Have Instrumental Variables Brought Us Closer to the Truth

Wei Jiang
Columbia Business School

A survey of 255 papers that rely on the instrumental variable (IV) approach for identifying causal effects published in the "Big Three" finance journals reveals that IV estimates are larger than their corresponding uninstrumented estimates in about 80% of the studies, regardless of whether the potential endogeneity is expected to create a positive or negative bias based on economic reasoning. The magnitude of the IV estimates is, on average, nine times of that of the uninstrumented estimates even when economic insights do not suggest a downward bias of the latter. This study provides several explanations to the "implausibly large" IV estimates in finance research, and proposes best practices for identification-conscientious researchers. (*JEL* G30, C13)

Received January 20, 2017; editorial decision April 7, 2017 by Editor Gregor Matvos

The most important hallmark of contemporary empirical finance research is its emphasis on causal inferences beyond statistical relations. As a result, empirical results derived from archival or observational data are inevitably challenged for potential endogeneity, which prevents a causal interpretation. While wine makers are free to advertise that "a glass of red wine a day keeps atherosclerosis at bay," any good economist would immediately cry "foul," pointing to the inherent endogeneity behind this overly promotional statement. Yes, studies show that people who drink a glass of red wine every day have a lower incidence of vascular disease, but there are innumerable—and unmeasured—variables also at play that directly affect the research. For example, marketers associate drinking a glass of wine with living a "good life," which itself would be positively correlated with good health.

Endogeneity typically arises from two main sources: First, researchers are never able to observe all variables affecting an outcome because

The paper is based on the author's keynote speech at the Society of Financial Studies (SFS) 2015 Cavalcade hosted by Georgia Tech Scheller College of Business. The author thanks the audience at the Cavalcade, Columbia Brownbag, and an anonymous referee for valuable feedback, and, in particular, Matt Spiegel for his encouragement for turning the speech into a paper. Discussions with Alon Brav over the past decade have contributed crucially to the study. Cici Xiao Cen provided valuable research assistance. Send correspondence to Wei Jiang, Columbia Business School, Uris Hall 101, New York, NY 10027; telephone: (212) 854-9002. E-mail: wj2006@columbia.edu.

doi:10.1093/rcfs/cfx015 Advance Access publication June 13, 2017

certain missing variables (e.g., other elements related to living a good life) might simultaneously affect the outcome ("atherosclerosis") and the key independent variable of interest ("red wine"), leading to an "omitted variable bias." Second, economic agents might be more inclined to take an action (e.g., sipping red wine) in anticipation of a likely outcome (a "good life," which could include financial security or personal development), leading to reverse causality (which some references refer to as a "simultaneous causality").[1] In the absence of an exogenous variation in the independent variable of interest, it is generally not possible to separate a causal or treatment effect from alternative hypotheses driven by missing variables or reverse causality.

When denied of the luxury of truly random variations, such as those generated by controlled experiments or exogenous shocks that happen naturally, researchers increasingly resort to the instrumental variable (IV) to "semi-randomize" the treatment variable. In fact, the number of studies that rely mainly on IVs for causal inferences has grown dramatically in the past two decades. Until now though, there has been no research to tell us whether IV has, indeed, brought us closer to truth. This study presents some evidence that might encourage researchers to think more carefully about this question so that the tool could help us achieve the intended goal.

This analysis surveys 255 papers that rely on the IV method for identifying causal effects published in the "Big Three" finance journals (*Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*). Our most notable finding was that, among the published works, the IV estimates almost always exponentially exceed the uninstrumented estimates. That is true even in the majority of the cases in which the uninstrumented estimates were already expected, either by the authors or by the economic context, to have overestimated the "truth" based on the ex ante nature of the endogeneity. As a result, it is hard to argue that the IV estimates in the published works, on the whole, are closer to the true (and unknown) parameters than the simple regression estimates that are potentially tainted by endogeneity.

