JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12032

WHAT'S TO KNOW ABOUT THE CREDIBILITY OF EMPIRICAL ECONOMICS?

John Ioannidis

Stanford University

Chris Doucouliagos

Deakin University

Abstract. The scientific credibility of economics is itself a scientific question that can be addressed with both theoretical speculations and empirical data. In this review, we examine the major parameters that are expected to affect the credibility of empirical economics: sample size, magnitude of pursued effects, number and pre-selection of tested relationships, flexibility and lack of standardization in designs, definitions, outcomes and analyses, financial and other interests and prejudices, and the multiplicity and fragmentation of efforts. We summarize and discuss the empirical evidence on the lack of a robust reproducibility culture in economics and business research, the prevalence of potential publication and other selective reporting biases, and other failures and biases in the market of scientific information. Overall, the credibility of the economics literature is likely to be modest or even low.

Keywords. Bias; Credibility; Economics; Meta-research; Replication; Reproducibility

1. Introduction

Research is a public good produced by unavoidably self-interested researchers. The credibility of research and the market for evidence have been critically questioned (see Ioannidis, 2005, 2012a; Young *et al.,* 2008). Is evidence distorted? Is wrong evidence produced at a faster rate than correct evidence? Is research produced and replicated efficiently enough? Can we do better? These and other questions apply as much to economics as they do to other sciences. Over the years, investigators have identified several problems that affect the credibility of empirical economics research. These include but are not limited to: publication bias (DeLong and Lang, 1992); overstating the level of statistical significance (Caudill and Holcombe, 1999); mistaking statistical significance for economic significance (Ziliak and McClosky, 2004); growing positive-outcome bias (Fanelli, 2012); fraud and dishonesty (Bailey *et al.,* 2001; List *et al.,* 2001); funding and promotion inefficiencies (Oswald, 2007); unsupported claims and false beliefs (Levy and Peart, 2012); editors reluctant to publicize plagiarism (Enders and Hoover, 2004) and potentially distorting refereeing process (Frey, 2003).

In this paper, we discuss first some key parameters that may influence the credibility of research findings and how these parameters seem to perform in the science of economics (Section 2). Then we catalogue some of the evidence about unproductive research cultures and inefficient processes and potential biases in empirical economics (Section 3). Finally, we conclude with discussion of some unanswered questions and possible interventions that could be adopted to improve the market for research and eventually raise the credibility of economics (Section 4).

2. Key Parameters Affecting Research Credibility

In a previous paper one of us has argued that 'a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser pre-selection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance' (Ioannidis, 2005, e124, p. 1). We recognize that economics research can be very diverse and generalized descriptions cannot fit all. However, we focus on some common patterns of research in economics.

2.1 *Sample Size*

Many empirical economics studies are of relatively modest sample size, although sample sizes are increasing in many subfields. Depending on the unit of observation, sample sizes are sometimes unavoidably limited (e.g. macroeconomic studies with ecological analyses at the country-level). In other subfields where analyses are performed at the level of individuals, households or stocks, sample sizes can be large or even very large.

2.2 *Magnitude of Effect Size*

Most of the focus in empirical economics has been on statistical significance rather than the practical (economic) significance (Ziliak and McCloskey, 2004). Some effects are large, and many are small. Rather upsettingly, meta-analyses are finding that effect sizes in economics appear to be declining over time. For some economics phenomenon, effects may be tiny or even barely distinguishable from the null hypothesis. For example, some mainstream theories imply that predictive effects for stock market behaviour and stock values may be very small or non-existent.

2.3 *Number and Pre-selection of Tested Relationships*

Databases used in economics research are increasingly data-rich and contain more and more variables. Massive household surveys (e.g. the PSID) and financial databases (e.g. Compustat) contain data on tens of thousands of individuals and firms with hundreds of variables. Much empirical research is exploratory, hypothesis-generating, and this may not be visibly acknowledged in the published papers. Economics does a poor job accounting for the multiplicity of testing, and this is likely to exaggerate the importance of much of empirical economics (Leamer, 1983).

2.4 *Flexibility in Designs, Definitions, Outcomes and Analytical Modes*

Economics does not have a strong tradition in experimental, randomized designs, similar to the randomized controlled trials in medicine. Randomized experimentation is often considered time-consuming, costly, or, in many cases, even impossible. Thus, most empirical studies are observational and association research, which, by default, is likely to have low to very low credibility. Definitions of variables of interest are standardized in some fields. For instance, there is wide consensus among economists on how key variables such as economic activity or inflation rate are defined. However, for many other areas, there can be much flexibility in how key economic concepts should be defined or measured. There can also be substantial flexibility in the analyses performed. Much economics research depends on regression modelling that has a notorious flexibility in model construction (Leamer, 1983).

