

Transparency, Reproducibility, and the Credibility of Economics Research

Garret Christensen
University of California, Berkeley

Edward Miguel¹
University of California, Berkeley and NBER

May 2017

Abstract: There is growing interest in enhancing research transparency and reproducibility in economics and other scientific fields. We survey existing work on these topics within economics, and discuss the evidence suggesting that publication bias, inability to replicate, and specification searching remain widespread in the discipline. We next discuss recent progress in this area, including through improved research design, study registration and pre-analysis plans, disclosure standards, and open sharing of data and materials, drawing on experiences in both economics and other social sciences. We discuss areas where consensus is emerging on new practices, as well as approaches that remain controversial, and speculate about the most effective ways to make economics research more credible in the future.

¹ We thank the editor Steven Durlauf and four anonymous referees for useful comments.

Contact Information: Department of Economics, Evans Hall #3880, University of California, Berkeley, CA 94720-3880, USA. Email: emiguel@berkeley.edu.

TABLE OF CONTENTS

1	INTRODUCTION.....	3
2	EVIDENCE ON PROBLEMS WITH THE CURRENT BODY OF RESEARCH.....	4
2.1	A Model for Understanding the Issues	4
2.2	Publication Bias	7
2.2.1	Publication bias in several empirical economics literatures	12
2.2.2	Publication Bias and Effect Size	14
2.3	Specification Searching	15
2.3.1	Sub-Group Analysis.....	17
2.4	Inability to Replicate Results	18
2.4.1	Data Availability.....	18
2.4.2	Types of Replication Failures and Examples.....	24
2.4.3	Fraud and Retractions	30
3	NEW RESEARCH METHODS AND TOOLS.....	33
3.1	Improved analytical methods: research designs and meta-analysis	33
3.1.1	Understanding Statistical Model Uncertainty	34
3.1.2	Improved Publication Bias Tests.....	36
3.1.3	Multiple Testing Corrections.....	37
3.2	Study Registration	41
3.3	Pre-Analysis Plans	42
3.3.1	Examples of Pre-analysis Plans (PAPs).....	43
3.3.2	Strengths, Limitations, and Other Issues Regarding Pre-Analysis Plans	46
3.3.3	Observational Studies.....	47
3.4	Disclosure and reporting standards	51
3.4.1	Fraud and Retractions	55
3.5	Open data and materials, and their use for replication	56
3.5.1	Computational Issues	59
3.5.2	The Limits of Open Data	60
4	FUTURE DIRECTIONS AND CONCLUSION	61

1 Introduction

Openness and transparency have long been considered key pillars of the scientific ethos (Merton 1973). Yet there is growing awareness that current research practices often deviate from this ideal, and can sometimes produce misleading bodies of evidence (Miguel et al. 2014). As we survey in this article, there is growing evidence documenting the prevalence of publication bias in economics and other scientific fields, as well as specification searching, and widespread inability to replicate empirical findings. Though peer review and robustness checks aim to reduce these problems, they appear unable to solve the problem entirely. While some of these issues have been widely discussed within economics for some time (for instance, see Leamer 1983; Dewald, Thursby, and Anderson 1986; DeLong and Lang 1992), there has been a notable recent flurry of activity documenting these problems, and also generating new ideas for how to address them.

The goal of this piece is to survey this emerging literature on research transparency and reproducibility, and synthesize the insights emerging in economics as well as from other fields – awareness of these issues has also recently come to the fore in political science (Gerber, Green, and Nickerson 2001; Franco, Malhotra, and Simonovits 2014), psychology (Simmons, Nelson, and Simonsohn 2011; Open Science Collaboration 2015), sociology (Gerber and Malhotra 2008a), finance (Harvey, Liu, and Zhu 2015), and other research disciplines as well, including medicine (Ioannidis 2005). We also discuss productive avenues for future work.

The potential flexibility in analysis described in Leamer (1983) has increased with the vastly greater computing power of recent decades and the ability to run a nearly infinite number of regressions (as in Sala-i-Martin 1997), and there is renewed concern that null-hypothesis statistical testing is subject to both conscious and unconscious manipulation. At the same time, technological progress has also facilitated various new tools and potential solutions, including by streamlining the online sharing of data, statistical code, and other research materials, as well as the creation of easily accessible online study registries, data repositories, and tools for synthesizing research results across studies. Data-sharing and replication activities are certainly becoming more common within economics research. Yet, as we discuss below, the progress to date is partial, with some journals and fields within economics adopting new practices to promote transparency and reproducibility and many others not (yet) doing so.

The rest of the paper is organized as follows: Section 2 focuses on documenting the problems, first framing them with a simple model of the research and publication process (subsection 2.1), then discussing publication bias (subsection 2.2), specification searching (subsection 2.3), and the inability to replicate results (subsection 2.4). Section 3 focuses on possible solutions to these issues: improved analytical methods (subsection 3.1), study registration (subsection 3.2) and pre-analysis plans (subsection 3.3), disclosure and reporting standards (subsection 3.4), and open data and materials (subsection 3.5). Section 4 discusses future directions for research as well as possible approaches to change norms and practices, and concludes.

2 Evidence on problems with the current body of research

Multiple problems have been identified within the body of published research results in economics. We focus on three that have come under greater focus in the recent push for transparency: publication bias, specification searching, and an inability to replicate results. Before describing them, it is useful to frame some key issues with a simple model.

2.1 A Model for Understanding the Issues

A helpful model to frame some of the issues discussed below was developed in the provocatively titled “Why Most Published Research Findings Are False” by Ioannidis (2005), which is among the most highly cited medical research articles from recent years. Ioannidis develops a simple model that demonstrates how greater flexibility in data analysis may lead to an increased rate of false positives and thus incorrect inference.

Specifically, the model estimates the positive predictive value (PPV) of research, or the likelihood that a claimed empirical relationship is actually true, under various assumptions. A high PPV means that most claimed findings in a literature are reliable; a low PPV means the body of evidence is riddled with false positives. The model is similar to that of Wacholder et al. (2004), which estimates the closely related false positive report probability (FPRP).²

For simplicity, consider the case in which a relationship or hypothesis can be classified in a binary fashion as either a “true relationship” or “no relationship”. Define R_i as the ratio of true

² We should note that there is also a relatively small amount of theoretical economic research modeling the researcher and publication process, including Henry (2009), which predicts that, under certain conditions, more research effort is undertaken when not all research is observable, if such costs can be incurred to demonstrate investigator honesty. See also Henry and Ottaviani (2014) and Libgober (2015).

relationships to no relationships commonly tested in a research field i (e.g., development economics). Prior to a study being undertaken, the probability that a true relationship exists is thus $R_i/(R_i+1)$. Using the usual notation for statistical power of the test $(1-\beta)$ and statistical significance level (α) , the PPV in research field i is given by:

$$(eqn. 1) \quad PPV_i = (1-\beta)R_i / ((1-\beta)R_i + \alpha)$$

Clearly, the better powered the study, and the stricter the statistical significance level, the closer the PPV is to one, in which case false positives are largely eliminated. At the usual significance level of $\alpha = 0.05$ and in the case of a well-powered study ($1 - \beta = 0.80$) in a literature in which half of all hypotheses are thought to be true ex ante ($R_i/(R_i+1) = 0.5$), the PPV is relatively high at 94%, a level that would not seem likely to threaten the validity of research in a particular economics subfield.

However, reality is considerably messier than this best case scenario and, as Ioannidis describes, this could lead much high rates of false positives in practice due to the presence of underpowered studies, specification searching and researcher bias, and the possibility that only a subset of the analysis in a research literature is published. We discuss these extensions in turn.

We start with the issue of statistical power. Doucouliagos and Stanley (2013), Doucouliagos, Ioannidis, and Stanley (2016), and others have documented that many empirical economics studies are actually quite underpowered. With a more realistic level of statistical power for many studies, say at 0.50, but maintaining the other assumptions above, the PPV falls to 91%, which is beginning to potentially look like more of a concern. For power = 0.20, fully 20% of statistically significant findings are false positives.

This concern, and those discussed next, are all exacerbated by bias in the publication process. If all estimates in a literature were available to the scientific community, researchers could begin to undo the concerns over a low PPV by combining data across studies, effectively achieving greater statistical power and more reliable inference, for instance, using meta-analysis methods. However, as we discuss below, there is growing evidence of a pervasive bias in favor of significant results, in both economics and other fields. If only significant findings are ever seen by the researcher community, then the PPV is the relevant quantity for assessing how credible an individual result is likely to be.

Ioannidis extends the basic model to account for the possibility of what he calls researcher bias. Denoted by u , researcher bias is defined as the probability that a researcher presents a non-finding as a true finding, for reasons other than chance variation in the data. This researcher bias could take many forms, including any combination of specification searching, data manipulation, selective reporting, and even outright fraud; below we attempt to quantify the prevalence of these behaviors among researchers. There are many checks in place that attempt to limit this bias, and through the lens of empirical economics research, we might hope that the robustness checks typically demanded of scholars in seminar presentations and during journal peer review manage to keep the most extreme forms of bias in check. Yet we believe most economists would agree that there remains considerable wiggle room in the presentation of results in practice, in most cases due to behaviors that fall far short of outright fraud.

Extending the above framework to incorporate the researcher bias term (u_i) in field i leads to the following expression:

$$(eqn. 2) \quad PPV_i = ((1 - \beta)R_i + u_i\beta R_i) / ((1 - \beta)R_i + \alpha + u_i\beta R_i + u_i(1-\alpha))$$

Here the actual number of true relationships (the numerator) is almost unchanged, though there is an additional term that captures the true effects that are correctly reported as significant only due to author bias. The total number of reported significant effects could be much larger due to both sampling variation and author bias. If we go back to the case of 50% power, $R_i/(R_i+1) = 0.5$, and the usual 5% significance level, but now assume that author bias is low at 10%, the PPV falls from 91% to 79%. If 30% of authors are biased in their presentation of results, the PPV drops dramatically to 66%, meaning that over a third of reported significant effects are actually false positives.

In a further extension, Ioannidis examines the case where there are n_i different research teams in a field i generating estimates to test a research hypothesis. Once again, if only the statistically significant findings are published, so there is no ability to pool all estimates, then the likelihood that any published estimate is truly statistically significant can again fall dramatically.

In Table 1 (a reproduction of Table 4 from Ioannidis (2005)), we present a range of parameter values and the resulting PPV. Different research fields may have inherently different levels of the R_i term, where presumably literatures that are in an earlier stage and thus more exploratory presumably have lower likelihoods of true relationships.

This simple framework brings a number of the issues we deal with in this article into sharper relief, and contains a number of lessons. Ioannidis (2005) himself concludes that the majority of published findings in medicine are likely to be false, and while we are not prepared to make a similar claim for empirical economics research – in part because it is difficult to quantify some of the key parameters in the model – we do feel that this exercise does raise important concerns about the reliability of findings in many literatures.

First off, literatures characterized by statistically under-powered (i.e., small $1-\beta$) studies are likely to have many false positives. A study may be under-powered both because of small sample sizes, and if the underlying effect sizes are relatively small. A possible approach to address this concern is to employ larger datasets or estimators that are more powerful.

Second, the hotter a research field, with more teams (n_i) actively running tests and higher stakes around the findings, the more likely it is that findings are false positives. This is both due to the fact that multiple testing generates more false positives (in absolute numbers) and also because author bias (u_i) may be greater when the stakes are higher. Author bias is also a concern when there are widespread prejudices in a research field, for instance, against publishing findings that contradict core theoretical concepts or assumptions.

Third, the greater the flexibility in research design, definitions, outcome measures, and analytical approaches in a field, the less likely the research findings are to be true, again due to a combination of multiple testing concerns and author bias. One possible approach to address this concern is to mandate greater data sharing so that other scholars can assess the robustness of results to alternative models. Another is through approaches such as pre-analysis plans that effectively force scholars to present a certain core set of analytical specifications, regardless of the results.

With this framework in mind, we next present empirical evidence from economics and other social science fields regarding the extent of some of the problems and biases we have been discussing, and then in section 3 turn to potential ways to address them.

2.2 Publication Bias

Publication bias arises if certain types of statistical results are more likely to be published than other results, conditional on the research design and data used. This is usually thought to be most relevant in the case of studies that fail to reject the null hypothesis, which are thought to generate less support for publication among referees and journal editors. If the research community is

unable to track the complete body of statistical tests that have been run, including those that fail to reject the null (and thus are less likely to be published), then we cannot determine the true proportion of tests in a literature that reject the null. Thus it is critically important to understand how many tests have been run. The term “file drawer problem” was coined decades ago (Rosenthal 1979) to describe this problem of results that are missing from a body of research evidence. The issue was a concern even earlier, see, for example, Sterling (1959) which warned of “embarrassing and unanticipated results” from type I errors if not significant results went unpublished.

Important recent research by Franco, Malhotra, and Simonovits (2014) affirms the importance of this issue in practice in contemporary social science research. They document that a large share of empirical analyses in the social sciences are never published or even written up, and the likelihood that a finding is shared with the broader research community falls sharply for “null” findings, i.e., that are not statistically significant (Franco, Malhotra, and Simonovits 2014).

Cleverly, the authors are able to look inside the file drawer through their access to the universe of studies that passed rigorous peer review for inclusion in a nationally representative social science survey administered at no cost to the researchers, namely, the NSF-funded Time-sharing Experiments in the Social Sciences, or TESS³. The same survey research firm conducted nearly all of the studies, and all studies included power calculations as part of the application process. TESS funded experimental studies across research fields, including in economics, e.g., Walsh, Dolfen, and DiNardo (2009) and Allcott and Taubinsky (2015), as well as political science, sociology and other fields. Franco, Malhotra, and Simonovits successfully tracked nearly all of the original studies over time, keeping track the nature of the empirical results as well as the ultimate publication of the study, across the dozens of studies that participated in the original project.

They find a striking empirical pattern: studies where the main hypothesis test yielded null results are 40 percentage points less likely to be published in a journal than a strongly statistically significant result, and a full 60 percentage points less likely to be written up in any form. This finding has potentially severe implications for our understanding of findings in whole bodies of social science research, if “zeros” are never seen by other scholars, even in working

³ See <http://tessexperiments.org>.

paper form. It implies that the positive predictive value (PPV) of research is likely to be lower than it would be otherwise, and also has negative implications for the validity of meta-analyses, if null results are not known to the scholars attempting to draw broader conclusions about a body of evidence. Figure 1 reproduces some of the main patterns from Franco, Malhotra, and Simonovits (2014), as described in (Mervis 2014b).

Consistent with these findings, other recent analyses have documented how widespread publication bias appears to be in economics research. Brodeur et al. (2016) collected a large sample of test statistics from papers in three top journals that publish largely empirical results (the *American Economic Review*, *Quarterly Journal of Economics*, and *Journal of Political Economy*) from 2005-2011. They propose a method to differentiate between the journal's selection of papers with statistically stronger results and inflation of significance levels by the authors themselves. They begin by pointing out that a distribution of z-statistics under the null hypothesis would have a monotonically decreasing probability density. Next, if journals prefer results with stronger significance levels, this selection could explain an increasing density, at least on part of the distribution. However, Brodeur et al. hypothesize that observing a local minimum density before a local maximum is unlikely if only this selection process by journals is present. They argue that a local minimum is consistent with the additional presence of inflation of significance levels by the authors.