There could be several possible explanations for this discovery. First, in theory an IV estimate could actually deviate even more than the corresponding uninstrumented estimate from the population average treatment effect in the same direction if an exogenous shock changes the probability of treatment but does not result in uniform assignment between the treated and not-treated. Second, a weak instrument is prone to producing implausibly large estimates, especially when combined with the next explanation. Third, researchers have the incentive to search for

[1] A third potential source of bias is that due to a "measurement error," especially one that is of a non-classical nature (Bound and Krueger 1991). However, measurement error does not arise as a major endogeneity concern in the literature surveyed in this study.

specifications that produce the most significant (or even dramatic and memorable) results.

This paper is not a critique of individual papers. Rather, we hope that the pattern uncovered and reviewed in this paper will draw attention to the limitations of the instrumental variable approach (especially in published studies) while acknowledging its important role as a research tool to uncover causality from observational data. We also hope that our findings will spur individual researchers to go the extra step in interpreting their IV results after the method successfully provides the desired results.

1. Survey of Papers Relying on the IV Method

1.1 The publication history in the "Big Three," 2003–2014

This is an unusual literature review in that the papers reviewed herein are actually the subjects of this project. The following chart plots the number of papers using IV in *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* between 2003 and 2014. The sample includes studies that rely on regression discontinuity as instruments, but does not include shock-based studies built on difference-in-difference analyses.

1.2 Decoding the IV-OLS divergence

1.2.1 Different types of endogeneity. For ease of notation, we use the term ordinary least squares (OLS) throughout this study to refer to the broader class of estimation methods that do not explicitly control for endogeneity and that do not have the benefit of a random shock or natural experiment. Most of the papers surveyed actually use OLS as the main regression model. For each of the 255 papers studied, we try to assess the direction of the OLS bias, or $SignBias_{OLS}$, depending on the a priori direction of the OLS bias due to potential endogeneity. We also sign the true effect of the treatment to be positive.[2] We classify all papers into one of the following three categories. Without loss of generality, we assume that the true treatment effect is non-negative.

1. Affirmative endogeneity: $SignBias_{OLS} > 0$
   In this case, there is a convincing argument that an OLS estimate overestimates the population average treatment effect. The classic example of this is a regression of earnings on years of education. There are two problems with a reduced-form analysis. First, there is the unobserved variable of "ability," which tends to affect both

[2] That is, if the reported coefficient on the key independent variable is negative, we would multiply both the dependent variable and all the coefficients by –1 before proceeding.

earnings and education in the same direction. Second, agents who anticipate better career prospects tend to go to school longer and attain a better education. Both the missing variables and the reverse causality lead the OLS estimate to overstate the true effect. The same analogy can be made when analyzing a company's board of directors. The benefits of specific (e.g., finance) expertise on the board to the company are likely overestimated by OLS because companies that will benefit more from such oversight and guidance are more likely to seek out board members with the specific background needed.

We term this relation as "affirmative endogeneity." Given that in most real-world situations agents are more incentivized to take an action from which they benefit more, a positive $SignBias_{OLS}$ should arise in most cases.

2. Corrective endogeneity: $SignBias_{OLS} < 0$
In this case, the underlying economics suggest that the sample correlation between the outcome and treatment variables understates the true effect. For example, a regression of health conditions on the number of hospital visits during the past year would dramatically underestimate the benefits of hospital care because hospital visits are mostly motivated (or necessitated by) health issues. Similarly, if public school teachers are incentivized to maximize the proportion of students who pass standardized tests or to minimize the incidence of students being "left behind," then they will invest their attention disproportionately on academically weaker students, leading to a spurious low (or even negative) correlation between teacher attention and student academic performance.

We term this relation as "corrective endogeneity." In this case, the underlying economics suggests that the sample correlation between the outcome and treatment variables understates the true effect, or that $SignBias_{OLS}$ is negative.

3. Unclear: $SignBias_{OLS} = 0$
If there is not enough information or there are conflicting forces regarding the nature of the endogeneity, we classify the case to be "unclear," or $SignBias_{OLS} = 0$.

