SASCHA O. BECKER
University of Warwick, UK, and IZA, Germany

I Z A
World of Labor
Evidence-based policy making

Using instrumental variables to establish causality

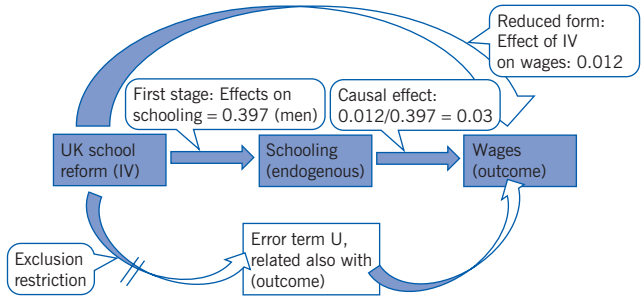
Even with observational data, causality can be recovered with the help of instrumental variables estimation

Keywords: natural experiments, quasi-natural experiments, treatment effects, local average treatment effect, omitted variable bias, reverse causality

ELEVATOR PITCH

Randomized control trials are often considered the gold standard to establish causality. However, in many policy-relevant situations, these trials are not possible. Instrumental variables affect the outcome only via a specific treatment; as such, they allow for the estimation of a causal effect. However, finding valid instruments is difficult. Moreover, instrumental variables estimates recover a causal effect only for a specific part of the population. While those limitations are important, the objective of establishing causality remains; and instrumental variables are an important econometric tool to achieve this objective.

Schematic depiction of IV estimation – Effect of education on wages



Note: A UK reform that increased minimum school leaving age is used as the Instrumental variable (IV); it should affect the outcome only via its effect on the endogenous variable but not in other ways. Numbers are based on [1].

Source: Author's own.

I Z A
World of Labor

KEY FINDINGS

Pros

- Valid instrumental variables help to establish causality, even when using observational data.
- Using instrumental variables helps to address omitted variable bias.
- Instrumental variables can be used to address simultaneity bias.
- To address measurement error in the treatment variable, instrumental variables can be used.

Cons

- Finding strong and valid instrumental variables that affect participation in the treatment but do not have a direct effect on the outcome of interest is difficult.
- Estimated treatment effects do not generally apply to the whole population, nor even to all the treated observations.
- Estimated treatment effects may vary across different instruments.
- For small sample sizes, and in case of "weak" instruments, instrumental variable estimates are biased.

AUTHOR'S MAIN MESSAGE

When treatment is not randomly assigned to participants, the causal effect of the treatment cannot be recovered from simple regression methods. Instrumental variables estimation—a standard econometric tool—can be used to recover the causal effect of the treatment on the outcome. This estimate can be interpreted as a causal effect only for the part of the population whose participation in the treatment was affected by the instrument. Finding a valid instrument that satisfies the two conditions of (i) affecting participation to the treatment, and (ii) not having a direct effect on the outcome, is however far from trivial.

MOTIVATION

Instrumental variables (IV) estimation originates from work on the estimation of supply and demand curves in a market were only equilibrium prices and quantities are observed [2]. A key insight being that in a market where, at the same time, prices depend on quantities and vice versa (reverse causality), one needs instrumental variables (or instruments, for short) that shift the supply but not the demand (or vice versa) to measure how quantities and prices relate. Today, IV is primarily used to solve the problem of "omitted variable bias," referring to incorrect estimates that may occur if important variables such as motivation or ability that explain participation in a treatment cannot be observed in the data. This is useful so as to recover the causal effect of a treatment. In a separate line of enquiry, it is demonstrated that IV can also be used to solve the problem of (classical) measurement error in the treatment variable [3].