2.5 *Financial and Other Interest and Prejudice*

It is not clear whether financial conflicts of interest are overall less problematic in economics than in the health sciences (Ioannidis, 2011). Non-disclosure and inconsistent disclosure of conflicts seems to be common among prominent academic economists (Carrick-Hagenbarth and Epstein, 2012). In some economics fields where empirical models are developed for proprietary application, model-builders definitely do not want to distort the performance of the models, because having the models work (that is, actually make money) is the real hallmark of success. This type of economics research may have something important to teach other sciences. Imagine if medical researchers had to pay money each time their research was not replicated or did not work. It is likely however that academic self-interest (publishing for promotion, funding or professional advancement) is as common in economics and some empirical research may be performed with the explicit 'mission' of proving theories correct, no matter what (i.e. 'confirmation' and 'allegiance' biases). Evidence on specific biases that may distort economics research is more specifically discussed in the next section. See Section 3, below.

2.6 *Multiplicity of Teams*

There are probably far fewer researchers working in economics than in the life or physical sciences. For example, as of 12/26/2012, Microsoft Academic Search lists 512,895 author names under Economics & Business, compared with 6,010,966 in Medicine and 1,847,184 in Physics. However, the paradigm of large-scale collaboration of multiple teams under common projects with emphasis on data and resources sharing and replication is not common in economics research, while conversely it is very popular in the physical sciences and several biomedical fields, e.g. genomics. Although the number of co-authors has been increasing over time, most economics papers are written by a single or a very few authors. Several fields have many teams working on the same or similar questions, but most of the time there is no overarching collaboration. Even retrospective meta-analysis is far less common in economics than in life sciences. Many economic sub-fields and specific areas of research have very few researchers actively involved, perhaps leading to some information monopolies or inbreeding (Ioannidis, 2012b). However, while collaboration is important to the advancement of science, so is competition. Doucouliagos and Stanley (2013) show that competition between rival economics researchers actually increases the credibility of research by reducing publication bias; the greater are theoretical contests the less distorted is economics research and empirical economic inference.

Based on the profile sketched above, it is likely that the credibility of economics research is not very high, but substantial variability is likely to exist across diverse sub-disciplines. Biases and deficiencies in the scientific process of self-correction may compound this picture further, as we discuss below.

3. Unproductive Cultures, Inefficient Processes and Biases in Economics

3.1 *Reproducibility Culture*

Reproducibility is essential for validating empirical research. Reproducibility has several levels, ranging from the ability of repeating and conforming analyses on existing data ('repeatability') to replication performed with new data. Replication may be performed using the same design, methods and question as the original study. Or, it may represent conceptual replication, where methods and research questions deviate from the original. On the one hand, conceptual replication may offer corroboration for a research claim, but, on the other, the new methods and research questions may also need to be subjected to strict replication themselves.

There is relatively very little replication performed in empirical economics, business, and marketing research (Hubbard and Vetter, 1992; Evaschitzky *et al*., 2007; Hubbard and Armstrong, 1994; Evanschitzky and Armstrong, 2010).[1] Most replication efforts are conceptual rather than strict replications. Moreover, there is a high rate of replications that fail to support original findings, ranging from 20% to 65%, depending on the field and journal (See Hubbard and Vetter, 1992; Evanschitzky and Armstong, 2010). These may actually be underestimates since some replications are performed either by the same scientists who proposed the original research findings, or by affiliates and scientists who share the same beliefs and thus may be under the influence of allegiance bias.

Repeatability is also known to be problematic in economics. In their famous study, Dewald *et al.* (1986) found that errors in empirical papers were common, although without necessarily invalidating the main conclusions of the studies. Most errors are inadvertent or due to suboptimal research practices and deficient quality control, but falsification may also be occasionally involved (Fanelli 2009). Tödter (2009) tested Benford's Law on data from two economics journals and found violations in about one-quarter of the papers, consistent with falsification. Bailey *et al*. (2001) conduct a survey of the most prolific researchers in accounting and report that 4% of the respondents confessed to research falsification. List *et al*. (2001) find a similar rate of falsification among economists and they also find that 7–10% of economists surveyed confessed to taking credit for graduate students' work or giving unjustified co-authorship. According to Fanelli (2009) up to 72% of scientists are thought to adopt questionable research practices (not necessarily unconditional falsification). John *et al*. (2012) find very high rates of questionable research practices among psychologists. The rates of fabrication and falsification may nevertheless be lower in economics and social sciences than in medical and pharmaceutical related research.