Brodeur et al. (2016) document a rather disturbing two-humped density function of test statistics, with a relative dearth of reported p-values just above the standard 0.05 level (i.e., below a t-statistic of 1.96) cutoff for statistical significance, and greater density just below 0.05 (i.e., above 1.96 for t-statistics). This is a strong indication that some combination of author bias and publication bias is fairly common. Using a variety of possible underlying distributions of test statistics, and estimating how selection would affect these distributions, they estimate the residual ("the valley and the echoing bump") and conclude that between 10 to 20% of marginally significant empirical results in these journals are likely to be unreliable. They also document that the proportion of misreporting appears to be lower in articles without "eye-catchers" (such as asterisks in tables that denote statistical significance), as well as in papers written by more senior authors, including those with tenured authors.

A similar pattern strongly suggestive of publication bias also appears in other social science fields including political science, sociology, psychology, as well as in clinical medical

research. Gerber and Malhotra (2008a) have used the caliper test, which compares the frequency of test statistics just above and below the key statistical significance cutoff, which is similar in spirit to a regression discontinuity design. Specifically, they compare the number of z-scores lying in the interval $[1.96 - X\%, 1.96]$ to the number in $(1.96, 1.96 + X\%]$, where X is the size of the caliper, and they examine these differences at 5%, 10%, 15%, and 20% critical values.⁴

These caliper tests are used to examine reported empirical results in leading sociology journals (the *American Sociological Review*, *American Journal of Sociology*, and *The Sociological Quarterly*) and reject the hypothesis of no publication bias at the 1 in 10 million level (Gerber and Malhotra 2008a). Data from two leading political science journals (the *American Political Science Review* and *American Journal of Political Science*) reject the hypothesis of no publication bias at the 1 in 32 billion level (Gerber and Malhotra 2008b).

Psychologists have recently developed a related tool called the “p-curve,” describing the density of reported p-values in a literature, that again takes advantage of the fact that if the null hypothesis were true (i.e., no effect), p-values should be uniformly distributed between 0 and 1 (Simonsohn, Nelson, and Simmons 2014a). Intuitively, under the null of no effect, a p-value < 0.08 should occur 8% of the time, a p-value < 0.07 occurs 7% of the time, etc., meaning a p-value between 0.07 and 0.08, or between any other 0.01-wide interval, should occur 1% of the time. In the case of true non-zero effects, the distribution of p-values should be right-skewed (with a decreasing density), with more low values (0.01) than higher values (0.04) (Hung et al. 1997).⁵ In contrast, in bodies of empirical literature suffering from publication bias, or “p-hacking” in their terminology, in which researchers evaluate significance as they collect data and only report results with statistically significant effects, the distribution of p-values would be left-skewed (assuming that researchers stop searching across specifications or collecting data once the desired level of significance is achieved).

To test whether a p-curve is right or left skewed, one can construct what the authors call a “pp-value,” or p-value of the p-value—the probability of observing a significant p-value at

⁴ Note that when constructing z-scores from regression coefficients and standard errors, rounding may lead to an artificially large number of round or even integer z-scores. Brodeur et al. (2016) reconstruct original estimates by randomly redrawing numbers from a uniform interval (i.e. a standard error of 0.02 could actually be anything in the interval $[0.015, 0.025]$). This does not alter results significantly.

⁵ Unlike economics journals, which often use asterisks or other notation to separately indicate p-values (0,.01), $[0.01 < .05]$, and $[\.05, .1]$, psychology journals often indicate only whether a p-value is $< .05$, and this is the standard used throughout (Simonsohn, Nelson, and Simmons 2014a).

least as extreme if the null were true—and then aggregate the pp-values in a literature with Fisher’s method and test for skew with a χ^2 test. The authors also suggest a test of comparing whether a p-curve is flatter than the curve that would result if studies were (somewhat arbitrarily) powered at 33%, and interpret a p-curve that is significantly flatter or left skewed than this as lacking in evidentiary value. The p-curve can also potentially be used to correct effect size estimates in literatures suffering from publication bias; corrected estimates of the “choice overload” literature exhibit a change in direction from standard published estimates (Simonsohn, Nelson, and Simmons 2014b).⁶

Thanks to the existence of study registries and ethical review boards in clinical medical research, it is increasingly possible to survey nearly the universe of studies that have been undertaken, along the lines of Franco, Malhotra, and Simonovits (2014). Easterbrook et al. (1991) reviewed the universe of protocols submitted to the Central Oxford Research Ethics Committee, and both Turner et al. (2008) and Kirsch et al. (2008) employ the universe of tests of certain anti-depressant drugs submitted to the FDA, and all found significantly higher publication rates when tests yield statistically significant results. Turner et al. found that 37 of 38 (97%) of trials with positive, i.e., statistically significant, results were published, while only 8 of 24 (33%) with null (or negative) results were published; for a meta-meta-analysis of the latter two studies, see Ioannidis (2008).

A simple model of publication bias described in McCrary, Christensen, and Fanelli (2015) suggests that, under some relatively strong assumptions regarding the rate of non-publication of statistically non-significant results, readers of research studies could potentially adjust their significance threshold to “undo” the distortion by using a more stringent t-test statistic of 3.02 (rather than 1.96) to infer statistical significance at 95% confidence. They note that approximately 30% of published test statistics in the social sciences fall between these two cutoffs. It is also possible that this method would break down and result in a “t-ratio arms race” if all researchers were to use it, so it is mostly intended for illustrative purposes.

As an aside, it is also possible that publication bias could work *against* rejection of the null hypothesis in some cases. For instance, within economics in cases where there is a strong theoretical presumption among some scholars that the null hypothesis of no effect is likely to

⁶ For an online implementation of the p-curve, see <http://p-curve.com>. Also see a discussion of the robustness of the test in Ulrich and Miller (2015); Simonsohn, Simmons, and Nelson (2015a).

hold (e.g., in certain tests of market efficiency) the publication process could be biased by a preference among editors and referees for non-rejection of the null hypothesis of no effect. This complicates efforts to neatly characterize the nature of publication bias, and may limit the application of the method in McCrary, Christensen, and Fanelli (2015).

Taken together, a growing body of evidence indicates that publication bias is widespread in economics and many other scientific fields. Stepping back, these patterns do not appear to occur by chance, but are likely to indicate some combination of selective editor (and referee) decision-making, the file-drawer problem alluded to above, and/or widespread specification searching (the focus of the next subsection), which is closely related to what the Ioannidis (2005) model calls author bias.

2.2.1 Publication bias in several empirical economics literatures

Scholars working in several specific literatures within economics have argued for the presence of considerable publication bias, including labor economics literatures on minimum wage impacts and on the value of a statistical life, and we discuss both briefly here, as well as several other bodies of evidence in economics.

Card and Krueger (1995) conducted a meta-analysis of the minimum wage and unemployment literature, and test for the “inverse-square-root” relationship between sample size and t-ratio that one would expect if there was a true effect and no publication bias, since larger samples should generally produce more precise estimates (for a given research design).⁷ They find that t-statistics from the 15 studies using quarterly data available at the time of writing are actually *negatively* correlated with sample sizes. A possible explanation is that a structural change in the effect of the minimum wage (a decline over time) has taken place, but the authors consider publication bias and specification searching a more likely explanation. Neumark and Wascher (1998) construct an alternative test for publication bias, which produces an attenuation of the effect size with larger sample sizes (as sample sizes increased over time) that is

⁷ Card and Krueger explain: “A doubling of the sample size should lower the standard error of the estimated employment effect and raise the absolute t ratio by about 40 percent if the additional data are independent and the statistical model is stable. More generally, the absolute value of the t ratio should vary proportionally with the square root of the number of degrees of freedom, and a regression of the log of the t ratio on the log of the square root of the degrees of freedom should yield a coefficient of 1.” In a similar test in political science, Gerber, Green, and Nickerson (2001) document likely publication bias in the voter mobilization campaign literature, showing that studies with larger sample sizes tend to produce smaller effect size estimates.

qualitatively similar to that in Card and Krueger (1995), but Neumark and Wascher thus place more emphasis on the structural change explanation (i.e., actual effects declined over time) and discount the possibility of publication bias. Another explanation has been proposed for Card and Krueger's findings: the simple lack of a true effect of the minimum wage on unemployment. If the null hypothesis of no effect is true, the t-statistic would have no relationship with sample size. Studies that advance this alternative explanation (Stanley 2005; Doucouliagos and Stanley 2009) argue that the minimum wage literature does likely suffer from some publication bias, since many studies' t-statistics hover around 2, near the standard 95% confidence level, and other tests, described below in Section 3, indicate as much.

Several studies have also documented the presence of publication bias in the literature estimating the value of a statistical life (VSL). As government regulations in health, environment and transportation are frequently based on this value, accurate estimation is of great public importance, but there is growing consensus that there is substantial publication bias in this literature, leading to a strong upward bias in reported estimates (Ashenfelter and Greenstone 2004). Using the collection of 37 studies in Bellavance, Dionne, and Lebeau (2009), Doucouliagos, Stanley, and Giles (2012) find that correcting for publication bias (using an approach we discuss below in section 3.1.2) reduces the estimates of VSL by 70-80% from that produced by a standard meta-analysis regression. Similar analysis shows that, correcting for publication bias, the VSL also appears largely inelastic to individual income (Doucouliagos, Stanley, and Viscusi 2014). An updated analysis of publication bias in the VSL literature (Viscusi 2015) shows that although publication bias is large and leads to meaningfully inflated estimates, he argues much of it may stem from early studies in the literature that used voluntary reporting of occupational fatalities, while more recent studies estimates employing the Census of Fatal Occupational Injuries (CFOI) suffer from less measurement error and tend to produce larger estimates.

Evidence for publication bias has been documented in many other economics research literatures, although not in all. See Longhi, Nijkamp, and Poot (2005) and Knell and Stix (2005) for notable examples. Table 2 describes a number of related publication bias studies that might be of interest to readers, but for reasons of space they are not discussed in detail here. In the most systematic approach to date (to our knowledge), Doucouliagos and Stanley (2013) carry out a meta-meta-analysis of 87 meta-analysis papers (many of which are reported in Table 2), and find

that over half of the literatures suffer from “substantial” or “severe” publication bias, with particularly large degrees of bias in empirical macroeconomics and in empirical research based on demand theory, and somewhat less publication bias in subfields with multiple contested economic theories. (Of course, and not to be facetious, one cannot completely rule out publication bias even among this body of publication bias studies.)

The *Journal of Economic Surveys* has published many meta-regression papers, including a special issue devoted to meta-regression and publication bias (Roberts 2005). The statistical techniques for assessing publication bias are summarized in Stanley (2005), and many of these are applied in the articles listed in Table 2. One common data visualization approach is the use of funnel graphs, see Stanley and Doucouliagos (2010), Light and Pillemer (1984), and our discussion in section 3 below.

2.2.2 Publication Bias and Effect Size

Another important issue related to publication bias and null hypothesis testing is the reporting of the magnitude of effect sizes. Although it appears that economics may fare somewhat better than other social science disciplines in this regard, since economics studies typically report regression coefficients and standard errors while articles in some other disciplines (e.g., psychology) have historically only reported p-values, there is some evidence that under-reporting of effect magnitudes is still a concern. In a review in this Journal, McCloskey and Ziliak (1996) find that 70% of full-length *American Economic Review* articles did not distinguish between statistical and practical significance. Follow-up reviews in 2004 and 2008 conclude that the situation had not meaningfully improved (Ziliak and McCloskey 2004; Ziliak and McCloskey 2008).

DeLong and Lang (1992) is an early contribution that addresses the issue of publication of null findings and effect sizes. They show that only 78 of 276 null hypotheses tested in empirical papers published in leading economics journals at the time were not rejected. This is generally equivalent to affirming an underlying economic model, since null hypotheses typically test for a zero effect. However, using the uniform distribution of p-values under a true null hypothesis, and the startling lack of published p-values close to 1, they conclude it is likely that practically all tested statistical null hypotheses in empirical economics are indeed false. They also conclude that the null results that actually do get published in journals may also result from publication bias: a null result is arguably more interesting if it contradicts previous statistically significant results. DeLong and Lang go on to suggest that since almost all tested null hypotheses

in economics are false, empirical evidence should pay more attention to practical significance and effect size rather than statistical significance alone, as is too often the case.

2.3 Specification Searching

While publication bias implies a distortion of a body of multiple research studies, bias is also possible within any given study (for instance, as captured in the author bias term u in Ioannidis (2005)). In the 1980's and 1990's, expanded access to computing power led to rising concerns that some researchers were carrying out growing numbers of analyses and selectively reporting econometric analysis that supported pre-conceived notions – or were seen as particularly interesting within the research community – and ignoring, whether consciously or not, other specifications that did not.

One the most widely cited articles from this period is Leamer's (1983), “Let's Take the Con Out of Econometrics”, which discusses the promise of improved research design (namely, randomized trials) and argues that in observational research, researchers ought to transparently report the entire range of estimates that result from alternative analytical decisions. Leamer's illustrative application employs data from a student's research project, namely, U.S. data from 44 states, to test for the existence of a deterrent effect of the death penalty on the murder rate. (These data are also used in McManus 1985.) Leamer classifies variables in the data as either “important” or “doubtful” determinants of the murder rate, and then runs regressions with all possible combinations of the doubtful variables, producing a range of different estimates. Depending on which set of control variables, or covariates, were included (among state median income, unemployment, percent population nonwhite, percent population 15-24 years old, percent male, percent urban, percent of two-parent households, and several others), the main coefficient of interest—the number of murders estimated to be prevented by each execution—ranges widely on both sides of zero, from 29 lives saved to 12 lives lost. Of the five ways of classifying variables as important or doubtful that Leamer evaluated, three produced a range of estimates that included zero, suggesting that inference was quite fragile in this case.

Leamer's recommendation that observational studies employ greater sensitivity checks, or extreme bounds analysis (EBA), was not limited to testing the effect of including different combinations of covariates, as in Leamer (1983). More detailed descriptions of EBA in Leamer (1978) and Leamer and Leonard (1983) explain that, if provided two “doubtful” control variables z_1 and z_2 , and an original regression $y_t = \beta x_t + \gamma_1 z_{1t} + \gamma_2 z_{2t} + u_t$, researchers should define a

composite control variable $w_t(\theta) = z_{1t} + \theta z_{2t}$, should allow θ to vary, and then report the range of estimates produced by the regression $y_t = \beta x_t + \eta w_t(\theta) + u_t$. The recommendations that flowed from Leamer's EBA were controversial, at least partly because they exposed widespread weaknesses in the practice of applied economics research at the time, and perhaps partly due to Leamer's often pointed (or humorous, we think) writing style. Few seemed eager to defend the state of applied economics, but many remained unconvinced that sensitivity analysis, as implemented with EBA, was the right solution. In "What Will Take the Con out of Econometrics" (McAleer, Pagan, and Volker 1985), critics of EBA sensibly considered the choice of which variables to deem important and which doubtful just as open to abuse by researchers as the original issue of covariate inclusion.

Echoing some of Leamer's (1983) recommendations, a parallel approach to bolstering applied econometric inference focused on improved research design instead of sensitivity analysis. LaLonde (1986) applied widely-used techniques from observational research to data from a randomized trial and showed that none of the methods reproduced the experimentally identified, and thus presumably closer to true, estimate.⁸

Since the 1980's, empirical research practices in economics have changed significantly, especially with regards to improvements in research design. Angrist and Pischke (2010) make the point that improved experimental and quasi-experimental research designs have made much econometric inference more credible. However, Leamer (2010) argues that researchers retain a significant degree of flexibility in how they choose to analyze data, and that this leeway could introduce bias into their results.