DISCUSSION OF PROS AND CONS

Advantages of using instrumental variables to demonstrate causality

As an example, consider the issue of estimating the effect of education on earnings. The simplest estimation technique, ordinary least squares (OLS), generates estimates indicating that one additional year of education is associated with earnings that are 6–10% higher [4]. However, the positive relationship may be driven by self-selection into education; i.e. individuals who have the most to gain from more education are more likely to stay. This will be the case, for example, if pupils with higher ability find studying easier, and would likely receive higher wages anyway. As such, the positive correlation observed between years of education and wages would partially reflect the premium on ability, and could not be interpreted as the returns from an additional year of education, as intended. OLS estimates would thus not be informative about the effect of a policy designed to increase years of education. This problem is called "omitted variable bias." It occurs when a variable (such as ability) that is not observed by the researcher is correlated both with the treatment (more education) and with the outcome (earnings). The direction (over- or underestimation) and size of the bias in OLS estimates is a function of the sign and strength of the correlations.

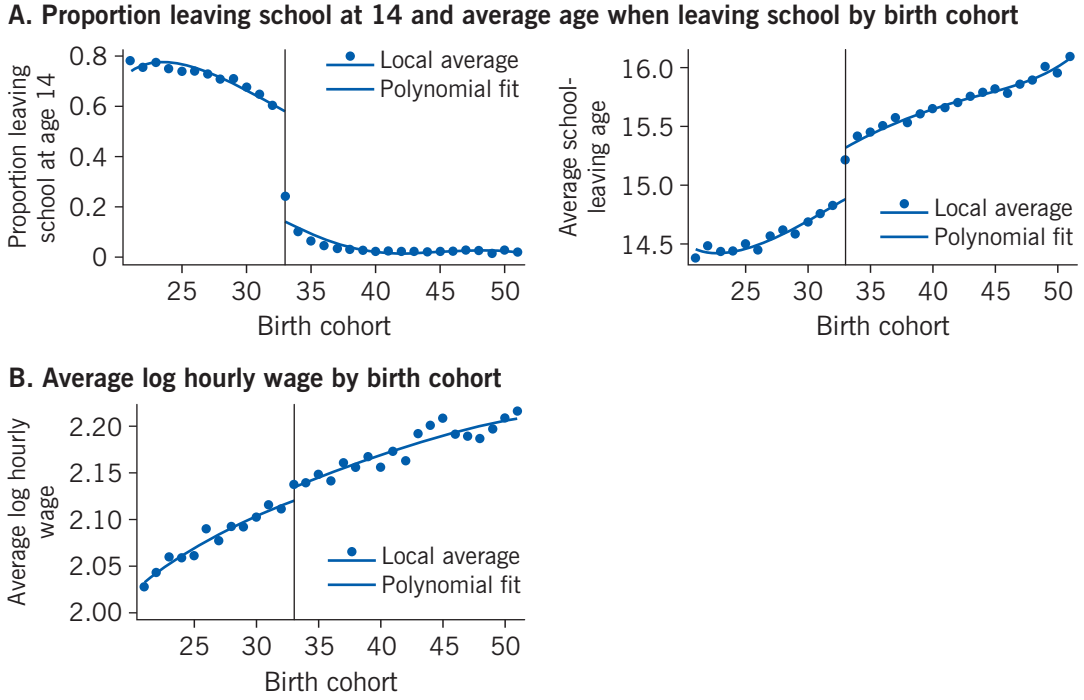
In this example, a randomized control trial (RCT), which would entail allocating education randomly to individuals and observing the differences in their wages over their lifetime, is simply not feasible on ethical grounds. However, some natural or quasi-natural experiments can come close to altering educational choice for some groups of individuals, and as such, can be used as instruments. One such natural experiment is a change in the legal minimum age at which pupils may leave school (school leaving age). This type of change affects all pupils, independent of their ability. It therefore acts like an external shock that cannot be influenced by the individual student.