Since Dewald *et al.* (1986) it has become much more common for authors to make their data available (Hamermesh, 2007). However, corrections of errors and mistakes remain uncommon. Even when errors are recognized, they are not prominently shared in public view. Independent replication or error correction has not increased significantly, and they are not frequently found in economic journals.[2] Replication is a public good and hence prone to market failure.[3]

3.2 *Are the Bad Days Over? Does Experimental Design Increase Credibility?*

Angrist and Pischke (2010) are optimistic about current empirical economics. They argue that the shift towards randomized trials and quasi-experimental studies has transformed econometrics; all researchers have to do is focus on better research design and credibility is increased. Their main focus is on Leamer's critique (1983). However, the issues noted by Leamer are only part of the problem. For example, even though experimental designs have inherently better protection from many confounding biases than observational data, we do not know how much of the experimental economics literature is contaminated by publication or other selective reporting biases. Roth (1994) argues that experimental economics is vulnerable to data selection and confirmation bias.

Small-study effects suggestive of possible selective reporting biases remain common in randomized trials in other disciplines (Dwan *et al*., 2008).[4] Similarly, we do not know how vulnerable these studies are to the winner's curse, i.e. inflated effect sizes when effects are discovered by relatively small studies (Ioannidis, 2008). Maniadis *et al.* (2012) argue that experimental economics is plagued by inflated initial results and false positives. Most randomized and quasi-experimental studies in econometrics tend to have modest sample sizes. For example, among all studies published in the four issues of *Experimental Economics* in 2012, sample sizes range from 67 to 1175 subjects with a median of 184. So publication bias and winner's curse may be an issue in such underpowered settings. Even the largest observed effects may be spurious, as has been shown in the medical sciences (Pereira, Horwitz, and Ioannidis, 2012). Finally, experimental randomized studies still represent a small minority. Hamermesh (2012) finds that in 2011, 8.2% of the studies in three leading economics journals were experimental studies, compared to 0.8% in

1983.[5] Quasi-experimental studies and those using instrumental variables may not be as protected from biases as experimental randomized studies.

3.3 *Other Empirical Studies of Diverse Biases in Economics*

The economics literature seems to have too many results that confirm the authors' expectations (Fanelli, 2010; 2012). Selective reporting biases cumulatively create literature where there are just too many nominally significant research findings, a situation that can be probed with an excess significance test (Ioannidis and Trikalinos, 2007). The proportion of studies that reported support for the tested hypothesis in economics was found to be 88%, one of the highest across all sciences (Fanelli, 2010). After controlling for differences between pure and applied disciplines and between papers testing one or several hypotheses, the odds of reporting in favour of the tested hypothesis were five times higher among papers in Economics and Business compared to Space Science (Fanelli, 2010).

Frey (2003)[6] has charged that publishing in the economic sciences is equivalent to a form of 'prostitution'. This highly charged word is used to denote the way that investigators have to compromise and submit to the wishes of reviewers and editors to get their work published. The conformity effects of peer-review may apply also at the proposal funding level rather than just the peer-review of journal articles (Nicholson and Ioannidis, 2012). There are few experimental studies of the impact of different forms of peer-review in the economic sciences, but there is an emerging literature in other sciences.[7] On-line publishing and the recasting of the review process with emphasis on post-publication review and open, crowd-sourced review need to be more fully explored and developed.

The interpretation and adoption of research theories and results can also be problematic. Rogeberg and Melberg (2011) have evaluated the rational addiction literature and they conclude that 'absurd and unjustified claims can be made and accepted in even highly-ranked journals' (p. 29). There is also some evidence that investigators and interested policy users may cherry-pick results and promote or publicize only those that fit into their agendas. This type of dissemination bias has been documented, for example, in the field of alcohol marketing and adolescent drinking (Nelson, 2011).

4. Conclusions: Unanswered Questions and Possibilities for Improvement

Despite the aforementioned empirical evidence, much remains unknown. How credible is economics research in different subfields? Is credibility rising or falling over time? How can this market be improved? What is the impact of corporate and institutional interests, journals, funding agencies, and other stakeholders and how can we utilize this potential impact to improve the credibility of economics?