This flexibility was highlighted in Lovell (1983), who shows that with a few assumptions regarding the variance of the error terms, searching for the best k of c explanatory variables means that a coefficient that appears to be significant at the level $\hat{\alpha}$ is actually only significant at the level $1 - (1 - \hat{\alpha})^{c/k}$. In the case of $k = 2$ and 5 candidate variables, this risks greatly overstating significance levels, and the risk is massive if there are, say, 100 candidate variables. Lovell (1983) goes on to argue for the same sort of transparency in analysis as Leamer (1983). Denton (1985) expands on Lovell's work and shows that data mining can occur as a collective

⁸ In a similar spirit, researchers have more recently called attention to the lack of robustness in some estimates from random-coefficient demand models, where problems with certain numerical maximization algorithms may produce misleading estimates (Knittel and Metaxoglou 2011; Knittel and Metaxoglou 2013); McCullough and Vinod (2003) contains a more general discussion of robustness and replication failures in nonlinear maximization methods.

phenomenon even if each individual researcher tests only one pre-stated hypothesis, if there is selective reporting of statistically significant results, an argument closely related to the file drawer publication bias discussion above (Rosenthal 1979).

Related points have been made in other social science fields in recent years. In psychology, Simmons, Nelson, and Simonsohn “prove” that listening to the Beatles’ song “When I’m Sixty-Four” made listeners a year and a half younger (Simmons, Nelson, and Simonsohn 2011). The extent and ease of this “fishing” in analysis is also described in political science by Humphreys, Sierra, and Windt (2013) who use simulations to show how a multiplicity of outcome measures and of heterogeneous treatment effects (sub-group analyses) can be used to generate a false positive, even with large sample sizes. In statistics, Gelman and Loken (2013) agree that “[a] dataset can be analyzed in so many different ways (with the choices being not just what statistical test to perform but also decisions on what data to [include] or exclude, what measures to study, what interactions to consider, etc.), that very little information is provided by the statement that a study came up with a $p < .05$ result.”

The greater use of extra robustness checks in applied economics is designed to limit the extent of specification search, and is a shift in the direction proposed by Leamer (1983), but it is unclear how effective these changes are in reducing bias in practice. As noted above, the analysis of 641 articles from three top economics journals in recent years presented in Brodeur et al. (2016) still shows a disturbing two-humped distribution of p-values, with relatively few p-values between 0.10 and 0.25 and far more just below 0.05. Their analysis also explores the correlates behind this pattern, and finds that this apparent misallocation of p-values just below the accepted statistical significant level was less pronounced for articles written by tenured authors, and tentatively find it less pronounced among studies based on randomized controlled trials (suggesting that improved research design itself may partially constrain data mining), but they did not detect any discernible differences in the pattern based on whether the authors had publicly posted the study’s replication data in the journal’s public archive.

2.3.1 Sub-Group Analysis

One area of analytical flexibility that appears particularly important in practice is sub-group analysis. In many cases, there are multiple distinct interaction effects that could plausibly be justified by economic theory, and current datasets have a growing richness of potential

covariates. Yet it is rare for applied economics studies to mention how many different interaction effects were tested, increasing the risk that only statistically significant false positives are reported.

While there are few systematic treatments of this issue in economics, there has been extensive discussion of this issue within medical research, where the use of non-prespecified subgroup analysis is strongly frowned upon. The FDA does not use subgroup analysis in its drug approval decisions (Maggioni et al. 2007). An oft repeated, and humorous, case comes from a trial of aspirin and streptokinase use after heart attacks conducted in a large number of patients (N=17,187). Aspirin and streptokinase were found to be beneficial, except for patients born under Libra and Gemini, for whom there was a harmful (but not statistically significant) effect (ISIS-2 COLLABORATIVE GROUP 1988). The authors included the zodiac subgroup analysis because journal editors had suggested that 40 subgroups be analyzed, and the authors relented under the condition that they could include a few subgroups of their own choosing to demonstrate the unreliability of such analysis (Schulz and Grimes 2005).

2.4 Inability to Replicate Results

2.4.1 Data Availability

There have been longstanding concerns within economics over the inability to replicate the results of specific published papers. The pioneering example is a project undertaken by the *Journal of Money, Credit, and Banking* (JMCB) (Dewald, Thursby, and Anderson 1986). The journal launched the JMCB Data Storage and Evaluation Project with NSF funding in 1982, which requested data and code from authors who published papers in the journal.⁹ Despite the adoption of an explicit policy of data sharing by the JMCB during the project, only 78% of authors provided data within six months after multiple requests, although this was certainly an improvement over the 34% data sharing rate in the control group, namely, those who published before the new journal policy went into effect. Of the papers that were still under review by the JMCB at the time of the requests for data, one quarter did not even respond to the request,

⁹ Note that the NSF has long had an explicit policy of expecting researchers to share their primary data, though there seems to be minimal enforcement. “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing”; see <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.

despite the request coming from the same journal considering their paper. The data that was submitted was often an unlabeled and undocumented mess, a problem that has persisted with recent data sharing policies, as discussed below. Dewald, Thursby, and Anderson (1986) attempted to replicate nine empirical papers, and despite extensive assistance from the original authors, they were often unable to reproduce the papers' published results.

Little changed for a long time after the publication of this landmark article. A decade later, in a follow-up piece to the JMCB Project published in the *Federal Reserve Bank of St. Louis Review*, Anderson and Dewald (1994) note that only two economics journals other than the *Review* itself, namely, the *Journal of Applied Econometrics* and the *Journal of Business and Economic Statistics*, systematically requested replication data from authors, though neither requested the associated statistical code. The JMCB itself had discontinued its policy of requesting replication data in 1993 (though it reinstated it in 1996). The authors repeated their experiment with papers presented at the St. Louis Federal Reserve Bank conference in 1992, and obtained similarly discouraging response rates as in the original JMCB Project.

The first “top five” general interest economics journal to systematically request replication data was the *American Economic Review* (AER), which began requesting data in 2003. After a 2003 article (McCullough and Vinod 2003) showed that nonlinear maximization methods from different software packages often produced wildly different estimates, that not a single *AER* article had tested their solution across different software packages, and that fully half of queried authors from a chosen issue of the *AER*, including a then-editor of the journal, had failed to comply with the policy of providing data and code, editor Ben Bernanke made the data and code sharing policy mandatory in 2004 (Bernanke 2004; McCullough 2007). The current *AER* data policy states:

“It is the policy of the *American Economic Review* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the *Review*, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the *AER* Web site. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.”¹⁰

¹⁰ <https://www.aeaweb.org/aer/data.php>

In addition to all the journals published by the American Economic Association (including this *Journal*, the *American Economic Journals*, and the *Journal of Economic Perspectives*), several other leading journals, including *Econometrica*, the *Journal of Applied Econometrics*, the *Journal of Money Credit and Banking*, *Journal of Political Economy*, *Review of Economics and Statistics*, and the *Review of Economic Studies*, now explicitly require data and code to be submitted at the time of article publication. The last of what are typically considered the leading general interest journals in the profession, the *Quarterly Journal of Economics*, finally adopted a data sharing requirement (that of the American Economic Association Journals) in April 2016¹¹.

Table 3 summarizes journal policies regarding data sharing, publication of replications or comments, and funding or conflict of interest disclosures at twelve of the top economics and finance journals (according to Scientific Journal Rankings). There has clearly been considerable progress along all of these dimensions over the past decade, but journal policies remain a mixed bag. Among these leading journals, most but not all now have some data sharing requirements, and are officially open to publishing papers that could be considered “replications.”¹² There is also greater use of disclosure statements.

The *AER* conducted a self-review and found relatively good, though still incomplete, compliance with its data sharing policy (Glandon 2010). Despite this positive self-assessment, others observers believe that much work remains to ensure greater access to replication data in economics. Recent studies document that fewer than 15 of over 150 articles in the JMCB archive could be replicated; there is typically little to no verification that the data and code submitted to journals actually generate the published results; and the majority of economics journals still have no explicit data sharing requirements (McCullough, McGeary, and Harrison 2006; Anderson et al. 2008; McCullough 2009).

The uneven nature of progress along these dimensions across economics journals is mirrored in the patterns observed in other research disciplines. Medical research tends to have relatively little public data sharing, partly due to the stringency of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), although it is thought that some researchers may use the law as a pretext for avoiding greater transparency (Annas 2003; Malin, Benitez, and

¹¹ http://www.oxfordjournals.org/our_journals/qje/for_authors/data_policy.html

¹² Though leading journals are officially open to publishing replications, they appear to publish few replication studies in practice.

Masys 2011). An increasing number of political science journals are now requiring data sharing (Gherghina and Katsanidou 2013), with a few journals (e.g. *International Interactions*, *Political Science Research and Methods*) doing at least some degree of in-house verification of results, and the *American Journal of Political Science* contracting out the verification to a third party¹³. A leading group of political scientists created the Data Access and Research Transparency (DART) statement, which includes data sharing requirements. That statement has been incorporated into the ethics guidelines of the American Political Science Association, and has since been adopted by nearly thirty political science journals.¹⁴ In psychology, one leading journal, *Psychological Science*, undertook drastic policy changes in early 2014 to increase transparency and reproducibility under editor Eric Eich (Eich 2014) and these have continued under the current editor (Lindsay 2015). The changes include the introduction of “badges” included in the article itself signifying open data, open materials, and pre-registration of hypotheses, which has helped spawn an increase in data availability.¹⁵

2.4.1.1 Proprietary Data

The American Economic Association’s journal data sharing policy – which has been adopted by several other journals and organizations nearly verbatim, as shown in Table 3 – allows for some exceptions, importantly, for proprietary data. In particular, the policy reads: “The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.”

In practice, this exemption is requested fairly often by empirical researchers, and the rate is increasing over time. During the past decade, the May *American Economic Review Papers & Proceedings* issue has featured a “Report of the Editor” which details the number of submissions to the journal, as well the number of papers published, those with data, and those that were granted exemptions. Figure 2 presents the percentage of papers in each issue of the *AER* since 2005 (when information becomes available) through 2015. A few patterns are noteworthy. First, proportion of papers that include data has risen over time, starting at roughly 60% and since

¹³ The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, see <https://ajpsblogging.files.wordpress.com/2015/03/ajps-guide-for-replic-materials-1-0.pdf>.

¹⁴ See <http://www.dartstatement.org/>. Accessed October 10, 2016.

¹⁵ More information on badges can be found here: http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/badges or here: <https://osf.io/tvyxz/wiki/home/>, and information on their influence on *Psychological Science* here: <http://www.psychologicalscience.org/index.php/publications/observer/obsonline/open-practice-badges-in-psychological-science-18-months-out.html>.

increasing into the 70-80% range, capturing the shift towards empirical research in the discipline as a whole. During this period, the proportion of papers using data that received exemptions from the data-sharing policy has risen rapidly, from roughly 10% to 40% over time. Thus replication data is in practice not available in practice for nearly half of all empirical papers published in the *AER* in recent years.

There are many common sources of proprietary or otherwise non-sharable data driving this trend. One of the most common is U.S. government data. There are currently 23 Federal Statistical Research Data Centers (RDC), which provide researchers access to sensitive federal government data that cannot simply be shared publicly on a journal website, typically due to individual or corporate privacy concerns (e.g., IRS tax records).¹⁶ We do not believe that research conducted with this data should be penalized in any way, and in fact, studies employing administrative data may be particularly valuable both intellectually and in terms of public policy decisions. However, despite the exemption from data sharing, it would still be useful for researchers (and journals) to make their work as reproducible as possible given the circumstances, for instance, by at least posting the associated statistical code and providing details about how other scholars could gain similar access to the data. Beyond government data, there are, of course, also an increasing number of proprietary datasets created by corporations or other entities that are willing to share sensitive commercial data with researchers, but not with the public at large where similar issues arise.

Beyond commercially proprietary or legally restricted government data, there is also the important issue of norms regarding the sharing of original data collected by scholars themselves. Given the years of effort and funding that goes into creating an original dataset, what special intellectual property rights (if any) do scholars involved in generating data have?

Economists are likely aware of the incentives created by temporary monopoly rights to intellectual property, and in many ways the issues regarding original data collection are closely linked to traditional arguments around granting private patents. Such monopoly rights, even if temporary, could theoretically be socially beneficial if they help to drive the creation of innovative new data sources, such as the explosion of original new survey datasets in development economics over the past two decades. Yet we know of no empirical research that

¹⁶ For more information on researcher access to, and National Science Foundation (NSF) funding for, US administrative data, see (Card et al. 2010; Mervis 2014a; Moffitt 2016) and Cowen and Tabarrok (2016), the latter of which also calls for NSF funding of replications, open data, and greater dissemination of economics research.

discusses the optimal length of such “research dataset” patents; this is an area that demands further attention, especially around the optimal length of exclusive access afforded to originators of new data.¹⁷

The increasingly common requirement to share data at the time of journal publication is a cause for concern in some fields. For example, in response to a proposal from the International Committee of Medical Journal Editors to require data sharing within six months after the publication of an article (Taichman et al. 2016) an editorial in the leading *New England Journal of Medicine* caused an outcry when they responded by describing those who do secondary analysis without the co-authorship and cooperation of the original data collecting author as “research parasites” (Longo and Drazen 2016). The journal re-affirmed their commitment to data-sharing (Drazen 2016) and published a supporting piece by Senator Elizabeth Warren (Warren 2016), but also a separate piece calling for a longer embargo period after publication: “2 years after publication of the primary trial results and an additional 6 months for every year it took to complete the trial, with a maximum of 5 years before trial data are made available to those who were not involved in the trial” (The International Consortium of Investigators for Fairness in Trial Data Sharing 2016). Presumably the increasing “patent length” here for each additional year it took to complete data collection is an attempt to reward research effort in collecting unusually rich longitudinal data. Yet these sorts of rules regarding timeframes seem quite ad hoc (to us, at least), further highlighting the need for a more serious examination of how best to balance the research community’s right to replicate and extend existing research with scholars’ incentives to invest in valuable original data.

In political science, many journals have recently adopted policies similar to the AEA policy described above. For example, the current policy of the *American Journal of Political Science* states: “In some limited circumstances, an author may request an exemption from the replication and verification policy. This exemption would allow the author to withhold or limit public access to some or all of the data used in an analysis. All other replication materials (e.g., software commands, etc.) still must be provided. The primary reasons for such exemptions are restricted access datasets and human subjects protection.”¹⁸ We lack data on how often this exemption is granted, however. Additionally, this journal goes much further than economics

¹⁷ Unlike a long line of empirical research on the optimal patent length for research and design such as Mansfield, Schwartz, and Wagner (1981),

¹⁸ See <https://ajps.org/ajps-replication-policy/>, accessed October 10, 2016.

journals in one important way: instead of simply collecting and publishing data and code from authors, the editors use a third party research center (namely, the [Odum Institute for Research in Social Science](#), at the [University of North Carolina, Chapel Hill](#) for quantitative analysis and the [Qualitative Data Repository](#) (QDR), at [Syracuse University](#) for qualitative analyses) to verify that the data and statistical code produce the published results.

2.4.2 Types of Replication Failures and Examples

There have been multiple high-profile examples in economics of cases where replication authors have claimed they are unable to replicate published results, including on topics of intense public policy interest.

It is unclear (to us, at least) exactly how pervasive the issues of lack of replicability are in economics, and thus how much confidence we should have in the body of published findings, and this is a topic on which future research should aim to gather more systematic evidence. It could certainly be the case that researchers – as well as graduate students in their courses, in a growing number of Ph.D. training programs – usually are able to successfully replicate published results, but that this unremarkable exercise of successfully verifying published results escapes our notice because researchers do not seek to publish their work (or that editors choose not to publish it). Yet in the absence of systematic standards regarding data sharing and replication, and given examples such as those discussed below in which there are discrepancies between the original published findings and later replication results, it remains possible that the high-profile cases of failed replication may simply be the tip of the iceberg. Thankfully, a few recent papers have begun to provide some evidence on this question, which we highlight below.