Numerous countries have legislation stipulating the age at which pupils can leave the educational system. For example, say that a child can leave school on the last day of the school year if she is 14 by the end of August. Let us assume now that the legislation is altered, so that children have to be 15 by the end of August to be allowed to leave school. Children who wanted to leave school at 14 are prevented from doing so, and have to remain for an additional year of schooling. Under the (strong) assumption that children under the two legislations are similar and face similar labor markets conditions,

the legislation change creates a quasi-natural experiment: independently of their ability, some individuals will be affected by the change in school leaving age and have to remain for an additional year of schooling, while pupils with similar preferences from the previous cohort will not. If researchers knew who wanted to leave school at 14, they could compare the outcomes of individuals who left school at 14 to the outcomes of individuals who were forced to stay until 15. This simple difference would then be the causal effect of remaining in school between the ages of 14 and 15. Unfortunately, observational data do not allow us to identify individuals whose educational choice was affected by the reform; so, under the new legislation, individuals who wanted to leave school at 15 are indistinguishable from those who wanted to leave at 14 but had to remain for another year. What the reform does, nonetheless, is to alter the probability of staying in school, and can thus be used as an instrument as it affects the probability of treatment (another year of schooling) without affecting the outcome of interest (e.g. earnings).

In 1947, a legislative change in the UK increased the minimum school leaving age from 14 to 15, affecting children born in 1933 and after. This change in the law provides an opportunity to evaluate the effect of (additional) schooling on earnings [1]. In Figure 1, panel A shows that the reform affected both the fraction of children leaving school at the earliest opportunity (left-hand chart) and the total amount of schooling completed (right-hand chart). Estimates indicate that the reform increased the average years of schooling for men by 0.397 years. This estimate of the effect of the reform (the IV) on

Figure 1. Effect of minimum school leaving age in the UK on men's education and earnings



Note: The vertical line refers to a UK reform that increased the minimum school leaving age from 14 to 15. The reform led to fewer students leaving school at 14, increased the average school leaving age, and increased the average log hourly wages.

Source: Devereux, P., and R. Hart. "Forced to be rich? Returns to compulsory schooling in Britain." *Economic Journal* 120:549 (2010): 1345–1364 [1].

I Z A
World of Labor

the treatment (education) is known as the "first-stage regression." If education has any causal effect on earnings, we should observe that the average earnings of individuals affected by the reform are also higher. This is indeed the case as shown in panel B of Figure 1, which reports the average log earnings for men. This series shows a clear break in 1933, the magnitude of which implies that individuals affected by the reform earn, on average, 1.2% higher wages. This second estimate of the effect of the reform (the IV) on the outcome (earnings) is known as the "reduced form estimate." A simple IV strategy, in this case using a binary instrument that takes on only two values (1 for being affected by the reform, and 0 for not being affected by the reform), is the ratio of the reduced form estimate over the first stage estimate. (This ratio is also known as the Wald estimate.) In this case the causal effect of additional education on earnings would be 0.012/0.397 = 0.030 and thus about 3%.

The intuition of this approach is that the effect of one more year of education on wages is basically the effect of the reform (the IV) on wages (the outcome)—which is given in the reduced form—scaled up by the effect that the reform has on years of education (the treatment)—which is what the first stage estimate is about. *If* the instrument is "relevant," i.e. has an effect on education (the treatment), and *if* the instrument affects wages "exclusively" through its effect on education, then the IV estimates can be interpreted as the causal effect of the treatment on the outcome. These two conditions are called "instrument relevance" and "exclusion restriction."

To summarize, when an unobserved variable such as ability correlates both with the treatment *and* the outcome, a simple estimate like OLS will be biased due to self-selection into the treatment. Similarly, if the treatment variable is measured with error, the OLS estimate will be biased toward zero. However, a causal estimate of a treatment on an outcome can be recovered if a credible instrument can be found. A credible instrument must satisfy two conditions:

- *Relevance*: the instrument must affect the probability of treatment. In a regression of the treatment on the instrument, also known as the first stage equation, the coefficient on the IV must be sufficiently strong.
- *Exclusion restriction*: the instrument affects the outcome exclusively via its effect on the treatment.

If such an IV can be found (i.e. both relevance and exclusion restriction are fulfilled), then an IV strategy can be implemented to recover a causal effect of the treatment on the outcome.