1.2.2 Distribution of endogeneity by type. Admittedly, it is not always obvious if a particular situation falls into "affirmative" or "corrective endogeneity" because researchers often face complicated situations where multiple factors work in different directions. Moreover, the omitted variable bias may not afford a clear directional prediction when it is correlated with multiple independent variables in the regression.

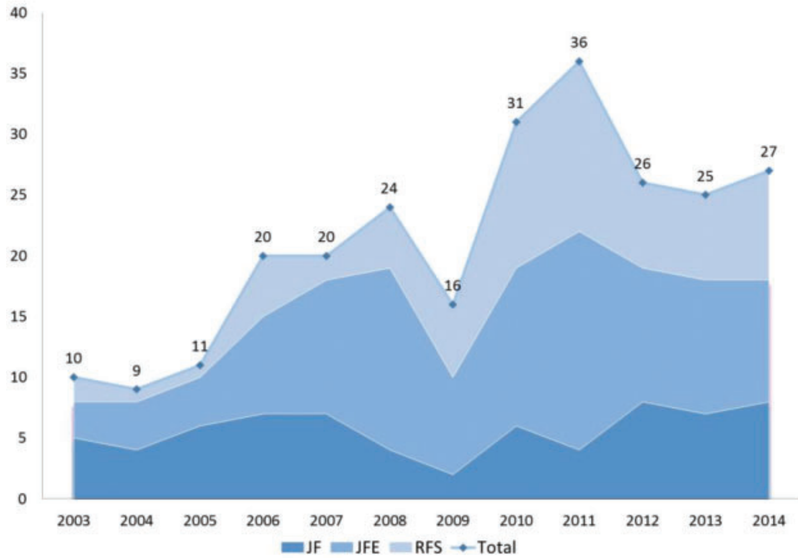

Figure 1
Number of published papers.
This figure plots the number of papers published in *Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies* during 2003–2014 that rely on the IV method for identifying causal relations.

Our classification is based on the economic context that the authors themselves provide to motivate their use of the instruments. For example, if the authors resorted to an instrument because of their concerns that the key independent variable is positively correlated with the residual, then we consider the case to be one of affirmative endogeneity. Based on our reading, 67.1% of the 255 papers are classified as affirmative endogeneity, 18.0% are corrective endogeneity, and 14.9% could be argued either way. Figure 2 shows the percentage of papers with $|\beta_{IV}| > |\beta_{OLS}|$ for each group, as classified by the sign of $SignBias_{OLS}$.

A striking discovery from this investigation is that regardless of what the a priori circumstance is, β_{IV} is predominantly greater than β_{OLS} in magnitude. Even if we acknowledge that the classification of papers in terms of $SignBias_{OLS}$ might be subject to discretion or even prone to errors, there is no reason to expect that the causal effects in close to 85% of all the cases studied by researchers should be predominantly higher than the simple correlational effect; in fact, in a great majority of the studies, the authors motivated their adoption of the instruments because they were concerned that in not doing so the uninstrumented estimates would likely overestimate the true effect.