We ourselves are no strangers to replication and re-analysis debates: papers by one of the authors of this article, described below, have been part of lively debates on replication and re-analysis using data that we shared publicly. These debates have led us to appreciate the great promise of replication research, as well as its potential pitfalls: exactly like original research studies, replication studies have their own particular strengths and weaknesses, and may serve to either advance the intellectual debate or could obscure particular issues. Yet there is no doubt in our minds that an overall increase in replication research will serve a critical role in establishing the credibility of empirical findings in economics, and in equilibrium, will create stronger incentives for scholars to generate more reliable results.

Further complicating matters, an imprecise definition of the term “replication” itself often leads to confusion. A taxonomic proposal in Hamermesh (2007) distinguished between “pure”, “statistical”, and “scientific” replications, while a more recent effort (Clemens 2015) uses the terms “verification,” “reproduction,” “reanalysis,” and “extension” to distinguish between replications (the first two) and robustness exercises (the latter two). We first present some existing evidence on the replicability of economics and social science research in the next subsection, and then provide examples of each of Clemens’ categories.

2.4.2.1 Evidence on Replication in Economics

The articles in the 1986 *Journal of Money Credit and Banking* project and the 1994 St. Louis Federal Reserve follow-up mentioned above provided some of the first attempts at systematic replication in economics, with fairly discouraging results. Have things improved in the last few decades?

New evidence is emerging about the reliability of empirical economics research. One of the most important recent studies is Camerer et al. (2016), which repeated 18 behavioral economics lab experiments originally published between 2011 and 2014 in the *American Economic Review* and the *Quarterly Journal of Economics* to assess their replicability. Figure 3 below reproduces a summary of their findings. Their approach is similar in design to a large scale replication of 100 studies in psychology known as the “Replication Project: Psychology”, which we discuss in detail below. The replication studies were designed with sample sizes that aimed to have 90% power to detect the original effect size at the 5% significance level. In all, the estimated effects were statistically significant with the same sign in 11 of the 18 replication studies (61.1%). This is a moderate, though perhaps not entirely demoralizing, rate of replicability. Yet there is still no single accepted standard of what it means for a study to successfully replicate another, and different definitions provide somewhat more positive assessments of replicability. For instance, in 15 of the 18 replication studies (83.3%), estimated effects lie within a 95% “prediction interval” (which acknowledges sampling error in both the original study and the replication); one further replication estimate was far larger in magnitude than the original estimate, arguably raising the replication rate to 89%.¹⁹ Overall, it is reasonable

¹⁹ See Patil, Peng, and Leek (2016) and the discussion below regarding prediction intervals. An interesting, if sad, detail of the difficulties of replication is highlighted in the *Science* news article covering the results of the Camerer

to conclude from this study that the body of recent experimental economics lab studies (at least in the leading journals) is unlikely to be riddled with spurious findings.

Camerer et al. (2016) also included both a survey and a novel prediction market to assess observers' (mostly Ph.D. students and post-doctoral researchers, as well as professors, recruited via e-mail) priors on whether the studies would in fact successfully replicate. Both the survey and market measures were somewhat more optimistic about replicability than the actual outcomes (described above), and the prediction market did not significantly outperform the survey beliefs. Statistical tests of the correlation of a successful replication outcome with the p-value and sample size of the original study reveal significant relationships in the expected directions, namely, a negative correlation with the p-value (in other words, studies with smaller p-values were more likely to replicate) and a positive correlation with sample size, where the latter result presumably implies that original results based on larger samples were less likely to have been spuriously driven by sampling variation.

Beyond experimental economics, a recent working paper by Andrew Chang and Phillip Li systematically tested the reproducibility of 67 macroeconomics papers (Chang and Li 2015). Chang and Li deliberately sampled a wider variety of journals, choosing 13 journals and articles from July 2008 to October 2013 and for comparability all papers that have an empirical component, model estimation with only U.S. data, and have a key result based on U.S. GDP figures. Of the 67 papers, 6 use proprietary data and are thus excluded from consideration. Thirty five articles are published in journals with data and code sharing requirements, but Chang and Li could obtain data for only 28 of these (80%) from the journal archives, suggesting limited enforcement of this requirement in many cases. Web search and e-mails to authors netted only one of the remaining seven missing datasets. Of the 26 papers in journals without data sharing requirements, Chang and Li were unable to obtain 15 datasets (58%).

With this data in hand, the overall replication success rate is 29 of 67 (43%) overall, or 29 of 61 (48%) among those using non-proprietary datasets, so roughly half. Though missing data is the largest source of replication failures, "incorrect data or code" accounts for the inability to replicate 9 papers. It should be noted that Chang and Li use a qualitative definition of replication, and test only key results of the paper, and this appears to lead to a fairly generous interpretation

et al study (Bohannon 2016). One of the replicated studies (Ifcher and Zarghamee 2011) originally showed subjects a clip of comedian Robin William to test if happiness (positive affect) impacts time preference. The replication took place after William's tragic suicide, so the video could easily induce a different emotional state in the replication.

of replicability. They write: “For example, if the paper estimates a fiscal multiplier for GDP of 2.0, then any multiplier greater than 1.0 would produce the same qualitative result (i.e., there is a positive multiplier effect and that government spending is not merely a transfer or crowding out private investment).” To our minds, this is evidence that even when data is available (which it sometimes not) a non-negligible fraction of empirical economics research cannot be reproduced, even when using the original data and a relatively non-stringent conceptual understanding of what constitutes replication success.

Other examples of replication failures abound. Clemens (2015) provides a useful taxonomy, and we provide an example of from each of the categories there to help distinguish between them, namely the two types of replication he discusses (verification and reproduction), and the two types of robustness exercises (reanalysis and extension). Of course not all papers fit easily into one of these categories as most tend to include elements from multiple categories.

2.4.2.2 Verification

Perhaps the most straightforward type of replication in economics involves using the same specification, the same sample, and the same population. Essentially, this is running the same code on the same data and testing if you get the same results. Hamermesh (2007) referred to this as a “pure replication.” We believe this basic standard should be expected of all published economics research, and hope this expectation is universal among researchers. One tiny tweak to the definition of verification is that it also includes errors in coding. If an author describes a statistical test in the paper, but the code indisputably does not correctly carry out the test as described, this is also considered a verification failure.

One of the earliest cases of quantitative economics research failing a verification test comes from an investigation of the effect of social security on private savings. Feldstein (1974) estimates a life cycle model showing that social security reduces private savings by as much as 50%. There were significant theoretical challenges to carrying out this exercise related to assumptions about the intergenerational transfer of wealth, but Leimer and Lesnoy (1982) discovered that a flaw in Feldstein’s computer program that overestimated the growth rate of Social Security wealth for widows led to larger effects of Social Security wealth than when the mistake was corrected.

Feldstein replied to the critique saying he was grateful for having the error corrected, but that the central conclusion of the study remains largely unchanged (namely, that Social Security

decreased private savings by 44%) (Feldstein 1982). Much of the change in coefficients in the replication exercise resulted from Leimer and Lesnoy including an expanded time series of data—this is not a failure of verification, but rather an extension, which we discuss below. Feldstein asserted that this was unwise because of an important 1972 change in Social Security law that bookended the original sample period. When including post-1972 data and modifying the Social Security wealth variable in a way to account for the change, Feldstein estimated a slightly larger deterrent effect of Social Security on private savings.

Clemens (2015) contains a larger selection of examples (see his Table 3).²⁰ In many (but not all) cases discussed in Clemens, the original authors clearly admit to the failure of verification, but there is vigorous, and we think, healthy scholarly debate about how important those mistakes are and whether the results are still significant – statistically and/or practically – when the code or data are corrected. Of course, authors whose papers are subject to replication debates should be commended for providing other scholars with access to their data and code public in the first place, especially for these earlier articles published before journal data-sharing requirements were established.

2.4.2.3 Reproduction

The other type of replication in Clemens' taxonomy is a reproduction. This approach uses the same analytical specification and the same population, but a different sample. Hamermesh (2007) refers to this as a statistical replication.

In economics, this approach would be exhibited in a study that generated a certain set of results using a 5% sample of the census, while a different 5% census sample produced different results, or an experimental economics lab study that produced one set of results with a certain sample while the reproduction study analyzed a different sample from broadly the same population (e.g., U.S. university students).

There is, of course, some gray area and room to debate as to the definition of what constitutes a given population. If we consider U.S. college undergraduates the population (and do

²⁰ Other well-known recent examples of verification debates in empirical economics include Donohue and Levitt (2001), Foote and Goetz (2008) and Donohue and Levitt (2008) on legalized abortion and crime rates; and Reinhart and Rogoff (2010) and Herndon, Ash and Pollin (2014) on growth rates and national debt. In the debate over Hoxby's (2000) results regarding school competition in Rothstein (2007) and Hoxby (2007), the possibility is discussed that one factor contributing to lack of verification is that intermediary datasets constructed from raw data were over-written when the raw data was updated, as sometimes happens with U.S. government data. The work of one of the authors of this paper could be included on this list; see Miguel and Kremer (2004), Aiken et al. (2015) and Hicks, Kremer and Miguel (2015) on the impact of school-based deworming in Kenya.

not differentiate by campus), or Amazon MTurk-ers, some of the failures of replication in Camerer et al. (2016) could be better classified as failures of reproduction, as long as the samples were in fact collected in broadly the same manner (i.e., in person vs. online).

Reproduction failures are perhaps more precisely defined in the hard sciences where experimenters routinely attempt to do the exact same physical process as another lab, albeit with a different sample of molecules, or in the biological sciences where experiments may employ a different sample of animal subjects. For instance, in defining reproduction, Clemens mentions the infamous case of the “discovery” of cold fusion by Fleischmann and Pons (1989), which failed to reproduce in Lewis et al. (1989).

2.4.2.4 Reanalysis

Robustness exercises come in two varieties, reanalysis and extensions.

Reanalysis uses a different analytical specification on the same population (with either the same or a different sample). Many economics replication studies include both a verification aspect as well as some re-analysis. For instance, Davis (2013) conducts a successful verification of Sachs and Warner (1997), but concludes that reanalysis shows the estimates are somewhat sensitive to different statistical estimation techniques. Other well-known recent reanalysis debates in empirical economics include Miguel, Satyanath, and Sergenti (2004), Ciccone (2011) and Miguel and Satyanath (2011) on civil conflict and GDP growth using rainfall as an instrumental variable; and Acemoglu, Johnson, and Robinson (2001), Albouy (2012), and Acemoglu, Johnson, and Robinson (2012) on institutions and GDP growth with settler mortality as an instrumental variable.

The debates over these and other studies makes it clear that reanalysis does not typically settle all key research questions, and the exercise often reveals that empirical economists have considerable flexibility in their analytical choices. This insight makes the development of methods to account for – and possibly constrain – this flexibility, which we discuss below in section 3, all the more important.

2.4.2.5 Extension

Under Clemens’ classification system, an extension uses the same analytical specification as an original study but a different population and a different sample. Most often this would be conducting the same analysis carried out in a different time or place.

A well-known example of an extension involves Burnside and Dollar (2000), which showed that foreign aid seemed to be effective in increasing GDP if the recipient country was well-governed. However, using the exact same regression specification but including additional countries and years to the dataset, Easterly, Levine, and Roodman (2004) do not obtain the same result. Burnside and Dollar (2004) discuss the differences between the findings and conclude that they occur largely because of the additional countries, rather than lengthening the time series.

One widely debated topic in economics that has features of both replication and robustness exercises is the topic of minimum wage impacts on unemployment. In early work, Welch (1974) concluded that early minimum wage legislation decreased teenage employment, increased the cyclical nature of teenage employment with respect to the business cycle, and shifted teenage employment towards sectors not covered by the law. However, in the course of using Welch's data, Siskind (1977) discovered that Welch had used data for teenagers 16-19 years old instead of 14-19 years old for certain years, and once this was corrected, the minimum wage did not appear to reduce teenage employment. This was a fairly easy mistake to understand since the Current Population Survey was undergoing changes at the time, and table headings for unpublished data had not even been updated. Welch graciously acknowledged the error, and used the corrected data to extend the analysis to probe impacts by industry sector (Welch 1977).

Scholars working on this important topic have, for several decades now, continued to find significant room for disagreement on key issues of sampling, data sources, and statistical analysis methods²¹, matters on which well-intended researchers may well disagree. In this and other similarly contentious debates, we believe that the use of pre-specified research designs and analysis plans could be useful for advancing scientific progress, a point we return to below.

2.4.3 Fraud and Retractions

Though we believe (or at least, would prefer to believe) that most instances in which economics studies cannot be replicated are due to inadvertent human error or analytical judgment calls, fraud cannot be completely discounted in all cases.

Popular books such as Broad and Wade's *Betrayers of the Truth* (1983) make it clear that scientists are not always saints. A survey of 234 economists at the 1998 ASSA/AEA meeting

²¹ See, for instance, Card and Krueger (1994), Neumark and Wascher (2000), and Card and Krueger (2000) the latter two of which extend the analysis by using new datasets with the original specifications, as well as new econometric specifications. The Pennsylvania/New Jersey comparison from these papers was extended to the set of all cross-state minimum wage differences in Dube, Lester, and Reich (2010), and Neumark, Salas, and Wascher (2014).

investigated falsification of research, inappropriate inclusion or omission of co-authors, and exchange of grades for gifts, money, or sexual favors (List et al. 2001). Both a randomization coin-toss technique to elicit true responses to sensitive questions, as well as a more standard question design, indicate that 4% of respondents admit to having at some time falsified research data, 7-10% of respondents admit to having committed one of four relatively minor research infractions, while up to 0.4% admitted to exchange of grades for gifts, money, or sexual favors. Given the seriousness of some of these offenses, an obvious concern is that these figures understate the actual incidence of fraudulent research practices.

A more recent survey of members of the European Economics Association described in Necker (2014) asks individuals about the justifiability of certain practices as well as their behavior regarding those practices. Necker shows that 2.6% of researchers admit to having falsified data, while 94% admit to at least one instance of a practice considered inappropriate by the majority of the survey, and there is a clear positive correlation between justifiability and behavior, as well as between perceived professional publication pressures and questionable research practices.

Similar surveys in other fields such as Anderson, Martinson, and Vries (2007), which surveyed researchers across disciplines funded by the U.S. National Institutes of Health, and John, Loewenstein, and Prelec (2012) in psychology, as well as a meta-analysis of 18 surveys of academic misbehavior, do not paint a very rosy picture, with 2% of respondents admitting to data fabrication, and 34% admitting to lesser forms of academic misconduct (Fanelli 2009).