The previous example presented the Wald estimate, i.e. the ratio of estimates from two regressions: the reduced form estimate, coming from a regression of the outcome on the instrument; and the first stage estimate, coming from a regression of the treatment on the instrument. This can easily be computed when the instrument takes only two values. In the more general case, a so-called "two stage least squares" (2SLS) estimate will be computed, whereby predictions of the treatment from the first stage equation are used in a regression of the outcome on the treatment, rather than the true value of the treatment. As such, only the variation in the treatment coming from the instrument is used to explain the variance in the outcome. This then solves the self-selection bias. In the case of a binary (two-value) instrument, the Wald and 2SLS estimates will be identical (see [5], for example). However, the difficulty is not in the implementation

of such a 2SLS estimate, all statistical packages can compute IV estimates, but in (a) finding a valid instrument and (b) interpreting the results. The discussion will now focus on these two points.

Finding a valid instrument

To understand the search process for a valid instrument, the two necessary conditions mentioned above (relevance and exclusion restriction) must be satisfied. The first condition is, in general, easier to satisfy. As illustrated by the previous example, public policy changes can often be a source of promising instruments since they affect the allocation to treatment independently of preferences, like in an RCT. For a policy change to be used as an instrument, it must not have been announced too far in advance of implementation, to ensure that allocation to the treatment is as close to random as possible. In our example, the change in the allocation of the treatment is based on day of birth, and could not have been manipulated after the announcement of the policy change. As such, the allocation to the treatment generated by the instrument is as good as random, at least in the proximity of the policy change; i.e. individuals born in August 1933 are very similar to individuals born in September 1933.

The remaining concern to satisfy the first condition of a credible instrument is that the correlation between the instrument and the change in treatment allocation is strong. An important example of the caveat of relying on "weak instruments" is provided in [6]. Weak instruments, i.e. instruments that are only weakly correlated with the treatment, do not solve the omitted variable bias of OLS estimates [6]. Very weak instruments may induce a bias of the IV/2SLS estimates, which can be even larger than the bias of the OLS estimates. A further study suggests a simple test to reject weak instruments [7].

The second condition (exclusion restriction) for a valid instrument is that the instrument affects the outcome exclusively via its effect on the treatment. Unfortunately, this condition cannot, in general, be statistically tested. It is exactly for this reason that finding a valid instrument is so difficult. Here, econometrics cannot escape economics: Econometric analysis needs to be supported by a convincing economic narrative, which provides credibility to the exclusion restriction. Following our example, one may believe that the change in minimum school leaving age had no direct effect on earnings. However, if we assume that young and old workers are not very good substitutes, employers wanting to recruit 14-year-old workers in 1947 would have faced a severely reduced supply of such workers, and may have had to subsequently increase wages in order to recruit new employees. If starting wages have long-term effects on career development, one could argue that the change in school leaving age is not a good instrument, because the higher starting wage of the few 14-year-olds who left school despite the increased school leaving age would lead to higher wages throughout their career independently of their schooling. However, since worker substitutability is likely to be high, such concerns are probably limited. Yet, the argument shows that instrument validity is not a given, but depends on the context.

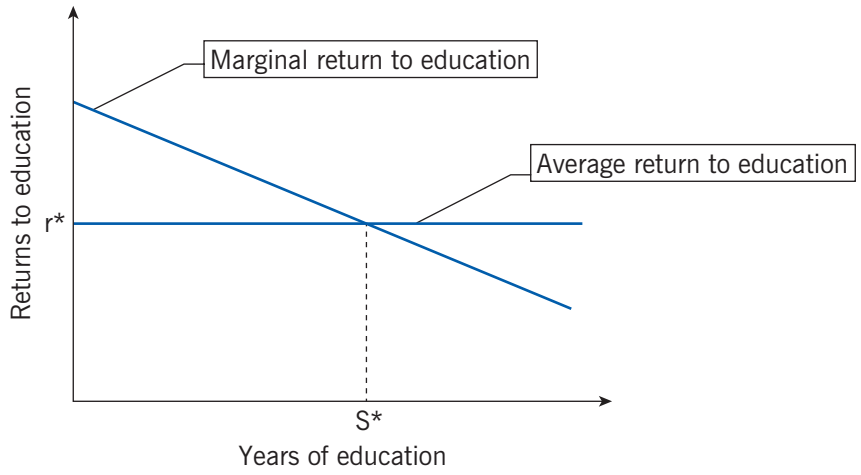
Interpreting IV estimates: Local average treatment effect (LATE)