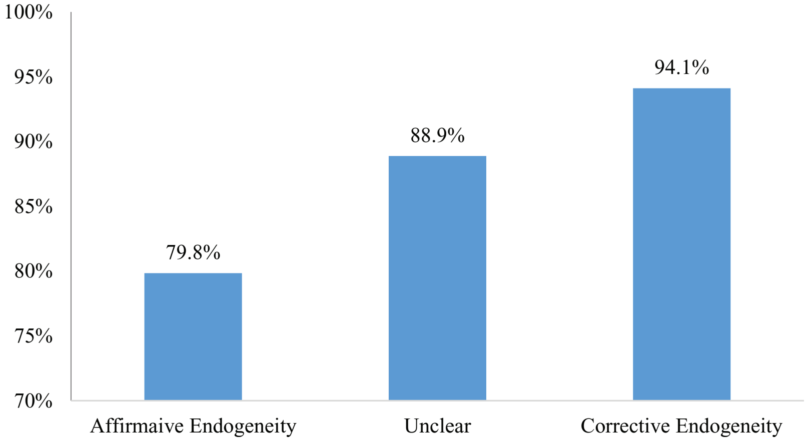

Figure 2
Percentage of papers with $|\beta^{IV}| > |\beta^{OLS}|$.
This chart shows the percentage of papers where the magnitude of IV estimates exceeds that of the OLS estimates, separately for the three categories of endogeneity, based on a priori information and economic reasoning.

An examination of the ratio $|\beta^{IV}/\beta^{OLS}|$ only deepens the puzzle.[3] Figure 3 shows the average ratio of the coefficients in all papers in each category. While studies that fall into the affirmative endogeneity category are expected to have low IV estimates relative to OLS estimates (because the IV method presumably filters out a positive selection effect embedded in the latter), the IV estimates in this category are strikingly larger, on average, in published papers. Even after winsorizing the extremes (at 1%), the magnitude of the IV estimates is, on average, 9.2 times that of the OLS estimates. The average ratios in the other two categories are also significantly higher than unity (three to four times), but they pose no immediate contradiction to the underlying economics.

While our classification method could be subject to human error, it is highly unlikely that the classification based on the authors' stance was systematically perverse. As long as the probability of an affirmative endogeneity in the category is higher than neutral, it is very puzzling that IV estimates turn out to be nine times larger than their OLS counterparts (and larger than the same ratio in the other two categories), despite the fact that the expected positive bias in the OLS estimate was the stated reason that researchers resorted to the IV method in the first place.

[3] We exclude the few cases where β^{OLS} has the opposite sign from β^{IV} and β^{OLS} is statistically significant. We re-sign β^{OLS} if it is of the opposite sign but also insignificant because, presumably, the true correlation could be of either sign given the low significance.

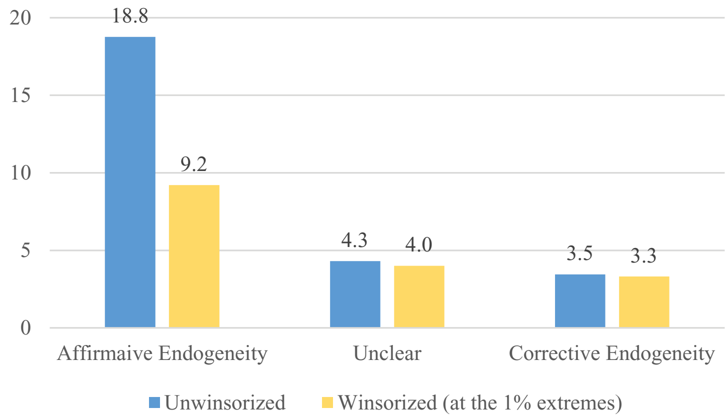

Figure 3
Ratio of $|\beta^{IV}/\beta^{OLS}|$.
This figure shows the average ratio of IV and OLS estimates across all papers by each endogeneity category.

2. Reconciling the "Implausibly Large" IV Estimates

2.1 Can the "local" effect be even less representative of the population treatment effect?

Published research often gives the impression that IV estimates are consistent as long as conditions for identification (notably the exclusion restriction) are satisfied. However, consistency is defined relative to the goal of the research. IV estimates could produce an effect that is larger than the true population average treatment effect for legitimate econometric reasons due to the fact that they are uncovering a "local average treatment effect" (LATE). In fact, under plausible conditions regarding the heterogeneous treatment effect across a population, IV estimates could be farther away from the true population average treatment effect than their uninstrumented counterparts, even when they satisfy the exclusion restriction by the conventional standards.