However some paths forward seem clear: strengthen the reproducibility culture with emphasis on independent replication; conduct larger, better studies; promote collaborative efforts rather than siloed, one-investigator research; and reduce biases and conflicts. The exact interventions, which might achieve these changes are not clear, and perhaps there is room for conducting experimental studies on different potential interventions. Progress may be difficult to achieve unless the rewards and incentives system of conducting and publishing research is modified. Little progress is likely if investigators get rewarded and promoted for publishing significant results and for perpetuating theories and claims even when they are wrong. Conversely, one might expect better outcomes, if replication research is encouraged, reproducibility is rewarded, and/or irreproducibility is penalized.

Answering these questions requires more meta-research (i.e. empirical research on research). Empirical studies that have been performed in other scientific disciplines may be extrapolated and conducted in economics. Tests for bias may also be adapted for economics and embraced by leading journals. Of course, meta-research is also susceptible to bias and errors of its own. The extent of bias and possible errors needs to be evaluated also for emerging meta-research tools. Nonetheless, tests for small-study

effects and publication or selective reporting biases need to be applied increasingly, further developed and evaluated (Stanley and Doucouliagos, 2012).

Notes

1. One may argue that it is not optimal to replicate all empirical studies. Moreover, it is impossible to tell how many replications are attempted, only how many are reported. But, only the reported ones matter in shaping the literature and future efforts by the scientific community.
2. There has been a revolution in software and data archives (Koenker and Zeileis, 2009). Despite this, Hamermesh (2007) finds that few scientists actually contact publishing authors for their data even in cases where they state that their data is available. Making data available is not sufficient to induce replication.
3. As Dewald *et al.* note (1986, p. 589): 'A single researcher faces high costs in time and money from undertaking replication of a study and finds no ready marketplace which correctly prices the social and individual value of the good'.
4. In their review of different experimental practices, Hertwig and Ortmann (2001) conclude that economics has adopted a stricter methodology than psychology, which serves as a regulatory framework that makes economics results relatively more credible.
5. The largest change has been in papers that use self-assembled data, which has risen from 2.4% in 1983 to 34% in 2011 of all papers published in these journals. At the same time, pure theory papers have fallen from 57.6% to 19.1% of the total.
6. In an interesting turn of events, Autor (2011) reveals that Frey had published essentially the same paper several times.
7. An exception is Blank (1991) who compares single- to double-blind reviewing and finds more critical referee reports and lower acceptance rates with double-blind reviewing.

References

Angrist, J.D. and Pischke, J.-S. (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24(2): 3–30.

Autor, D.H. (2011) Letter to Professor Bruno Frey. *Journal of Economic Perspectives* 25(3): 239–240.

Bailey, C.D., Hasselback, J.R. and Karcher, J.N. (2001) Research misconduct in accounting literature: a survey of the most prolific researchers' actions and beliefs. *Abacus* 37(1): 26–54.

Blank, R.M. (1991) The effects of double-blind versus single-blind reviewing: experimental evidence from the American Economic Review. *American Economic Review* 81(5): 1041–1067.

Carrick-Hagenbarth, J. and Epstein, G.A. (2012) Dangerous interconnectedness: economists' conflicts of interest, ideology and financial crisis. *Cambridge Journal of Economics* 36: 43–63.

Caudill, S.B. and Holcombe, R.G. (1999) Specification search and levels of significance in econometric models. *Eastern Economic Journal* 25: 289–300.

DeLong, J. B. and Lang, K. (1992) Are all economic hypotheses false? *Journal of Political Economy* 100: 1257–1272.

Doucouliagos, H. and Stanley, T.D. (2013) Theory competition and selectivity: are all economic facts greatly exaggerated? *Journal of Economic Surveys* 27(2): 316–339.

Dewald, W.G., Thursby, J.G. and Anderson, R.G. (1986) Replication in empirical economics. *The Journal of Money, Credit and Banking Project. American Economic Review* 76: 587–603.

Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J.P., Simes, J. and Williamson, P.R. (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 3: e3081.

Enders, W. and Hoover, G.A. (2004) Whose line is it? Plagiarism in economics. *Journal of Economic Literature* 42(2): 487–493.