Assuming that a valid instrument has been found, the remaining difficulty is the interpretation of the IV estimate. Going back to our example of the returns to years of

education in the UK, the IV estimate obtained from using the change in school leaving age was 3% higher wages, only about half the OLS estimate. What could explain this much lower estimate of the returns? The probable answer is that the OLS estimate suffers from omitted variable bias if, for instance, information regarding ability is unobservable. Since ability is positively correlated with both years of education and earnings, its omission from the OLS regression means that the effect of ability on earnings is picked up by the education variable, overestimating the direct effect of education on earnings (upward bias). However, the literature reports several cases of IV estimates of the returns to education that are greater than the OLS estimate (see the review in [4]), how is this possible? One reason is that education is often measured with error, especially in surveys, and that this measurement error in the treatment biases the OLS estimate of the treatment effect toward zero (OLS estimates are "too small"). Since the IV estimate is unaffected by the measurement error in the treatment variable, they tend to be larger than the OLS estimates.

However, the main reason why the IV estimate might be larger than the OLS estimate, even in cases were the omitted variable bias is expected to be the other way round, is that while the OLS estimate describes the *average* difference in earnings for those whose education differs by one year, the IV estimate is the effect of increasing education *only* for the population whose choice of the treatment was *affected* by the instrument (in our example, those 14-year-olds forced to stay in school an additional year who would not otherwise have). This is known as the "local average treatment effect" (LATE). Economic theory predicts that the marginal returns to education (return to one additional year of schooling) decrease with the level of education: so, learning to read has very high returns, but doing a PhD might not do much to increase earnings. This concept is made clearer in Figure 2. At low levels of education (below the average level S*), the return to one additional year of education is greater than the average return (r*). The reverse is true at higher (above average) levels of education. These decreasing returns to education are important when trying to understand why the IV estimate may be larger than the OLS estimate, even in a case where we expect OLS estimates to be upward biased due to omitted variable bias.

Figure 2. Average and marginal return to education



Source: Author's own.

Let us assume that the instrument affects the educational choice of low achievers. The IV estimates indicate a positive effect of additional education for low achievers (below average, left of S* in the figure); for this group, the returns are even greater than for the average population. The situation is reversed when examining an instrument that affects high achievers (i.e. for people with above average education the IV estimate might be lower than the OLS estimate). As such, while it is possible to have one OLS estimate of the returns to education for a given population, different instruments will yield different IV estimates of the returns to education specific to the group affected by the instrument. Rephrasing this statement, we may say that IV estimates have strong "internal validity" (for specific groups) but may have little "external validity" (for the entire population): in our example, the IV recovers the returns to one additional year of education for individuals who wanted to finish school at age 14 in 1947, but were forced to stay for an additional year. This return might be very different from the return to one additional year of education for other cohorts or individuals with a greater taste for education, i.e. one additional year of education later on in life. While this interpretation of the IV estimate may appear very restrictive, it is in fact similar to the interpretation of an RCT, for instance (see [8] for an extensive discussion on external versus interval validity).

The difficulty in interpreting an IV estimate as a local characteristic (i.e. LATE) is that it is not possible to formally identify the individuals whose decision to participate in the treatment was affected by the instrument. Formally, in the case of a binary treatment (cohorts that are affected by the new higher compulsory schooling age, in contrast to cohorts born before 1933 and thus unaffected by the reform) and a binary instrument (school attendance until at least the age of 15, or school attendance only until the age of 14 or less), the population can be divided into four groups, as shown in Figure 3.