That's because an exogenous shock often dramatically changes the probability of a treatment but falls short of assigning the treated status completely randomly. As a result, the IV estimation built on the exogenous shock estimates the effects of the treatment for those who respond to the manipulation as intended. Such agents are called "compliers" in econometric terms. Even though the shock could be perfectly exogenous, whether to respond to the shock remains a choice to some extent. If agents are rational and have some information about their expected treatment effect, then those who anticipate stronger effects are more likely to respond, on the margin, if the shock relaxes some constraints for participation. Therefore, the selection effect from isolating an effect based on the behavior of the compliers is not necessarily less than the voluntary selection to be treated among the total population in the absence of an exogenous shock.

This issue is well recognized in labor economics. For example, although economists generally believe that the correlation between education and earnings overstates the true causal impact of education, published studies routinely present estimates from elaborate identification schemes that are even more dramatic than the simple uninstrumented estimates (see a review and analysis by Card 2001).

Consider the following classical example, featured in Angrist and Pischke (2009), in which a researcher is analyzing the relationship between earnings and college attendance. The structural model is:

$$Earnings_i = \beta_0 + \beta_{1,i} College_i + \beta_2 \boldsymbol{Control}_i + \epsilon_i. \tag{1}$$

If the conditions $E(\epsilon_i | College_i, \boldsymbol{Control}_i) = 0$ and $E(\beta_{1,i} | College_i) = E(\beta_{1,i})$ are satisfied, the OLS regression yields a population average treatment effect, that is, plim $\hat{\beta}_{1,OLS} = E(\beta_{1,i})$. Most researchers would probably agree that, a priori, it is likely that $E(\epsilon_i \cdot College_i | \boldsymbol{Control}_i) > 0$, in other words, people with better earnings prospects based on attributes unobservable to econometricians are more likely to choose to receive more education. Hence, in expectation, $\hat{\beta}_{1,OLS} > \hat{\beta}_{1,IV}$ if the compliers are representative of the population.

Realizing that $College_i$ is potentially endogenous (that is, $E(College_i \epsilon_i) \neq 0$), the researcher introduces an instrument $Proximity_i$, defined as the distance between person i and the closest college campus. Suppose we accept that $Proximity_i$ is a valid instrument; that is, $Proximity_i$ affects the decision to attend the college in a significant way but does not affect $Earnings_i$, except indirectly through actual college attendance. For simplicity, let both $College_i$ (whether one goes to college or not) and $Proximity_i$ (whether there is a college campus within commuting distance or not) be discrete $\{0, 1\}$ variables. Then $\hat{\beta}_{1,IV}$ identifies the average treatment effect of a subpopulation of students who attend college if and only if there is a campus close by.

If a careful execution produces a result indicating that $\hat{\beta}_{1,IV} > \hat{\beta}_{1,OLS}$, then we learn the following: the subgroups whose decisions are affected by the supply-side shock (i.e., campus proximity) were constrained by the marginal cost of schooling, rather than by the lack of either desire or ability to benefit from education. Hence, the return to education for this subgroup could be substantially higher than the subpopulation that has no interest in attending college, even if there is a campus on the next block. The return to education for these compliers may also be higher than the return to the subpopulation of people who go to college regardless of distance because they can afford it. As such, the local treatment effect on the margin for the IV-compliers could exceed that of the population average treatment effect by more than the force of affirmative endogeneity embedded in the OLS estimate.

Such a lesson could be applied to many business decisions that are the subject of finance research because a large class of instruments for decisions made by investors or firms is supply-side innovations, or exogenous shifts in the cost of taking the action. The compliers in the IV analysis are precisely the agents who are the most sensitive to the cost of supply. If these subjects' inability to take the action in the absence of the supply shock was due to constraints (which could be financial or institutional), then the benefit they could enjoy from the treatment could be substantially higher than that of the unconstrained groups.