We are not aware of a recent case in economics that received media attention similar to the Michael Lacour fraud scandal uncovered by Broockman, Kalla, and Aranow (2015) in political science, or the case of Diedrick Stapel (see Carey 2011; Bhattacharjee 2013) in psychology. However, there is considerable evidence of plagiarism and other forms of research malpractice in economics. This *Journal* itself published the results of a survey sent to 470 economics journal editors, which revealed significant problems (Enders and Hoover 2004). Among the 127 editors who responded, only 19% claimed that their journal had a formal policy on plagiarism, and 42 cases of plagiarism were discovered in an average year, with nearly 24% of editors encountering at least one case. A follow-up survey of rank-and-file economists

revealed a general lack of consensus on how to respond to cases of alleged plagiarism (Enders and Hoover 2006).²²

Article retraction is another useful indicator of research misconduct. A search of four popular article databases for terms related to article retractions identified by Karabag and Berggren (2012) found six retractions: (“Retraction Statement and Authors’ Apology” 2009; Berger 2009; Nofsinger 2009; “Statement of Retraction” 2010; “Redundant Publishing – Australasian Journal of Regional Studies” 2011; “Statement of Retraction” 2012) which all occurred in the last few years. The volunteer network Research Papers in Economics (RePEc) maintains a plagiarism committee, which as of August 2016, had documented 52 cases of plagiarism, 12 cases of self-plagiarism, and four cases of fraud involving 96 authors.²³

Some institutional journal policies in economics lag behind those of other disciplines. For instance, as documented by Karabag and Berggren (2012), many economics and business journals appear not to even have explicit policies regarding ethics, plagiarism, or retraction,²⁴ and in many cases articles that have been retracted continue to be available on the journal’s website without any indication that it has been retracted. For example, though Gerking and Morgan (2007) features “Retraction” in the title, the relevant earlier paper (Kunce, Gerking, and Morgan 2002) is still available and appears unchanged. If one happened to discover the webpage of the original²⁵ first (note that the original appears first in Google Scholar searches), one would have no reason to suspect that it had been retracted. For comparison, the webpage²⁶ for Maringer and Stapel (2009), which was retracted in 2015²⁷, clearly reads “THIS PAPER HAS BEEN RETRACTED,” the title has been altered to begin with “Retracted: ,” and the pdf features an obvious RETRACTED watermark on every page. This is also the case with all six of the

²² Well-known plagiarism cases involve an article published in 1984 in the *Quarterly Journal of Economics* (see Chenault 1984; “Notice to Our Readers” 1984) and a case of plagiarism of an original article from *Economics Innovation and New Technology* for re-publication in *Kyklos* (Frey, Frey, and Eichenberger 1999). The most recent incident that seemed to attract significant attention was the submission of a substantively identical article to multiple journals within economics, which is also a serious lapse (“Correspondence: David H Autor and Bruno S. Frey” 2011). Even if plagiarism of this manner would seem significantly easier to catch in the Internet age, the proliferation of journals partially counteracts this ease.

²³ <https://plagiarism.repec.org/index.html>

²⁴ Although note that journals may present these policies online as opposed to formally publishing them in the journal; for instance, see the *Quarterly Journal of Economics*’ formal ethics policy: http://www.oxfordjournals.org/our_journals/qje/for_authors/journal_policies.html.

²⁵ <https://www.aeaweb.org/articles.php?doi=10.1257/000282802762024656>, accessed October 10, 2016.

²⁶ <http://onlinelibrary.wiley.com/doi/10.1002/ejsp.569/abstract>, accessed October 10, 2016.

²⁷ See “Retraction Statement: ‘Correction or Comparison? The Effects of Prime Awareness on Social Judgments’, by M. Maringer and D. Stapel” (2015).

retractions in Karabag and Berggren (2012), as well as other notable recent retractions such as LaCour and Green (2014), which was retracted by *Science* (see McNutt 2015).

The bottom line is that there is little reason to believe that economists are inherently more ethical than other social scientists or researchers in other disciplines, so policies regarding fraud and retraction from other disciplines might potentially be beneficially applied to economics.

3 New research methods and tools

This section discusses several new methods and tools that have emerged in economics research over the past two decades – and more forcefully over the past ten years – to address the concerns discussed in Section 2. These approaches have in common a focus on greater transparency and openness in the research process. They include improved research design (including experimental designs and meta-analysis approaches), study registration and pre-analysis plans, strengthened disclosure and reporting practices, and new norms regarding open data and materials. We discuss each in turn.

3.1 Improved analytical methods: research designs and meta-analysis

There have been a number of different responses within economics to the view that pervasive specification searching and publication bias was affecting the credibility of empirical literatures. As mentioned above, there has been a shift towards a greater focus on prospective research design in several fields of applied economics work. Experimental (Duflo, Glennerster, and Kremer 2007) and quasi-experimental (Angrist and Pischke 2010) research designs arguably place more constraints on researchers relative to earlier empirical approaches, since there are natural ways to present data using these designs that researchers are typically compelled to present by colleagues in seminars and by journal referees and editors. Prospective experimental studies also tend to place greater emphasis on adequately powering an analysis statistically, which may help to reduce the likelihood of publishing only false positives (Duflo, Glennerster, and Kremer 2007).

There is also suggestive evidence that the adoption of experimental and quasi-experimental empirical approaches is beginning to address some concerns about specification search and publication bias: Brodeur et al. (2016) present tentative evidence that the familiar spike in p-values just below the 0.05 level is less pronounced in randomized control trial studies

than in studies utilizing non-experimental methods. Yet improved research design alone may not solve several other key threats to the credibility of empirical economics research, including the possibility that null or “uninteresting” findings never become known within the research community.

3.1.1 Understanding Statistical Model Uncertainty

In addition to improvements in research design, Leamer (1983) argued for greater disclosure of the decisions made in analysis, in what became known as “extreme bounds analysis” (described in Section 2). Research along these lines has dealt with model uncertainty by employing combinations of multiple models and specifications, as well as comparisons between them. Leamer himself has continued to advance this agenda (see Leamer 2016). We describe several related approaches here.

3.1.1.1 Model Averaging

A natural way to deal with statistical model uncertainty is through Bayesian model averaging. This approach has stronger decision-theoretic foundations than extreme bounds analysis (see Brock, Durlauf, and West 2003). In this approach, each model in the space of plausible models is assigned a probability of being true based on researcher priors and goodness of fit criteria. Averaging the resulting estimates generates a statistic incorporating model uncertainty:

(eqn. 3)
$$\hat{\delta}_M = \sum_m \mu(m | D) \hat{\delta}_m ,$$

where m refers to a particular statistical model, M is the space of plausible models, $\mu(m | D)$ is the posterior probability of a model being the true model given the data D , and $\hat{\delta}_m$ is the estimated statistic from model $m \in M$.

These weights must, of course, be chosen somehow. Cohen-Cole et al. (2009), from whom we borrow the above notation, study the deterrent effect of the death penalty with a model averaging exercise combining evidence from Donohue and Wolfers (2005) and Dezhbakhsh, Rubin, and Shepherd (2003) and use the Bayesian Information Criterion (BIC) (Schwarz 1978). The weighted average they generate implies a large but imprecisely estimated deterrent effect of executions on homicides in the U.S. Of course, even without employing explicit probability weights, simply visualizing the distribution of estimates across the entire space of statistical models can also be quite informative on its own.

Two well-cited examples of model averaging engage in a thorough investigation of the determinants of cross-country economic growth. Sala-i-Martin's (1997) famous "I Just Ran Two Million Regressions" article uses model weights proportional to the integrated likelihoods of each model, picks all possible three-variable combinations out of 60 covariates that have been reported as being significantly related to economic growth, and finds that only about one-third of the 60 variables can be considered robustly positively correlated with economic growth across models. Sala-i-Martin, Doppelhofer, and Miller (2004) conduct what they call Bayesian Averaging of Classical Estimates (BACE), weighting estimates using an approach analogous to Schwarz's BIC, and find that just 18 of 67 variables are significantly and robustly partially correlated with economic growth, once suggesting that many findings reported in the existing empirical literature may be spuriously generated by specification searching and selective reporting; see also Fernández, Ley, and Steel (2001) for a related exercise.

3.1.1.2 The LSE School, Data Mining, and Machine Learning

While specification searching or data mining often has a negative connotation in applied economic research, some scholars have taken a more favorable view of it, as long as the data mining is carried out appropriately (Pagan 1987; Phillips 1988). Advocates of this approach have sometimes been known as the "LSE school" of econometrics, or the general-to-specific modelling approach (Gilbert 1989; Hendry 1987; Hendry 1995). This approach is related in spirit to the idea of "encompassing," which is the principle that one statistical model can account for, or explain, another (Mizon and Richard 1986; Bontemps and Mizon 2008). To our knowledge, this approach is more often applied in time series econometrics than in applied microeconomics, but perhaps a closer consideration is warranted. There have also been recent calls within applied economics for greater application of machine learning methods and other data science techniques that share some features with these approaches (Kleinberg, Ludwig, et al. 2015; Kleinberg, Lakkaraju, et al. 2015). For a wide set of views on data mining more broadly and the LSE approach specifically, see the *Journal of Economic Methodology*, which devoted a special issue to the topic (Backhouse and Morgan 2000).

3.1.1.3 Specification Curve

Simonsohn, Simmons, and Nelson (2015b) propose a method, which they call the "specification curve," that is similar in spirit to Leamer's extreme-bounds analysis, but recommends researchers test the exhaustive combination of analytical decisions, not just decisions about

which covariates to include in the model. If the full exhaustive set is too large to be practical, a random subset can be used. After plotting the effect size from each of the specifications, researchers can assess how much the estimated effect size varies, and which combinations of decisions lead to which outcomes. Using permutation tests (for treatment with random assignment) or bootstrapping (for treatment without random assignment), researchers can generate shuffled samples with no true effect by construction, and compare the specification curves from these placebo samples to the specification curve from the actual data. Many comparisons are possible, but the authors suggest comparing the median effect size, the share of results with predicted sign, and share of statistically significant results with predicted sign. A key comparison, which is analogous to the traditional p-value, is the percent of the shuffled samples with as many or more extreme results.

The paper builds specification curves for two examples: Jung et al. (2014) which tested the effect of the gender of hurricane names on human fatalities, and Bertrand and Mullainathan (2004) which tested job application callback rates based on the likely ethnicity of applicant names included in job resumes. Jung et al. (2014) elicited four critical responses taking issue with the analytical decisions (Christensen and Christensen 2014; Maley 2014; Malter 2014; Bakkensen and Larson 2014). The specification curve shows that 46% of curves from permuted data show at least as large a median effect size as the original, 16% show at least as many results with the predicted sign, and 85% show at least as many significant results with the predicted sign. This indicates that the results are likely to have been generated by chance. The Bertrand and Mullainathan (2004) specification curve, on the other hand, shows that fewer than 0.2% of the permuted curves generate as large a median effect, 12.5% of permuted curves show at least as many results with the predicted sign, and less than 0.2% of permuted curves show at least as many significant results with the predicted sign, providing evidence that the results are very unlikely to have been generated by chance.

3.1.2 Improved Publication Bias Tests

There have been significant advances in the methodological literature on quantifying the extent of publication bias in a given body of literature. Early methods mentioned above include Rosenthal's (1979) method (the “fail-safe N”), while Galbraith (1988) advocated for radial plots of log odds ratios, and Card and Krueger (1995) tested for relationships between study sample sizes and *t*-statistics.

Statisticians have developed methods to estimate effect sizes in meta-analyses that control for publication bias (Hedges 1992; Hedges and Vevea 1996). The tools most widely used by economists tend to be simpler, including the widely used funnel plot, which is a scatter plot of some measure of statistical precision (typically the inverse of the standard error), versus the estimated effect size. Estimates generated from smaller samples should usually form the wider base of an inverted funnel, which should be symmetric around more precise estimates in the absence of publication bias. The method is illustrated with several economics examples in Stanley and Doucouliagos (2010), and two of these are reproduced in Figure 4. In addition to scrutinizing the visual plot, a formal test of the symmetry of this plot can be conducted using data from multiple studies and regressing the relevant t-statistics on inverse standard errors:

(eqn. 4)
$$t_i = \frac{\text{Estimated effect}_i}{SE_i} = \beta_0 + \beta_1 \left(\frac{1}{SE_i}\right) + v_i.$$

The resulting t-test on β_0 , referred to as the Funnel Asymmetry Test (FAT) (Stanley 2008), captures the correlation between estimated effect size and precision, and thus tests for publication bias. This analysis assumes that study sample size is uncorrelated with other research design features, an assumption that is debatable (Simonsohn 2017).

Using the FAT, Doucouliagos and Stanley (2009) find evidence of publication bias in Card and Krueger's (1995) sample of minimum wage studies ($\beta_0 \neq 0$), consistent with their own interpretation of the published literature at that time. β_1 here can also be interpreted as the true effect (called the precision effect test, PET) free of publication bias, and Doucouliagos and Stanley (2009) find no evidence of a true effect of the minimum wage on unemployment. The authors also conduct the FAT-PET tests with 49 additional more recent studies in this literature and find the same results: evidence of significant publication bias and no evidence of an effect of the minimum wage on unemployment. Additional meta-analysis methods, including this “FAT-PET” approach, are summarized in Stanley and Doucouliagos (2012).

3.1.3 Multiple Testing Corrections

Other applied econometricians have recently called for increasing the use of multiple testing corrections in order to generate more meaningful inference in study settings with many research hypotheses (Anderson 2008; Fink, McConnell, and Vollmer 2014). The practice of correcting for multiple tests is already widespread in certain scientific fields (e.g., genetics) but has yet to

become the norm in economics and other social sciences. Simply put, since we know that p-values fall below traditional significance thresholds (e.g., 0.05) purely by chance a certain proportion of the time, it makes sense to report adjusted p-values that account for the fact that we are running multiple tests, since this makes it more likely that at least one of our test statistics has a significant p-value simply by chance.

There are several multiple testing approaches, some of which are used and explained by Anderson (2008), namely, reporting index tests, controlling the family-wise error rate (FWER), and controlling the false discovery rate (FDR). These are each discussed in turn below.

3.1.3.1 Reporting Index Tests

One option for scholars in cases where there are multiple related outcome measures is to forego reporting the outcomes of numerous tests, and instead standardize the related outcomes and combine them into a smaller number of indices, sometimes referred to as a mean effect. This can be implemented for a family of related outcomes by making all signs agree (i.e., allowing positive values to denote beneficial outcomes), demeaning and dividing by the control group standard deviation, and constructing a weighted average (possibly using the inverse of the covariance matrix to weight each standardized outcome). This new index can be used as a single outcome in a regression model and evaluated with a standard t test. Kling, Liebman, and Katz (2007) implement an early index test in the Moving to Opportunity field experiment using methods developed in biomedicine by O'Brien (1984).

This method addresses some concerns regarding the multiplicity of statistical tests by simply reducing the number of tests. A potential drawback is that the index may combine outcomes that are only weakly related, and may obscure impacts on specific outcomes that are of interest to particular scholars, although note that these specific outcomes could also be separately reported for completeness.

3.1.3.2 Controlling the Family-Wise Error Rate

The family-wise error rate (FWER) is the probability that at least one true hypothesis in a group is rejected (a type I error, or false positive). This approach is considered most useful when the “damage” from incorrectly claiming *any* hypothesis is false is high. There are several ways to implement this approach, with the simplest method being the Bonferroni correction of simply

multiplying every original p-value by the number of tests carried out (Bland and Altman 1995), although this is extremely conservative, and improved methods have also been developed.

Holm's sequential method involves ordering p-values by class and multiplying the lower p-values by higher discount factors (Holm 1979). A related and more efficient recent method is the free step-down resampling method, developed by Westfall and Young (1993), which when implemented by Anderson (2008) implies that several highly cited experimental pre-school interventions (namely, the Abecedarian, Perry, and Early Training Project studies) exhibit few positive long-run impacts for males.

Another recent method improves on Holm by incorporating the dependent structure of multiple tests. Lee and Shaikh (2014) apply it to reevaluate the Mexican PROGRESA conditional cash transfer program and find that overall program impacts remain positive and significant, but are statistically significant for fewer subgroups (e.g., by gender, education) when controlling for multiple testing. List, Shaikh, and Xu (2016) propose a method of controlling the FWER for three common situations in experimental economics, namely, testing multiple outcomes, testing for heterogeneous treatment effects in multiple subgroups, and testing with multiple treatment conditions.²⁸

3.1.3.3 Controlling the False Discovery Rate

In situations where a single type I error is not considered very costly, researchers may be willing to use a somewhat less conservative method than the FWER approached discussed above, and trade off some incorrect hypothesis rejections in exchange for greater statistical power. This is made possible by controlling the false discovery rate (FDR), or the percentage of rejections that are type I errors. Benjamini and Hochberg (1995) details a simple algorithm to control this rate at a chosen level under the assumption that the p-values from the multiple tests are independent, though the same method was later shown to also be valid under weaker assumptions (Benjamini and Yekutieli 2001). Benjamini, Krieger, and Yekutieli (2006) describes a two-step procedure with greater statistical power, while Romano, Shaikh, and Wolf (2008) propose the first methods to incorporate information about the dependence structure of the test statistics.