Every student can only be one of four types. "Always-takers" are those who leave school at age 15 or above, independently of whether the compulsory schooling age is 14 or 15. "Never-takers" leave school at age 14, independently of whether the compulsory schooling age is 14 or 15; in the example, this group ignores the new legislation and drops out of school anyway. "Compliers" are students who leave school at age 14 when the compulsory schooling age is 14, but they continue to age 15 when the compulsory schooling age is 15. "Defiers" are students who leave school at age 15 or older even when the compulsory schooling age allows them to leave at age 14, but when the compulsory schooling age is 15, they drop out earlier. "Defiers" do the exact opposite of what the law prescribes: less if more is asked, and more if less is asked.

Figure 3. Population group description for binary treatment and binary instrument

		Old regime (school leaving age 14)	
		Leave school at 14	*Leave school after 14*
New regime (school leaving age 15)	*Leave school at 14*	Never-taker	Defier
	Leave school after 14	Complier	Always-taker

Source: Author's own.

I Z A
World of Labor

To be able to interpret an IV estimate as a LATE, an additional assumption must be made on the instrument: monotonicity [9]. The monotonicity assumption states that the instrument pushes some people from no-treatment into taking the treatment (compliers) but nobody in the opposite direction (defiers), i.e. individuals who react to the instrument at all do so in one (intuitive) direction only.

Accordingly, the IV is only informative about the effect of the treatment on the compliers, but cannot identify the effect on always-takers and never-takers, since for these two groups, the treatment choice is unaffected by the instrument (they leave school at 14 or after age 14 independently of the reform). As such, the IV can recover the average treatment effect (the average effect of the treatment on the population) only if the always-taker and never-taker groups are very small and thus (statistically) negligible.

LIMITATIONS AND GAPS

While IV estimates are very helpful tools to measure causal effects, they are not beyond controversy.

As mentioned before, different instruments will identify treatment effects for different subgroups, and we will therefore get numerically different treatment effects. This can also be considered good news if one looks at several different instruments that are informative about treatment effects for different sets of compliers. This point is nicely illustrated in the literature by looking at two different instruments for the same treatment (schooling) [10]. In this example, the first instrument is whether a child attending school during the Second World War had a father engaged in the war. The second instrument is the father's education. The father-in-war instrument is likely to (negatively) affect the schooling of smart children who are constrained because of their father's absence from home. The father's education instrument builds on an intergenerational correlation of education: smarter fathers can help their children get smarter. Having a smart father (as opposed to not) might make more of a difference for the schooling of rich children who are not very smart to begin with. These two instruments affect complier groups at opposite ends of the "returns to schooling" spectrum: the first one should recover the returns for individuals with low levels of schooling (their schooling was reduced due to the absence of the father), while the second identifies the returns for individuals with high levels of education [10]. IV estimates find that the returns to schooling are between 4.8% per year for the father's education IV and 14.0% per year for the father-in-war IV, showing a considerable heterogeneity in returns to schooling (as expected from Figure 2) [10].

While this local estimate (the LATE) helps to clarify what exactly IV estimates, some critics say that it is a controversial parameter because it is defined for an unknown subpopulation [11]. In fact, while we can observe who received the treatment, we cannot distinguish between always-takers and compliers, because we do not know what the treated would have done had the instrument taken a different value. In this case, we have a missing counter-factual. In other cases, the LATE is exactly the parameter policymakers may be interested in, as it reveals the effect of a policy for the individuals affected by the policy.

Another criticism of IV is that, often, one cannot rule out "mild" violations of the exclusion restriction. A recent study, using specific (Bayesian) methods, shows how to assess the

influence of violations of the exclusion restriction on parameter estimates [12]. Bayesian methods are beyond the scope of this article, but it is worth noting that researchers do not have to give up when facing mild violations of the exclusion restriction.

Finally, it is necessary to highlight an additional limitation of IV, which is a bit more on the technical side: IV is consistent but not unbiased. Consistency means that, as estimation samples get larger and larger, IV estimates will converge to the "true" population parameter. Unbiasedness means that, even in finite samples, on average, if we were to draw a series of independent samples from the same population, we would get the "true" population parameter. So, the fact that IV is consistent, but *not* unbiased is troublesome, because any sample is finite. In small samples, IV estimates are unlikely to recover the true effects, and will thus suffer from small sample bias.