In the end, there is nothing inherently wrong with the surveyed papers, in terms of econometrics, if being the LATE of "special locations" of the population is the primary reason for the prevalence of IV estimates being dramatically larger than OLS ones. While it is interesting and informative to uncover the treatment effects of various localities in the population, it does beg the question as to whether we are collectively reaching a fair and unbiased inference about an underlying economic relation. If the localities in which the causal effects are identified in published research are concentrated in subsample locations where the LATE is significantly larger than the ATE, then our learning about the economic relation would never converge to what is prevailing at the center of the population even as more and more studies built upon one another attempting at the same relation.

2.2 Weaker IV, stronger results?

Angrist and Pischke (2009) show that weak (but still exogenous) instruments yield IV estimates that are biased toward their corresponding OLS estimates. If so, why do we observe so many IV estimates that are multitudes larger than their uninstrumented counterparts? Perhaps, except for a godsend, it is challenging to find a valid instrument that could explain a significant portion of the variation in an endogenous independent variable of key interest. That's because such a powerful explanatory variable would, most likely, already be in the system, which means that it would, most likely, directly affect an outcome in most cases, thus violating the exclusion restriction.

Let us look at the simplest IV system in which all variables are demeaned and normalized to be of unit variance:

$$\begin{aligned} x_i &= \gamma z_i + u_i \text{ (First stage)} \\ y_i &= \beta x_i + \epsilon_i = \beta x_i + (\iota z_i + \eta_i) \text{ (Structural equation)} \end{aligned} \tag{2}$$

In general, it is hard to conclude (or to be fully convinced) that any variable with explanatory potency is fully exogenous. A more acceptable argument for a good (but not godsend) instrument is that the direct effect of the z_i on y_i is of a secondary order compared to that of x_i—that is, ι/β is minuscule. Without loss of generality, we assume that all parameters are positive. Then

$$\beta_{IV} = \frac{cov(\beta x_i + \iota z_i + \eta_i, z_i)}{cov(x_i, z_i)} = \frac{\beta cov(x_i, z_i) + \iota var(z_i)}{cov(x_i, z_i)} = \beta + \frac{\iota}{\gamma}. \quad (3)$$

Therefore, the bias is $\frac{\iota}{\gamma}$. A valid but weak instrument is one such that γ is not far – though strictly bounded away – from zero in magnitude. In any finite sample, the estimate $\hat{\gamma}$ would not be zero, but $cov(x_i, z_i)$ will be small (relative to $var(x_i)$ and $var(z_i)$). Hence an instrument that only weakly covaries with the endogenous independent variable will amplify a very small (and unknown) violation of the exclusion restriction.[4] In other words, the weaker the explanatory power of an instrument, that is, the smaller the γ, the more demanding the estimation system is of the "purity" of the IV—that is, the smaller the ι must be to keep the bias small enough to be second order.

2.3 Bias from specification search

The weak instrument problem does not, on its own, introduce directional bias. However, it does when combined with and exacerbated by researchers' search for specifications that produce the most striking and significant results. In realistic situations, a requirement for the IV estimate to be significant—an implicit precondition for publication of original and positive research as opposed to critiques and placebo tests—implies that β_{IV} must be many times higher than β_{OLS} in order to be viable.

Here is a simple example. Start with a univariate regression model as in equation (2). Let us assume that the t-statistic of β_{OLS} is 4.0 (a respectable number by today's publication records), and the effective number of independent observations is 1,000 (a large sample for which asymptotic theorems safely apply). Further assume that the instrument z explains 2% of the variation in the endogenous variable x (another respectable number if considered as the incremental explanatory power of the instrument on top of other exogenous regressors). In this case, the asymptotic standard errors for the two estimates (under the null that both are consistent) satisfy:

[4] For more technical details, please see Bound, Jaeger, and Baker (1995).

$$\frac{\sigma(\hat{\beta}_{OLS})}{\sqrt{n}} \stackrel{n\to\infty}{\longrightarrow} \frac{\sigma_\epsilon}{\sigma_x};$$

$$\frac{\sigma\left(\hat{\beta}_{IV}\right)}{\sqrt{n}} \stackrel{n\to\infty}{\longrightarrow} \frac{\sigma_\epsilon}{\rho_{x,z}\sigma_x}.$$