Multiple hypothesis testing adjustments have recently been used in finance (Harvey, Liu, and Zhu 2015) to re-evaluate 316 factors from 313 different papers that explain the cross-section

²⁸ Most methods are meant only to deal with the first and/or second of these cases. Statistical code to implement the adjustments in List, Shaikh, and Xu (2016) in Stata and Matlab is available at: <https://github.com/seidelj/mht>.

of expected stock returns. The authors employ the Bonferroni; Holm (1979); and Benjamini, Hochberg, and Yekutieli (2006) methods to account for multiple testing, and conclude that t-statistics greater than 3.0, and possibly as high as 3.9, should be used instead of the standard 1.96, to actually conclude that a factor explains stock returns with 95% confidence. Index tests and both the FWER and FDR multiple testing corrections are also employed in Casey, Glennerster, and Miguel (2012) to estimate the impacts of a community driven development program in Sierra Leone using a dataset with hundreds of potentially relevant outcome variables.

3.1.3.4 Bootstrap Reality Check

Another method that controls for, or acts as a reality check, for data snooping or data mining was developed in White (2000). The testing of multiple hypotheses, or repeated use of the same data, is a particularly central problem with time series data used over and over again by multiple scholars, such as data on stock returns, which makes this research quite important in empirical finance. Like the model averaging approach described above, the reality check requires a researcher to estimate the entire space of plausible models, but now compares the performance of the preferred model to a benchmark model (e.g., a model for stock market predictions based on the efficient market hypothesis), and does so repeatedly with bootstrapped samples.

To assess whether a certain preferred model actually outperforms the benchmark after accounting for snooping with multiple models, the econometrician first calculates the performance of the preferred model (using mean squared error improvement over the benchmark, or relative profit of the strategy). She then selects a bootstrap sample (with replacement), and calculates the mean squared error improvement (or profit) with the new sample for all of the different plausible statistical models, recording the best mean squared error improvement (or profit) across all the models. This approach can then be repeated 1,000 (or more) times, gathering the 1,000 best mean squared errors (or profits). In the final step, one must compare the original preferred model's mean squared error to the best performance from each of the 1,000 bootstraps. The p-value is the fraction of bootstrapped best fits that outperform the preferred model. (A truly predictive model would have returns higher than 95% of the best performing models from each of the bootstrapped samples.)

This method was implemented on a large number of trading rules in Sullivan, Timmermann, and White (1999) and a similar method that addresses the presence of poorly performing or irrelevant alternatives was developed in Hansen (2005).

3.2 Study Registration

A leading proposed solution to the problem of publication bias is the registration of empirical studies in a public registry. This would ideally be a centralized database of all attempts to conduct research on a certain question, irrespective of the nature of the results, and such that even null (not statistically significant) findings are not lost to the research community. Top medical journals have adopted a clear standard of publishing only medical trials that are registered (De Angelis et al. 2004). The largest clinical trial registry is clinicaltrials.gov, which helped to inspire the most high-profile study registry within economics, the AEA Randomized Controlled Trial Registry (Katz et al. 2013), which was launched in May 2013.²⁹

While recent research in medicine finds that the clinical trial registry has not eliminated all under-reporting of null results or other forms of publication bias and specification searching (Laine et al. 2007; Mathieu et al. 2009), they do allow the research community to quantify the extent of these problems and over time may help to constrain inappropriate practices. It also helps scholars locate studies that are delayed in publication, or are never published, helping to fill in gaps in the literature and thus resolving some of the problems identified in Franco, Malhotra, and Simonovits (2014).

Though it is too soon after the adoption of the AEA's trial registry to measure its impact on research practices and the robustness of empirical results, it is worth noting that the registry is already being used by many empirical researchers: since inception in 2013, over 850 studies conducted in over 100 countries have been registered, and the pace of registrations continues to rise rapidly. Figure 5, Panel A presents the total number of registrations over time in the AEA registry (through September 2016), and Panel B shows the number of new registrations per month. A review of the projects currently included in the registry suggests that there are a particularly large number of development economics studies, which is perhaps not surprising given the widespread use of field experimental methods in contemporary development economics.

In addition to the AEA registry, several other social science registries have recently been created, including by the International Initiative for Impact Evaluation's (3ie) Registry for International Development Impact Evaluations (RIDIE, <http://ridie.3ieimpact.org>), launched in September 2013 (Dahl Rasmussen, Malchow-Møller, and Barnebeck Andersen 2011), and the

²⁹ The registry can be found online at: <https://www.socialscienceregistry.org/>.

Experiments in Governance and Politics (EGAP) registry (<http://egap.org/content/registration>), also created in 2013. The Center for Open Science's Open Science Framework (OSF, <http://osf.io>) accommodates the registration of essentially any study or research document by allowing users to create a frozen time-stamped web URL with associated digital object identifier (DOI) for any materials uploaded to OSF. Several popular data storage options (including Dropbox, Dataverse, and GitHub) can also be synced with the OSF and its storage, creating a flexible way for researchers to register their research and materials. As of October 2016, over 7,300 public registrations have been created on OSF since the service launched in 2013.

3.3 Pre-Analysis Plans

In addition to serving as a useful way to search for research findings on a particular topic, most supporters of study registration also promote the pre-registration of studies, including pre-analysis plans (PAPs) that can be posted and time stamped even before analysis data are collected or otherwise available (Miguel et al. 2014). Registration is now the norm in medical research for randomized trials, and registrations often include (or link to) prospective statistical analysis plans as part of the project protocol. Official guidance from the U.S. Food and Drug Administration's Center for Drug Evaluation and Research (CDER) from 1998 describes what should be included in a statistical analysis plan, and discusses eight broad categories: pre-specification of the analysis; analysis sets; missing values and outliers; data transformation; estimation, confidence intervals, and hypothesis testing; adjustment of significance and confidence levels; subgroups, interactions, and covariates; and integrity of data and computer software validity (Food and Drug Administration, 1998).

While there were scattered early cases of pre-analysis plans being used in economics (most notably by Neumark 2001), the quantity of published papers employing pre-specified analysis has grown rapidly in the past few years, mirroring the rise of studies posted on the AEA registry.

There is ongoing discussion of what one should include in a PAP; detailed discussions include Glennerster and Takavarasha (2013), David McKenzie's World Bank Research Group blog post,³⁰ and a template for pre-analysis plans by Alejandro Ganimian (2014). Ganimian's template may be particularly useful to researchers themselves when developing their own pre-analysis plans, and instructors may find it useful in their courses. Building on, and modifying,

³⁰ <http://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist>

the FDA’s 1998 checklist with insights from these other recent treatments of pre-analysis plans, there appears to be a growing consensus that pre-analysis plans in economics should consider discussing at least the following list of ten issues:

1. Study Design
2. Study Sample
3. Outcome Measures
4. Mean Effects Family Groupings
5. Multiple Hypothesis Testing Adjustments
6. Subgroup Analyses
7. Direction of Effect for One-Tailed Tests
8. Statistical Specification and Method
9. Structural Model
10. Timestamp for Verification

Pre-analysis plans are relatively new to economics, and this list is likely to evolve in the coming years as researchers explore the potential, and possible limitations, of this new tool.

For those concerned about the possibility of “scooping” of new research designs and questions based upon a publicly posted pre-analysis plan or project description, several of the social science registries allow temporary embargoing of project details. For instance, the AEA registry allows an embargo until a specific date or project completion. At the time of writing, the OSF allows a four-year embargo until the information is made public.³¹

3.3.1 Examples of Pre-analysis Plans (PAPs)

Recent examples of economics papers based on experiments with PAPs include Casey, Glennerster, and Miguel (2012) and Finkelstein et al. (2012), among others. Casey, Glennerster, and Miguel (2012) discuss evidence from a large-scale field experiment on community driven development (CDD) projects in Sierra Leone. The project, called GoBifo, was intended to make local institutions in post-war Sierra Leone more democratic and egalitarian. GoBifo funds were spent on a variety of local public goods infrastructure (e.g., community centers, schools, latrines, roads), agriculture, and business training projects, and were closely monitored to limit leakage. The analysis finds significant short-run benefits in terms of the “hardware” aspects of infrastructure and economic wellbeing: the latrines were indeed built. However, a larger goal of the project, reshaping local institutions, making them more egalitarian, increasing trust, improving local collective action, and strengthening community groups, which the researchers

³¹ See <http://help.osf.io/m/registrations/l/524207-embargoes>. Accessed October 10, 2016.

call the “software effects,” largely failed. There are a large number of plausible outcome measures along these dimensions, hundreds in total, which the authors analyze using a mean effects index approach for nine different families of outcomes (with multiple testing adjustments). The null hypothesis of no impact cannot be rejected at 95% confidence for any of the nine families of outcomes.

Yet Casey et al (2012) go on to show that, given the large numbers of outcomes in their dataset, and the multiplicity of ways to define outcome measures, finding some statistically significant results would have been relatively easy”. In fact, the paper includes an example of how, if they had had the latitude to define outcomes without a pre-analysis plan, as has been standard practice in most empirical economics studies (and in other social science fields), the authors could have reported either statistically significant and positive effects, or significantly negative effects, depending on the nature of the “cherry-picking” of results. We reproduce their results here as Table 4, where Panel A presents the statistically significant positive impacts identified in the GoBifo data and Panel B highlights negative effects. This finding begs the question: how many empirical economics papers with statistically significant results are, unbeknownst to us, really just some version of either Panel A or Panel B?

Finkelstein et al. (2012) study the politically charged question of the impacts of health insurance expansion, using the case of Oregon’s Medicaid program, called Oregon Health Plan (OHP). In 2008, Oregon determined it could afford to enroll 10,000 additional adults, and it opted to do so by random lottery. Most of the analyses in the impact evaluation were laid out in a detailed pre-analysis plan, which was publicly posted on the National Bureau of Economic Research’s website in 2010, before the researchers had access to the data.

This is important because, as in Casey et al (2012), the researchers tested a large number of outcomes: hospital admissions through the emergency room (ER) and not through the ER; hospital days; procedures; financial strain (bankruptcy, judgments, liens, delinquency, medical debt, and non-medical debt, measured by credit report data); self-reported health from survey data, and so on. When running such a large number of tests, the researchers again could have discovered some “significant” effects simply by chance. The pre-analysis plan, in conjunction with multiple hypothesis testing adjustments, give us more confidence in the main results of the study: that recipients did not improve significantly in terms of physical health measurements, but

they were more likely to have health insurance, had better self-reported health outcomes, utilized emergency rooms more, and had better detection and management of diabetes.

Additional studies that have resulted from the experiment have also employed pre-analysis plans, and they show that health insurance increased emergency department use (Taubman et al. 2014), had no effect on measured physical health outcomes after two years, but did increase health care use and diabetes management, as well as leading to lower rates of depression and financial strain (Baicker et al. 2013). The health care expansion had no significant effect on employment or earnings (Baicker et al. 2014).

Other prominent early examples of economics studies that have employed pre-analysis plans include poverty targeting programs in Indonesia, an evaluation of the Toms shoe company donation program, and a job training program in Turkey, among many others (Olken, Onishi, and Wong 2012; Alatas et al. 2012; Wydick, Katz, and Janet 2014; Hirshleifer et al. 2015). The PAP tool is also spreading to other social sciences beyond economics. For instance, in psychology, a pre-specified replication of an earlier paper that had found a link between female conception risk and racial prejudice failed to find a similar effect (Hawkins, Fitzgerald, and Nosek 2015).

One issue that arises for studies that did register a pre-analysis plan is the question of characterizing the extent to which the analysis conforms to the original plan, or if it deviates in important ways from the plan. To appreciate these differences, scholars will need to compare the analysis to the plan, a step that could be seen as adding to the burden of journal editors and referees. Even if the analysis does conform exactly to the PAP, there is still the possibility that authors are consciously or unconsciously emphasizing a subset of the pre-specified analyses in the final study. Berge et al. (2015) develop an approach to comparing the distribution of p-values in the paper's main tables versus those in the PAP in order to quantify the extent of possibly selective reporting between the plan and the paper.

The Finkelstein et al (2012) study is a model of transparency regarding the presentation of results. To the authors' credit, all analyses presented in the published paper that were not pre-specified are clearly labeled as such; in fact, the exact phrase "This analysis was not prespecified." appears in the paper six times. Tables in the main text and appendix that report analyses that were not pre-specified are labelled with a "^" character to set them apart, and are clearly labeled as such.

3.3.2 Strengths, Limitations, and Other Issues Regarding Pre-Analysis Plans

There remain many open questions about whether, when, and how pre-analysis plans could and should be used in economics research, with open debates about how useful they are in different subfields of the discipline. Olken (2015), for example, highlights both their “promises and perils.” On the positive side, pre-analysis plans bind the hands of researchers and greatly limit specification searching, allowing them to take full advantage of the power of their statistical tests (even making one-sided tests reasonable).

A further advantage of the use of pre-analysis plans is that they are likely to help shield researchers from pressures to affirm the policy agenda of donors and policymakers, in cases where they have a vested interest in the outcome, or when research focuses on politically controversial topics (such as health care reform). This is especially the case if researchers and their institutional partners can agree on the pre-analysis plan, as a sort of evaluation contract.

On the negative side, PAPs are often complex and take valuable time to write. Scientific breakthroughs often come at unexpected times and places, often as a result of exploratory analysis, and the time spent writing PAPs may thus lead less time to spend on less structured data exploration.

Coffman and Niederle (2015) argue that there is limited upside from PAPs when replication (in conjunction with hypothesis registries) is possible. In experimental and behavioral economics, where lab experiments utilize samples of locally recruited students and the costs of replicating an experiment are relatively low, they argue that replication could be a viable substitute for pre-analysis plans. Yet there does appear to be a growing consensus, endorsed by Coffman and Niederle, that pre-analysis plans can significantly increase the credibility of reporting and analysis in large-scale randomized trials that are expensive or difficult to repeat, or when a study that relies on a particular contextual factor makes it impossible to replicate. For instance, Berge et al. (2015) carry out a series of lab experiments timed to take place just before the 2013 Kenya elections. Replication of this lab research is clearly impossible due to the unique context, and thus use of a pre-analysis plan is valuable.

Olken (2015) as well as Coffman and Niederle (2015) discuss another potential way to address publication bias and specification search: results-blind review. Scholars in psychology have championed this method; studies that are submitted to such review are often referred to as “registered reports” in that discipline. Authors write a detailed study protocol and pre-analysis

plan, and before the experiment is actually run and data is collected, submit the plan to a journal. Journals review the plan for the quality of the design and the scientific value of the research question, and may choose to give “in-principle acceptance”. This can be thought of as a kind of revise and resubmit that is contingent on the data being collected and analyzed as planned. If the author follows through on the proposed design, and the data are of sufficiently high quality (e.g., with sufficiently low sample attrition rates in a longitudinal study, etc.), the results are to be published regardless of whether or not they are statistically significant, and whether they conform to the expectations of the editor or referees, or to the conventional wisdom in the discipline.

Several psychology journals currently have begun using results-blind review, either regularly or in special issues (Chambers 2013; Chambers et al. 2014; Nosek and Lakens 2014).³² A forthcoming issue of *Comparative Political Studies* will be the first to feature results-blind review in political science (Findley et al. 2016), and it will include both experimental and observational research studies.