Evanschitzky, H., Baumgarth, C., Hubbard, R. and Armstrong, J.S. (2007) Replication research's disturbing trend. *Journal of Business Research* 60: 411–415.
Evanschitzky, H. and Armstrong, J.S. (2010) Replications of forecasting research. *International Journal of Forecasting* 26: 4–8.
Fanelli, D. (2009) How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* 4(5): e5738.
Fanelli, D. (2010) 'Positive' results increase down the hierarchy of the sciences. *PLoS ONE* 5: e10068.
Fanelli, D. (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics* 90: 891–904.
Frey, B.S. (2003) Publishing as prostitution? Choosing between one's own ideas and academic success. *Public Choice* 116: 205–223.
Hamermesh, D. (2007) *Replication in Economics*. IZA Discussion Paper No. 2760.
Hamermesh, D. (2012) Six decades of top economics publishing: who and how? National Bureau of Economic Research Working Paper Number 18635.
Hertwig, R. and Ortmann, A. (2001) Experimental practices in economics: a methodological challenge for psychologists? *Behavioral and Brain Sciences* 24: 383–451.
Hubbard, R. and Armstrong, J.S. (1994) Replication and extensions in marketing: rarely published but quite contrary. *International Journal of Research in Marketing* 11: 233–248.
Hubbard, R. and Vetter, D.E. (1992) The publication incidence of replications and critical commentary in economics. *The American Economist* 36(1): 29–34.
Ioannidis, J.P.A. (2005) Why most published research findings are false. *PLoS Medicine* 2: e124.
Ioannidis, J.P.A. (2008) Why most true discovered associations are inflated. *Epidemiology* 19: 640–648.
Ioannidis, J.P.A. (2011) An epidemic of false claims. Competition and conflicts of interest distort too many medical findings. *Scientific American* 304: 16.
Ioannidis, J.P.A. (2012a) Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7: 645–654.
Ioannidis, J.P.A. (2012b) Scientific inbreeding and same-team replication: type D personality as an example. *Journal of Psychosomatic Research* 73: 408–410.
Ioannidis, J.P.A. and Trikalinos, T.A. (2007) An exploratory test for an excess of significant findings. *Clinical Trials* 4: 245–253.
John, L.K., Loewenstein, G. and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23: 524–532.
Koenker, R. and Zeileis, A. (2009) On reproducible econometric research. *Journal of Applied Econometrics* 24: 833–847.
Leamer, E.E. (1983) Let's take the con out of econometrics. *American Economic Review* 73(1): 31–43.
Levy, D.M. and Peart, S.J. (2012) Tullock on motivated inquiry: expert-induced uncertainty disguised as risk. *Public Choice* 152: 163–180.
List, J.A., Bailey, C.D., Euzent, P.J. and Martin, T.L. (2001) Academic economists behaving badly? A survey of three areas of unethical behaviour. *Economic Inquiry* 39(1): 162–170.
Maniadis, Z., Tufano, F. and List, J.A. (2012) One swallow doesn't make a summer: how economists (mis-) use experimental methods and their results. Available at: http://www.aueb.gr/conferences/Crete2012/papers/papers%20more%20recent/Maniadis.pdf (last accessed 25 March 2013).
Nelson, J.P. (2011) Alcohol marketing, adolescent drinking and publication bias in longitudinal studies: a critical survey using meta-analysis. *Journal of Economic Surveys* 25(2): 191–232.
Nicholson, J.M. and Ioannidis, J.P. (2012) Research grants: conform and get funded. *Nature* 492: 34–36.
Oswald, A.J. (2007) An examination of the reliability of prestigious scholarly journals: evidence and implications for decision-makers. *Economica* 74: 21–31.
Pereira, T.V., Horwitz, R.I. and Ioannidis, J.P. (2012) Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 308: 1676–1684.
Roth, A.E. (1994) Lets keep the con out of experimental econ.: a methodological note. *Empirical Economics* 19: 279–289.
Stanley, T.D. and Doucouliagos, H. (2012) *Meta-Regression Analysis in Economics and Business*. Oxford: Routledge.

Tödter, K.-H. (2009) Benford's Law as an indicator of fraud in economics. *German Economic Review* 10(3): 339–351.

Young, N.S., Ioannidis, J.P. and Al-Ubaydli, O. (2008) Why current publication practices distort science. *PLoS Medicine* 5: e201.

Ziliak, S.T. and McCloskey, D.N. (2004) Size matters: the standard error of regressions in the American Economic Review. *Journal of Socio-Economics* 33: 527–546.