The comparison shows that the instrumentation would "blow up" the standard error by $\frac{1}{\rho_{x,z}} = \frac{1}{\sqrt{R^2}} = 7.1$ when the R^2 is 2%. Correspondingly, the β_{IV} needs to be "blown up" to at least 4.4 times as large as β_{OLS} to keep the t-statistic of β_{IV} above 2.5 (the typical magnitude of t-statistics for β_{IV} in published research)! On the other hand, with 1,000 effective independent observations, even a 2% first-stage R^2 translates into an F-statistic of nearly 20, which would safely pass a "weak instrument" test under the current industry standard (as summarized in Stock and Yogo 2005)[5].

It is an encouraging trend that authors of published studies increasingly include weak-instrument tests defending their choice of IVs. Nevertheless, we should still be aware that the test does not inform us about the relative "score" between the relevance and exogeneity of the IV, that is, the ratio of $\frac{\iota}{\gamma}$. As the sample size increases, even a given low explanatory power will produce higher and higher t- and F-statistics; on the other hand, the ratio $\frac{\iota}{\gamma}$, if reflecting an underlying economic relation, does not diminish in larger samples.

This could explain why, conditional on publication, we observe β_{IV} to be multiple times as large as β_{OLS} even though the economic insights would have predicted β_{IV} to be lower than β_{OLS} in most cases. Such a bias belongs to a general class of "publication bias," in that the published literature is systematically unrepresentative of the population of attempted studies. The issue has been long recognized in medical and clinical research, and has more recently been given attention in social science (Franco, Malhotra, and Simonovits 2014). Because of the de facto requirement of statistical significance for publication, and given the econometric structure, only the specifications (of which the critical step is the choice of a proper instrument) that yield β_{IV} estimates that are multiples of β_{OLS} can survive the blown-up standard error. As such, it is plausible that the effects they identify do not represent the true, global, or local effects that could be uncovered with a neutral mind.

5 Under the following parameters that are commonly applied in empirical finance research: one endogenous regressor, one excluded instrument, nine other exogenous regressors, maximal bias of β_{IV} relative to β_{OLS} of 10%, and a size of the test 5%, the required first-stage F-statistic to pass the Stock and Yogo (2005) test is 11.49. This is the basis on which researchers often simply resort to requiring $F = 10$ as a rule of thumb.

3. What Should Be Done?

We would like to reiterate that this meta-analysis is not meant to be a criticism of any individual research. Each study may well be the authors' best and most honest search for the truth. The goal is to alert the profession that a method that was introduced as a cure for endogeneity bias might not, in some cases, have brought us closer to the true population relation, as suggested by the direction of the pre-cure bias. We also do not attempt to come up with a better (and feasible) cure for endogeneity. Instead, we propose the following practice (in addition to what researchers have increasingly been doing already, such as discussing the difference between ATE and LATE) to enhance the transparency about the effectiveness of the IV method as a cure in each individual case.

3.1 Anticipate the relative magnitude of β_{IV} and β_{OLS} ex ante, and reconcile ex post

After authors obtain both β_{IV} and β_{OLS}, there should be an explicit discussion on what their relative magnitude says about the nature of endogeneity or, more specifically, the sign of the correlation between the potentially endogenous regressor and the error term. Once the authors spell out the mechanisms of endogeneity, they can take advantage of the cross-sectional variation in the strength of such endogeneity across different subsamples based on economic reasoning. The sequence of OLS estimates over the different subsamples could offer some cues or provide validation for the assessed direction of OLS bias. Such logic is behind the "identification at infinity" approach (Chamberlain 1986). In other words, the selection effect wanes when the probability of treatment approaches either zero or one.