In our view, it would also be useful to experiment with results-blind review and registered reports in economics journals. The rise in experimental studies and pre-analysis plans in economics, as evidenced by the rapid growth of the AEA registry, is likely to facilitate the eventual acceptance of this approach.

3.3.3 Observational Studies

An important open question is how widely the approach of study registration and hypothesis pre-specification could be usefully applied in non-prospective and non-experimental studies.

This issue has been extensively discussed in recent years within medical research but consensus has not yet been reached in that community. It actually appears that some of the most prestigious medical research journals, which typically publish randomized trials, are even more in favor of the registration of observational studies than the editors of journals that publish primarily in non-experimental research (see the dueling editorial statements in *Epidemiology* 2010; *The Lancet* 2010; Loder, Groves, and MacAuley 2010; Dal-Ré et al. 2014).

A major logical concern with the pre-registration of non-prospective observational studies using pre-existing data is that there is often no credible way to verify that pre-registration

³² A list of journals that have adopted registered reports is available at: <https://osf.io/8mpji/wiki/home/>.

took place before analysis was completed, which is different than the case of prospective studies in which the data has not yet been collected or accessed. In our view, proponents of the pre-registration of observational work have not formulated a convincing response to this obvious concern.

The only economics study of which we are aware that has used a pre-analysis plan on non-experimental data was undertaken in Neumark (2001). Based on conversations with David Levine, Alan Krueger appears to have suggested to Levine, who was the editor of the *Industrial Relations* journal at the time, that multiple researchers could analyze the employment effects of an upcoming change in the federal minimum wage with pre-specified research designs, in a bid to eliminate “author effects,” and that this could create a productive “adversarial collaboration” between authors with starkly different prior views on the likely impacts of the policy change (Levine 2001). (The concept of adversarial collaboration—two sets of researchers with opposing theories coming together and agreeing on a way to test hypotheses before observing the data—is often associated with Daniel Kahneman, see, for example Bateman et al. 2005).

The U.S. federal minimum wage increased in October 1996 and September 1997. Although Krueger ultimately decided not to participate, Neumark submitted a pre-specified research design consisting of the exact estimating equations, variable definitions, and subgroups that would be used to analyze the effect of the minimum wage on the unemployment of younger workers using October, November, and December Current Population Survey (CPS) data from 1995 through 1998. This detailed plan was submitted to journal editors and reviewers prior to the end of May 1997; the October 1996 data started to become available at the end of May 1997, and Neumark assures readers he had not looked at any published data at the state level prior to submitting his analysis plan.

The verifiable “time stamp” of the federal government’s release of data indeed makes this approach possible, but the situation also benefits from the depth and intensity of the minimum wage debate prior to this study. Neumark had an extensive literature to draw upon when choosing specific regression functional forms and subgroup analyses. He tests two definitions of the minimum wage, the ratio of the minimum wage to the average wage (common in Neumark’s previous work) as well as the fraction of workers who benefit from the newly raised minimum wage (used in David Card’s earlier work, Card 1992a and Card 1992b), and tests both models with and without controls for the employment rate of higher-skilled prime age adults (as

recommended by Deere, Murphy, and Welch 1995). The results mostly fail to reject the null hypothesis of no effect of the minimum wage increase: only 18 of the 80 specifications result in statistically significant decreases in employment (at the 90% confidence level), with estimated elasticities ranging from -0.14 to -0.3 for the significant estimates and others closer to zero.

A more recent study bases its analysis on Neumark's exact pre-specified tests estimate the effect of minimum wages in Canada and found larger unemployment effects, but they had access to the data before estimating their models and did not have an agreement with the journal, so the value of this "pre-specification" is perhaps less clear (Campolieti, Gunderson, and Riddell 2006). In political science, a pre-specified observational analysis measured the effect of the immigration stances of Republican representatives on their 2010 election outcomes (Monogan 2013).

It is difficult to see how a researcher could reach Neumark's level of pre-specified detail with a research question with which they were not already intimately familiar. It seems more likely that in a case where the researcher was less knowledgeable they might either pre-specify with an inadequate level of detail, or choose an inappropriate specification; this risk makes it important that researchers should not be punished for deviating from their pre-analysis plan in cases where the plan omits important details or contains errors, as argued in Casey et al (2012).

It seems likely to us that the majority of observational empirical work in economics will continue largely as is for the foreseeable future. However, for important, intensely debated, and well-defined questions, it would be desirable in our view for more prospective observational research to be conducted in a pre-specified fashion, following the example in Neumark (2001). Although pre-specification will not always be possible, the fact that large amounts of government data are released to the public on regular schedules, and that many policy changes are known to occur well in advance (such as in the case of the anticipated federal minimum wage changes discussed above, with similar arguments for future elections), will make it possible for the verifiable pre-specification of research analysis to be carried out in many settings.

3.3.3.1 Comparisons to Other Research Fields

Another frontier topic in this realm is the use of pre-specified algorithms, including machine learning approaches, rather than exact pre-analysis plans for prospective studies. For instance, the exact procedure to be used to determine which covariates should be included in order to

generate the most statistically precise estimates can be laid out in advance, even if those covariates are unknown (and unknowable) before the data has been collected. This approach has not yet been widely adopted in economics (to our knowledge), but has begun to be used in medical trials and biostatistics (van der Laan et al. 2007; Sinisi et al. 2007).

A proposal related to, but slightly different than, pre-analysis plans is Nobel Prize winning physicist Saul Perlmutter's suggestion for the social sciences to use "blind analysis" (MacCoun and Perlmutter 2015). In blind analysis, researchers add noise to the data while working with it and running the analysis, thus preventing them from knowing which way the results are turning out, and thus either consciously or unconsciously biasing their analysis, until the very end, when the noise is removed and the final results are produced. This technique is apparently quite common in experimental physics (Klein and Roodman 2005), but we are not aware of its use in economics or other social sciences.

Major differences are also beginning to emerge in the use of pre-analysis plans, and in the design and interpretation of experimental evidence more broadly, among economists versus scholars in other fields, especially health researchers, with a much greater role of theory in the design of economics experiments. Economists often design experiments to shed light on underlying theoretical mechanisms, to inform ongoing theoretical debates, and measure and estimate endogenous behavioral responses. These behavioral responses may shed light on broader issues beyond the experimental intervention at hand, and thus could contribute to greater external validity of the results. As a result, pre-analysis plans in economics are often very detailed, and make explicit reference to theoretical models. For example, Bai et al. (2015) pre-registered the theoretical microeconomic model and detailed structural econometric approach that they planned to apply to a study of commitment contracts in the Indian health sector.

This distinction between the types of studies carried out by medical researchers versus economists (including those working on health topics) has a number of important implications for assessing the reliability of evidence. One has to do with the quality standards and perceptions of the risk of bias in a particular design. For medical trialists accustomed to the CONSORT standards or other medical efficacy trial reporting guidelines (described below), studies that do not feature double-blinding, and thus run the risk of endogenous behavioral responses to the medical intervention, are considered less reliable than those studies that employ double-blinding (for a detailed discussion, see Eble, Boone, and Elbourne 2014). A double-blind study is one in

which neither the participants nor the experimenters know who is receiving a particular treatment. While a few studies conducted by economists do feature double-blinding (e.g., Thomas et al. 2003; Thomas et al. 2006), in nearly all settings blinding participants to their status is either logistically difficult (for instance, if government partners are unwilling to distribute placebo treatments to some of their population) or even impossible.

To illustrate, how would you provide a placebo treatment in a study investigating the impact of the distribution of cash transfers on household consumption patterns? Even in settings that might seem promising for placebo treatments, such as the community-level deworming treatments discussed in Miguel and Kremer (2004), blinding participants to their status is basically impossible: deworming generates side effects (mainly gastrointestinal distress) in roughly 10% of those who take the pills, so community members in a placebo community would quickly deduce that they were in fact not receiving real deworming drugs if there are few or no local cases of side effects.

As noted above, endogenous behavioral responses are often exactly what we economists (and other social scientists) set out to measure and estimate in our field experiments, as described in our pre-analysis plans, and thus are to be embraced rather than rejected as symptomatic of a “low quality” research design that is at “high risk of bias”. Taken together, it is clear to us that the experimental literature in economics (and increasingly in other social sciences such as political science) often has very different objectives than medical, public health and epidemiological research, and thus different research methodologies are often called for. Despite the value of learning from recent experience in biomedical research, and the inspiration that the experience of medical research has played to the rise of new experimental research methods in the social sciences, economists have not simply been able to import existing medical trial methods wholesale, but are developing new and tailored approaches to pre-registration, pre-analysis plans, reporting standards, and transparency more broadly.

3.4 Disclosure and reporting standards

Another approach to promoting transparency is to establish detailed standards for the disclosure of information regarding study design, data, and analysis. These could serve to limit at least some forms of data mining and specification searching, or at least might make them more apparent to the reader.

Detailed reporting standards have become widespread in medical research for both experimental and observational research. Most notably for clinical trials, the Consolidated Standards of Reporting Trials (CONSORT) was developed (Begg C et al. 1996). A before-and-after comparison showed improvement in some measures of study reliability (Moher D et al. 2001), and the standards have been twice revised (Moher, Schulz, and Altman 2001; Schulz et al. 2010) and since extended to at least ten specific types of research designs, interventions, or data. Among others, and possibly particularly relevant for some types of economics research, these include cluster randomized trials (Campbell, Elbourne, and Altman 2004; Campbell et al. 2012), non-pharmacological treatment interventions (Boutron et al. 2008), and patient-reported outcomes (Calvert et al. 2013). In addition to the requirement by the International Committee of Medical Journal Editors (ICMJE, a group comprised of editors of top medical journals such as the *British Medical Journal*, *The Lancet*, *JAMA*, etc.) that randomized trials be registered in a registry such as clinicaltrials.gov, it is now standard that these journals require authors to include a completed CONSORT checklist at the time of article submission³³.

Observational research in epidemiology is increasingly subject to its own set of guidelines, the so-called Strengthening the Reporting of Observational Studies in Epidemiology, or STROBE, standards (von Elm et al. 2007). In fact, developing reporting guidelines is a growth industry in medical research: at least 284 sets of guidelines have been developed for different types of health research. To deal with the proliferation of reporting standards, the Equator Network has been established to organize these guidelines and help researchers identify the most appropriate set of guidelines for their research.³⁴

There are obviously very strong, and well understood, norms regarding how to report empirical results in economics studies, but there are far fewer formal guidelines or reporting checklists than in medical research. One exception is the AEA policy, announced in January 2012³⁵, that its journals would require disclosure statements from authors regarding potential conflicts of interest. The AEA journals enforced the policy in July 2012, and the NBER working

³³ See for example <http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#two> and <http://jama.jamanetwork.com/public/instructionsForAuthors.aspx#ClinicalTrials>

³⁴ Equator: Enhancing the QUALity and Transparency Of health Research, see <http://www.equator-network.org/>

³⁵ See https://www.aeaweb.org/PDF_files/PR/AEA_Adopts_Extensions_to_Principles_for_Author_Disclosure_01-05-12.pdf.

paper series has since adopted a similar set of required disclosures³⁶. It appears the economics discipline may have been shamed into adopting these conflict of interest policies, at least in part, by the scathing Academy Award-winning documentary “Inside Job,” which argued that some leading economists with strong (and often undisclosed) ties to the financial services industry were at least somewhat complicit in promoting policy choices that contributed to the 2008 global financial crisis (Casselmann 2012).

Despite recent progress on conflict of interest disclosure, there has been less change within economics regarding other forms of disclosure or reporting guidelines. The only set of disclosure guideline specific to economics that we are aware of is the Consolidated Health Economic Evaluation Reporting Standards (CHEERS), although these appear to be more widely followed in health than in economics (Husereau et al. 2013). In this regard, there has been less movement within economics than in other social sciences, including political science, where a section of the American Political Science Association has developed guidelines for reporting of experimental research (Gerber et al. 2014). The American Political Science Association has formed committees that resulted in the Data Access and Research Transparency (DART) statement, which APSA adopted in both its Ethics Guide and Journal Editors’ Transparency Statement, with 27 journals choosing to enact data sharing, data citation, and analytical methods sharing standards starting January 15, 2016.³⁷

In psychology, researchers have created an extension of CONSORT for social and psychological interventions (CONSORT-SPI) (Montgomery et al. 2013; Grant et al. 2013). Others psychologists have proposed that an effective way to reform reporting and disclosure norms within their discipline is for referees to enforce desirable practices when reviewing articles (Simmons, Nelson, and Simonsohn 2011). These authors recommended six conditions for referees to consider include the following:

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.³⁸
3. Authors must list all variables collected in a study.

³⁶ See https://www.aeaweb.org/aea_journals/AEA_Disclosure_Policy.pdf and <http://www.nber.org/researchdisclosurepolicy.html>.

³⁷ See <http://www.dartstatement.org>.

³⁸ It is now widely acknowledged, including by the authors themselves, that 20 is typically far too few. More generally, this sort of ad hoc sample size guideline seems difficult to justify as a blanket rule across all settings.

4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

These disclosure rules are further simplified into a simple 21-word solution to be used by authors: “We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study” (Simmons, Nelson, and Simonsohn 2012). There is a corresponding statement to be used by reviewers: “I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science ([see http://osf.io/project/hadz3](http://osf.io/project/hadz3)). I include it in every review.”³⁹

Recently, we, the authors of this article, were part of an interdisciplinary group of researchers that developed a detailed set of journal guidelines called the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al. 2015). This modular set of guidelines for journals features eight categories, namely: citation standards, data transparency, analytic methods (code) transparency, research materials transparency, design and analysis transparency, preregistration of studies, preregistration of analysis plans, and replication – with four levels (0-3) of transparency that journals could choose to endorse or require. For example, with regards to data transparency, the level zero standard is that the journal either encourages data sharing or says nothing, while the level three standard is that “data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication”; levels 1 and 2 fall somewhere in between. Journals could choose to adopt higher standards in some categories than others, as they feel most appropriate for their research community.

In the six months after the guidelines were published in *Science*, 538 journals and 57 organizations across a wide variety of scientific disciplines, including many in the social sciences, expressed their support for the standards and agreed to evaluate them for potential adoption. *Science* has now announced that it will be implementing the standards, effective January 1, 2017 (M. McNutt 2016). However, none of the leading economics journals have yet

³⁹ See <http://centerforopenscience.github.io/osc/2013/12/09/reviewer-statement-initiative/>

chosen to endorse or implement the guidelines; we encourage economics journal editors to review the guidelines and seriously consider adopting high transparency and reproducibility standards for their journals, keeping in mind that the TOP standards are meant to be modular rather than one-size-fits-all.

One last issue is worth a brief mention. Another important dimension of research transparency related to disclosure has to do with the presentation of data and results in tables, figures and other display items. There is a flourishing literature on effective data visualization approaches, much of it inspired by the seminal work of political scientist Edward Tufte (2001). While beyond the scope of this survey article, we refer interested readers to Gelman, Pasarica, and Dodhia (2002) and Schwabish (2014) for detailed discussions.

3.4.1 Fraud and Retractions

Building on the discussion in section 2 above, it appears that the formulation of explicit economics journal standards for article retraction, and clearer communication on journal websites stating when an article is retracted, could also be beneficial. The RePEC tracking of offenses, mentioned above, is a helpful but only partial start.

There is mounting evidence from other research fields that could help inform the creation of new standards in economics. Evidence from article retractions catalogued in PubMed show that the rate of retractions in medical research is on the rise. Articles appear to be retracted sooner after publication, and it is not the case that fraud represents an increasing proportion of reasons for retractions (Steen 2010; Steen, Casadevall, and Fang 2013). With tracking of offenses, researchers can use the Retraction Index (simply the fraction of retracted articles per 1000 papers published in a journal) which Fang and Casadevall (2011) show to be positively correlated with journal impact factor.

