., & Fang, Y. (2012). The validity of instruments revisited. *Journal of Econometrics, 166,* 255-266.

Bound, J., & Jaeger, D. A. (2000). Do compulsory school attendance laws alone explain the association between quarter of birth and earnings?. In *Research in labor economics* (pp. 83-108). Emerald Group Publishing Limited.

Bound, J., & Jaeger, D. A. (1996). On the validity of season of birth as an instrument in wage equations: A comment on Angrist & Krueger's "Does compulsory school attendance affect schooling and earnings?" NBER Working Paper No. 5835.

Bound, J., Jaeger, D. A., & Baker, R. M. (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,* 443-450.

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

Conley, T. G., Hansen, C. B., & Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics, 94*, 260-272.

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

Ding, W., Lehrer, S. F., Rosenquist, J. N., & Audrain-McGovern, J. (2009). The impact of poor health on academic performance: New evidence using genetic markers. *Journal of Health Economics, 28*, 578-597.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., … & Baicker, K. (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics, 127*, 1057-1106.

Flores, C. A., & Flores-Lagunes, A. (2013). Partial identification of local average treatment effects with an invalid instrument. *Journal of Business & Economic Statistics, 31*, 534-545.

Hahn, J., & Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica, 70*, 163–189.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jones, D. (2015). The economics of exclusion restrictions in IV models. NBER Working Paper No. 21391.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association, 111*, 132-144.

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

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., & Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics, 33*, 474-484.

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

Nevo, A., & Rosen, A. M. (2012). Identification with imperfect instruments. *Review of Economics and Statistics, 94*, 659-671.

Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review, 101*, 3221-3252.

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, 1.

Slichter, D. (2015). Testing instrument validity and identification with invalid instruments. Department of Economics, University of Rochester. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxzbGljaHRlcmRhdmlkfGd4OjJkYWM5ZGNmMWFhYWM4MjA.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association,* 102, 1049-1058.

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

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 65*, 557-586.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. *Andrews DWK Identification and Inference for Econometric Models*. New York: Cambridge University Press; pp. 80-108.

Appendix to: Beyond Plausibly Exogenous

1. Analysis details example 1

In the first example we replicate the analysis results presented in Table 4 of Altonji, Elder and Taber (2005). For this purpose, we analyze 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, which were interviewed for the first time in 1988. There were follow-up interviews in 1990, 1992, 1994, and 2000. Altonji et al. use data from the first three waves, 1988-1994. In 1994, most sample members had completed high school. This dataset is referred to as NELS:88/94. We used the descriptions in Altonji, Elder and Taber (2005) and appendix B of Altonji, Elder and Taber (2000) to reproduce the variables as follows:

Outcomes:

- *Twelfth grade reading score*: Twelfth grade reading score. Based on variable F22XRTH.
- *Twelfth grade math score*: Twelfth grade math score. Based on variable F22XMTH.
- *Enrolled in college in 1994*: Dummy variables for whether student enrolled in a 4-year college as of April 1994. Based on variable ENRL0494.
- *High school graduation*: Dummy variable indicating whether student received high school diploma as of 1994. Based on variable HSSTAT.

Main explanatory variable:

- *Attending Catholic high school*: 1 if yes, 0 otherwise. Based on variable G10CTRL1.

Instrumental variable:

- *Catholic background*: 1 if Catholic, 0 otherwise. Based on variable BYP29.

Control variables:

- *Male*: 1 if true, 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 student lives only with his/her mother. Based on variable BYFCOMP.
- *Parent's marital status*: Dummy variable for whether parents are married or in marriage-like relationship, 0 otherwise. Based on variable BYPARMAR.
- *Urbanicity*: Dummy variables for 8th grade school in urban, suburban or rural area. Based on variable G8URBAN.
- *Fighting*: Student got in a fight in 8th 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 student rarely completes homework. Based on variable BYT1_3 and BYT4_3.
- *Student frequently disruptive in class*: Dummy for whether student is frequently disruptive in class. Based on variables BYT1_8 and BYT4_8.
- *Delinquency index*: Variable ranging from 0-4 indicating whether student misbehaved or whether parents were contacted because of a behavior 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 0-6 indicating the risk of dropping out of school. Based on variable BYRISK.
- *Unpreparedness index*: Variable ranging from 3-12 indicating whether the student comes unprepared to class. Based on variables BYS78A, BYS78B, and BYS78C.
- *Grade index*: Variable ranging from 0-4 indicating 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 A1 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, Elder and Taber (2005) exactly, our results are generally similar in sign, magnitude and significance. STATA code to reproduce the results is available upon request.

Table A1. Summary of the regression results for the effect of attending a Catholic high school on schooling outcomes. Robust standard errors are reported between parentheses.

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

*** 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 $\Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$. Remaining sample indicates the full sample bar the zero-first-stage group.

2. 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 (i) the municipality register for demographic information on gender, country of birth, and month of birth ("GBAPERSOONSTAB"); (ii) the intergenerational linkage register to link parents to their children ("KINDOUDERTAB"); (iii) marital status and partner registers ("GBABURGERLIJKESTAATBUS; PARTNERBUS"); (iv) 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. [6]

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. 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 percent was working, for an average of 782 hours per year.

Outcomes:

- *Working*: Binary indicator of employment status. 1 if main source of income throughout the year was work, 0 otherwise.
- *Log hours of work*: Natural logarithm of the hours of work, where hours of work is contractual hours plus paid overwork hours. Hours of work are set to zero 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.

Instrumental variable:

- *Two boys*: 1 if the first two children were both boys, 0 otherwise.

[6] https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research.

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, Lavy and Schlosser (2010) to include the following control variables:
    - *Year of birth*
    - *Age at birth of first child*
    - *Age at birth of second child*
    - *Whether the first child was a boy*

Zero-first-stage group:

- The zero-first-stage group is defined as mothers whose country of birth is one in which son preferences are strong. The OECD Gender, Institutions and Development Database 2014 (http://stats.oecd.org/index.aspx?datasetcode=GIDDB2014) ranks countries according to "fertility preference", 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 these from the list.

Table A2 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 reduced 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 upon request.

Table A2. Summary of the regression results for the effect of number of children on female labour supply. Robust standard errors are reported between parentheses.

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

*** 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 $\Omega_\gamma = \left(0.125\sqrt{S_0^2 + S_{-0}^2}\right)^2$. Remaining sample indicates the full sample bar the zero-first-stage group.