For example, if $\beta_{IV} \approx \beta_{OLS}$, does it imply that endogeneity, in the end, is not an issue or that there are opposing forces canceling each other out? If it is the latter, there should be an elaboration of what the competing forces are or, even better, an illustration with analyses on a subsample for which one force is expected to be dominant relative to the others. If $|\beta_{IV}| >> |\beta_{OLS}|$, as is the case in the majority of the published research, then the authors are obligated to explain why the force of endogeneity works against finding the desired result, despite the fact that in most cases it was the concern for affirmative endogeneity that led the authors to search for an instrument in the first place. Authors should also be allowed to modify their view about the directional nature of endogeneity if β_{IV} lies on the "wrong side" relative to β_{OLS}, based on economic reasoning; for example, when confounding factors affect the correlation between the key independent variable and the residual from opposing directions. Such a discussion will no doubt enrich a study.

3.2 Transparency regarding IV potency

We encourage the authors to calibrate the potential amplification effect. In published studies, authors almost never report the partial R^2 of the excluded instrumental variables in explaining the variation in the endogenous variable. Often, a seemingly high first-stage R^2 includes a long list of control variables, especially fixed effects with high dimensions, which masks the weak instrument problem—even with the gatekeeper of the weak-instrument tests.

If the partial R^2 is minuscule (which does not necessarily prevent highly significant t-statistics or respectable F-statistics in large samples), the readers need to be aware of two issues. First, the identification, even when consistent in its own right, is accomplished using a potentially very thin slice of "compliers." To what extent this thin slice represents the population of interest warrants its own discussion. Second, the readers should be aware that, as illustrated in Equation (3), the IV estimate is potentially a combination of a true treatment effect plus a blow-up of a second-order direct effect of the instrument on the outcome. The smaller the IV's incremental explanatory power, the greater the burden is on the authors to refute, or at least to qualify, any possible second-order direct effect of the instrument on the outcome, especially if both variables are from the same closed ecosystem.

3.3 Reality check

Compared with many other disciplines in social sciences, empirical economics and finance research takes pride in its greater reliance on rigorous data analysis in addition to presenting views with reasoning. However, a reality check should follow every analysis, no matter how rigorous it is. In general, OLS estimates mostly do not fall out of the bounds of reality because they represent an approximate linear relation among variables from the center of the data. The same cannot be said about the IV estimate. When $|\beta_{IV}| >> |\beta_{OLS}|$, the magnitude of the presumably consistent estimate often exceeds what one believes is plausible, where plausibility could build on basic intuition or reasonable calibration. Needless to say, common sense often fails to survive scientific scrutiny. Still, even in those cases a discussion of why common sense turns out to be wrong is worthwhile and educational.

In many cases, the bounds for a reality check could be more reliably substantiated by calibrating the cost for agents to undo or substitute out the effect of a treatment in the absence of major frictions. For example, if a dual-class share structure is found to causally destroy firm value by 5%, it could be perceived as plausible that empire-minded CEOs and their close insiders treat the value destruction as a cost of staying in power. However, if the finding is 50%, then one should at least raise the question as why the insiders in power are willing to make such a value sacrifice, or why outsider shareholders are not motivated to launch activism aiming at declassifying the dual class given the 100% potential return. If the discussion successfully and convincingly defends the large effect, then we learn something new (and shocking) about both the implied value of control and the associated friction in the financial markets that explains the deviation from rational value maximization. Then, armed with a careful analysis that tells us something true about the world, we can confidently recommend changes.

References

Angrist, J. D., and J.S. Pischke. 2009. *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.

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

Bound, J. and A. B. Krueger. 1991. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics* 9:1–24.

Card, D. 2001. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69:1127–60.

Chamberlain, G. 1986. Asymptotic efficiency in semiparametric model with censoring, *Journal of Econometrics* 32:189–218.

Franco, A., N. Malhotra, and G. Sinonovits. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science* 19:1502–05.

Stock, J. H., and M. Yogo. 2005. Test for weak instruments in linear IV regression. In *Identification and inference for econometric models: Essays in honor of Thomas J. Rothenberg*, ed. Donald W. Andrews and James H. Stock. Cambridge: Cambridge University Press.