SUMMARY AND POLICY ADVICE

Taxpayers support public policies with their own money and have a right to know whether their money is well-spent. Politicians have warmed up to the idea that public policy interventions need to be seriously evaluated. While RCTs are a promising avenue to study the causal effect of treatments on outcomes of interest, they cannot be universally applied to all relevant policy issues. Methods dealing with observational data are thus important, and IV estimation has been a workhorse for empirical research over the last decades. However, finding valid instruments is not easy. Instruments need to fulfill two crucial conditions: they need to be relevant, i.e. significantly correlated with the treatment of interest; and they need to satisfy the exclusion restriction, i.e. they should only affect the outcome via their effect on the treatment. The first condition is testable, but a weak correlation between instrument and treatment is not good enough. Thus, instruments should be sufficiently strong because, otherwise, IV is no better than standard OLS regression. The second condition is fundamentally untestable. We can never exclude the possibility that an instrument affects the outcome above and beyond its effect on the treatment. It is this point that makes IV estimation a matter of debate and controversy. These debates are not merely academic; they are, in fact, crucial if researchers and policymakers are keen to avoid drawing wrong inferences about the direction and size of treatment effects. Nevertheless, with a good instrument, we are able to get reliable estimates of treatment effects that can help influence effective policy.

Acknowledgments

The author thanks two anonymous referees and the IZA World of Labor editors for many helpful suggestions on earlier drafts. The author also thanks Wiji Arulampalam, Clément de Chaisemartin, Andreas Ferrara, and Fabian Waldinger.

Competing interests

The IZA World of Labor project is committed to the *IZA Guiding Principles of Research Integrity*. The author declares to have observed these principles.

REFERENCES

Further reading

Angrist, J. D., and A. B. Krueger. "Instrumental variables and the search for identification: From supply and demand to natural experiments." *Journal of Economic Perspectives* 15:4 (2001): 69–85.

Angrist, J. D., and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Key references

[1] Devereux, P., and R. Hart. "Forced to be rich? Returns to compulsory schooling in Britain." *Economic Journal* 120:549 (2010): 1345–1364.

[2] Wright, P. G. *The Tariff on Animal and Vegetable Oils*. New York: Macmillan, 1928.

[3] Wald, A. "The fitting of straight lines if both variables are subject to error." *Annals of Mathematical Statistics* 11:3 (1940): 284–300.

[4] Card, D. "The causal effect of education on earnings." In: Ashenfelter, O. C., and D. Card (eds). *Handbook of Labor Economics Vol. 3A*. Amsterdam: Elsevier, 1999; pp. 1801–1863.

[5] Angrist, J. D., and A. B. Krueger. "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics* 106:4 (1991): 979–1014.

[6] Bound, J., D. A. Jaeger, and R. M. Baker. "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak." *Journal of the American Statistical Association* 90:430 (1995): 443–450.

[7] Stock, J. H., and M. Yogo. "Testing for weak instruments in linear IV regression." In: Andrews, D. W. K., and J. H. Stock (eds). *Identification and Inference For Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge, UK: Cambridge University Press, 2005.

[8] Shadish, W. R., T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. 2nd edition. Andover, UK: Wadsworth Cengage Learning, 2002.

[9] Angrist, J. D., G. W. Imbens, and D. B. Rubin. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91:434 (1996): 444–455.

[10] Ichino, A., and R. Winter-Ebmer. "Lower and upper bounds of returns to schooling: An Exercise in IV estimation with different instruments." *European Economic Review* 43 (1999): 889–901.

[11] Heckman, J. J. "Identification of causal effects using instrumental variables: Comment." *Journal of the American Statistical Association* 91:434 (1996): 459–462.

[12] Conley, T. G., C. B. Hansen, and P. E. Rossi. "Plausibly exogenous." *Review of Economics and Statistics* 94:1 (2012): 260–272.

Online extras

The **full reference list** for this article is available from:
http://wol.iza.org/articles/using-instrumental-variables-to-establish-causality

View the **evidence map** for this article:
http://wol.iza.org/articles/using-instrumental-variables-to-establish-causality/map