Optimistically, perhaps, Fanelli (2013) argues that the evidence of an increasing rate of retractions points towards a stronger system, rather than an increasing rate of fraud. This claim is based on the facts that, though the rate of retractions is increasing, the rate of article corrections has not; that despite the increasing proportion of journals issuing retractions, the rate of retractions per-retracting journal has not increased; and that despite an increase in allegations made to the U.S. Office of Research Integrity, the rate of misconduct findings has not increased.

Researchers have also developed novel statistical tools that one can use to detect fraud, using the fact that humans tend to drastically underestimate how noisy real data is when they are

making up fraudulent data. Simonsohn (2013) used this forensic technique after observing summary statistics that were disturbingly similar across treatment arms to successfully combat fraud in psychology, resulting in the retraction of several papers by two prominent scholars.

Another potentially useful tool is post-publication peer review. Formalizing post-publication peer review puts us in relatively uncharted waters. Yet it is worth noting that all four of the AEA's *American Economic Journals* allow for comments to appear on every article's official webpage post-publication (anonymous comments are not allowed). The feature does not appear to be widely used, but in one case, Lundqvist, Dahlberg, and Mörk (2014), comments placed on the website have actually resulted in changes to the article between its initial online pre-publication and the final published version, suggesting that this could be a useful tool for the research community to improve the quality of published work in the future.⁴⁰

3.5 Open data and materials, and their use for replication

There has clearly been considerable progress on the sharing of the data and materials necessary for replication since the famous 1980's *Journal of Money, Credit, and Banking* project mentioned above. Today, all American Economic Association journals require sharing of data and code to at least make replication theoretically possible (Glandon 2010). However, many leading journals in economics only recently introduced similar requirements, most notably the *Quarterly Journal of Economics*, and even when a journal data sharing policies exist, they are rarely enforced in a serious way (McCullough, McGeary, and Harrison 2008; Anderson et al. 2008). Authors can share the bare minimal final dataset necessary to generate the tables in the paper—all merging, cleaning, and removal of outliers or observations with missing data already done. Stripping this dataset of any additional variables not used in the final analysis would meet journal sharing requirements, and is certainly a big step forward relative to sharing no data at all, but it does limit the usefulness of the dataset for other researchers hoping to probe the robustness of the published results, extend the analysis, or utilize the data for other purposes.

This means that in practice, we economists as a discipline are still in a situation in which replication attempts for most empirical studies are still relatively costly in terms of time and effort. Despite improved (if still imperfect) data availability, we also know of no mainstream

⁴⁰ <https://www.aeaweb.org/articles.php?doi=10.1257/pol.6.1.167>

journal in economics that systematically tests that submitted data and code actually produce the claimed results as a pre-condition of publication. An interesting new movement hoping to change this is the Peer Reviewer's Openness Initiative, whereby researchers can pledge that after a certain date (January 1, 2017) they will begin to require data sharing at the time of peer review in the articles they referee (Morey et al. 2015)⁴¹. If journal reviewers demand en masse to have access to the code and data that generated the results, and new norms develop around this expectation, this might lead to rapid changes in data sharing practices, given the central role that journal publication plays in scholars' individual professional success and standing.

As discussed above, the imprecise definition of the term "replication" itself often leads to confusion (Clemens 2015). Clarification of what authors mean when they say a replication "failed" (can the data not even produce the published results, are they not robust to additional specifications, or does a new sample or extended data set produce different results?) may be an important first step to mainstreaming replication research within economics.

Some economists have advocated for a *Journal of Replication* (and as many have called for a *Journal of Null Results*), including recently Coffman and Niederle (2015) and Zimmermann (2015), but the low status that would likely accompany these journals could limit submission rates and doom them to failure. In lieu of this, several alternative solutions have been proposed. Hamermesh (2007) urges top journals to commission a few replications per year from top researchers, on a paper of the authors' choice, with acceptance guaranteed but subject to peer review (not by the original author, though they would be allowed to respond).

In psychology, Nosek, Spies, and Motyl (2012) are also skeptical of creating new journals devoted to replications or null results, and instead suggest crowdsourcing replication efforts. This seems have to been extremely successful, with two large-scale replication efforts in which many researchers worked together to repeat classic experiments in psychology with new samples, the Many Labs project (Klein et al. 2014) and the Replication Project: Psychology (Open Science Collaboration 2012; Open Science Collaboration 2015). Both were published in prominent journals and widely covered in the popular media. A similar project in cancer biology is ongoing.⁴²

⁴¹ For more information, see <http://opennessinitiative.org>

⁴² <http://elifesciences.org/collections/reproducibility-project-cancer-biology>

The Many Labs project sought to reproduce 13 effects found in the literature, testing them in 36 samples with a total sample size of 6,344, and determining whether online samples produced different effects than lab samples, and also comparing international to U.S. samples. They find that two types of interventions failed to replicate entirely, while results for other replications relative to the original studies were more nuanced.

The Replication Project: Psychology (RPP) team repeated the experiments of 100 previous effects, finding that only 47% of the replications produced results in the original 95% confidence interval, and subjectively considered 39% of the original findings to have successfully been “reproduced.”

Some in psychology have taken issue with the claims of the RPP, most notably Gilbert et al. (2016), which argues that differences in implementation between original and replication experiments were inappropriate and introduces noise in addition to the expected sampling error. When taking this into account, one should actually expect the relatively low reported replication rate, and they thus argue there is no replication crisis. Some of the original RPP authors respond that differences between original and replication studies were in fact often endorsed by original study authors and take issue with the statistical analysis in Gilbert et al. (Anderson et al. 2016).

Simonsohn (2015) engages in further discussion of how one should evaluate replication results, suggesting that powering a replication based on the effect size of the original study is problematic, and to distinguish the effect size from zero, replications (at least in psychology, with their typically small sample and effect sizes) should have a sample at least 2.5 times as large as the original. An optimistic take by Patil, Peng, and Leek (2016) suggests that researchers should compare the effect in the replication study to a “prediction interval” defined as $\hat{r}_{orig} \pm z_{0.975} \sqrt{\frac{1}{n_{orig}-3} + \frac{1}{n_{rep}-3}}$ where \hat{r}_{orig} is the correlation estimate in the original study, n_{orig} and n_{rep} are the sample sizes in the original and replication studies, respectively; and $z_{0.975}$ is the 97.5% quantile of the normal distribution, which incorporates uncertainty in the estimates from both the original and replication study. Applying this approach leads to much higher estimates of study replication (75%) for the RPP.

Economists may be interested to know that the researchers behind the RPP also included a prediction market in their project, and the market did a fairly good job of predicting which of the effects studies would ultimately be reproduced (Dreber et al. 2015). Unlike the prediction

market in Camerer et al. (2016), the RPP prediction market outperformed a survey of researcher beliefs.⁴³

Despite the inability to replicate so many prominent empirical papers in economics (discussed above), there have been few systematic effort to replicate findings, with one exception in addition to Camerer et al. (2016) being the new 3ie replication program for development economics studies, which has replicated a handful papers to date, including one by an author of this article.⁴⁴ Few economics journal editors specifically seek to publish replications, and even fewer are willing to publish “successful” replications, i.e., papers that demonstrate that earlier findings are indeed robust, with the *Journal of Applied Econometrics* being a notable exception (Pesaran 2003). Despite the value to the research enterprise of more systematic evidence on which empirical results are actually reliable, and the fact that many scholars have advocated for changes in this practice over the years with a near constant stream of editorials (see among others Kane 1984; Mittelstaedt and Zorn 1984; Fuess 1996; Hunter 2001; Camfield and Palmer-Jones 2013; Duvendack and Palmer-Jones 2013; Duvendack, Palmer-Jones, and Reed 2015; Zimmermann 2015), as yet there has been little progress within the economics profession towards actually publishing replication studies on a more general basis (Andreoli-Versbach and Mueller-Langer 2014). In many ways, the patterns in economics are similar to those in the other social sciences, particularly in political science where prominent voices have long spoken out in favor of replication, but their publication remains rare (King 1995; Gherghina and Katsanidou 2013).

3.5.1 Computational Issues

Scholars’ ability to carry out replications and share data has been facilitated by new software and computational improvements. Some of these advances are described in Koenker and Zeileis (2009). They discuss what has come to be called Claerbout’s principle: “An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development

⁴³ For related research on expert predictions, see DellaVigna and Pope (2016) . Other psychology researchers have tried another way to crowdsource replication: instead of bringing different research groups together to all independently run the same classic experiment, other researchers have independently analyzed the same observational dataset and attempted to answer the same question, in this case, the question of whether darker skinned soccer players receive more red cards as a result of their race, conditional on other factors (Silberzahn and Uhlmann 2015).

⁴⁴ <http://www.3ieimpact.org/evaluation/impact-evaluation-replication-programme/>.

environment and the complete set of instructions which generated the figures.” Koenker and Zeileis recommend version control, using open source programming environments when possible (including for document preparation), and literate programming, which is defined below.

Version control software makes it easier to maintain detailed record-keeping of changes to statistical code even among multiple collaborators. Koenker and Zeileis discuss one such centralized system, Subversion (SVN, <http://subversion.apache.org>), but in recent years distributed forms of version control such as Git have become more widely used, and are well supported by a user community.⁴⁵

For document preparation, Koenker and Zeileis discuss LaTeX, which has a steep learning curve but has the advantage of being open source, and has the ability to intermix, or “weave” text, code, and output. Even more recently, dynamic documents (which Koenker and Zeileis refer to as literate programming, see also Knuth 1992) can be used to write statistical analysis code and the final paper all in a single master document, making it less likely that copying and pasting between programs will lead to errors, and making it possible in some cases to reproduce an entire project with a single mouse click. The knitr package for R, incorporated into R Studio, makes this relatively easy to implement (Xie 2013; Xie 2014). Jupyter notebooks (<http://jupyter.org>) also simplifies interactive sharing of computational code with over 40 popular open source programming languages (Shen 2014). Many programs that accommodate these approaches, including R, Python, and Julia, are open-source, making it easier for members of the research community to look under the hood and possibly reduce the risk of the software computational errors documented in McCullough and Vinod (2003).⁴⁶ Computational aspects of reproducibility are discussed at length in Stodden, Leisch, and Peng (2014).

3.5.2 The Limits of Open Data

While we believe that economics as a whole would benefit from stronger data sharing requirements and more widespread publication of replication research, there are also potential

⁴⁵ For a how-to manual on version control and other reproducibility tools, see Matthew Gentzkow and Jesse Shapiro’s Practitioner’s Guide at <http://web.stanford.edu/~gentzkow/research/CodeAndData.pdf> or the Best Practices Manual by Garret Christensen at <https://github.com/garretchristensen/BestPracticesManual>.

⁴⁶ The recommendations regarding checking the conditions of Hessians for non-linear solving methods proposed by McCullough and Vinod (2003) are quite detailed, and were modified after omissions were brought to light. See Shachar and Nalebuff (2004); McCullough and Vinod (2004a); Drukker and Wiggins (2004); McCullough and Vinod (2004b).

downsides to data sharing that cannot be ignored. Technological innovations, and in particular the explosion in Internet access over the past 20 years, have made the sharing of data and materials much less costly than was the case in earlier periods. However, the rise of “big data,” and in particular, the massive amounts of personal information that are now publicly available and simple to locate online, also mean that open data sharing raises new concerns regarding individual confidentiality and privacy.

For instance, it has been shown in multiple instances that it is often trivially easy to identify individuals in purportedly “de-identified” and anonymous datasets using publicly available information. In one dramatic illustration, MIT computer science Ph.D. student Latanya Sweeney sent then-Massachusetts Governor William Weld his own complete personal health records only days after anonymized state health records were released to researchers (Sweeney 2002). A new focus of computer science theorists has been developing algorithms for “differential privacy” that simultaneously protect individual privacy while allowing for robust analysis of datasets. They have established that there is inherently a trade-off between these two objectives (Dwork and Smith 2010; Heffetz and Ligett 2014), though few actionable approaches to squaring this circle are currently available to applied researchers, to our knowledge.

4 Future Directions and Conclusion

The rising interest in transparency and reproducibility in economics reflects broader global trends regarding these issues, both among academics and beyond. As such, we argue that “this time” really may be different than earlier bursts of interest in research transparency within economics (such as the surge of interest in the mid-1980’s following Leamer’s 1983 article) that later lost momentum and mostly died down.

The increased institutionalization of new practices – including through the new AEA RCT registry, which has rapidly attracted hundreds of studies, many employing pre-analysis plans, something unheard of in economics until a few years ago – is evidence that new norms are emerging. The rise in the use of pre-analysis plans has been particularly rapid in certain subfields, especially development economics, pushed forward by policy changes promoting pre-analysis plans in the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Center for Effective Global Action. Interest in pre-analysis plans, and more broadly in issues of research

transparency and openness, appears to be particularly high among Ph.D. students and younger faculty (at least anecdotally), suggesting that there may be a generational shift at work.

The Berkeley Initiative for Transparency in the Social Sciences (BITSS) is another new institution that has emerged in recent years to promote dialogue and build consensus around transparency practices. BITSS has established an active training program for the next generation of economists and other social scientists, as well as an award to recognize emerging leaders in this area, the Leamer-Rosenthal Prize for Open Social Science.⁴⁷ Other specialized organizations have also emerged in economics: the Replication Network aims promote the publication of replication studies, Project TIER has developed a curriculum to teach computational reproducibility to economics undergraduates, and MAER-NET has developed guidelines for meta-analysis (Stanley et al. 2013). Similar organizations play analogous roles in other disciplines, including the Center for Open Science (COS), which is most active within psychology (although it spans other fields), and the Evidence in Governance and Politics (EGAP) group⁴⁸.

At the same time, we have highlighted many open questions. The role that pre-analysis plans and study registration could or should play in observational empirical research – which comprises the vast majority of empirical economics work, even a couple of decades into the well-known shift towards experimental designs – as well as in structural econometric work, macroeconomics, and economic theory remains largely unexplored. There is also a question about the impact that the adoption of these new practices will ultimately have on the reliability of empirical research in economics. Will the use of study registries, pre-analysis plans, disclosure statements, and open data and materials lead to improved research quality in a way that can be credibly measured and assessed? To this point, the presumption among advocates (including ourselves, admittedly) is that these changes will indeed lead to improvements, but rigorous evidence on these effects, using meta-analytic approaches or other methods, will be important in determining which practices are in fact most effective, and possibly in building further support for their adoption in the profession.

⁴⁷ <http://www.bitss.org>. In the interest of full disclosure, Miguel is one of the founders of BITSS and currently its faculty director, and Christensen is a post-doctoral research fellow at BITSS. BITSS is an initiative of the Center for Effective Global Action at U.C. Berkeley.

⁴⁸ <http://cos.io>, <http://www.egap.org>.

There are many potential avenues for promoting the adoption of new and arguably preferable practices, such as the data sharing, disclosure and pre-registration approaches described at length in this article. One issue that this article does not directly address is how to most effectively – and rapidly – shift professional norms and practices within the economics research community. Shifts in graduate training curricula⁴⁹, journal standards (such as the Transparency and Openness Promotion Guidelines), and research funder policies might also contribute to the faster adoption of new practices, but their relative importance and the costs of adopting them remain open questions. The study of how social norms among economists have shifted, and continue to evolve, in this area is an exciting social science research topic in its own right, and one that we hope is also the object of greater scholarly inquiry in the coming years.

⁴⁹ See <http://emiguel.econ.berkeley.edu/teaching/12> for an example of a recent Ph.D. level course on research transparency methods at the University of California, Berkeley taught by the authors.

Table 1: Positive Predictive Value (PPV) of Research Findings for Various Combinations of Power ($1 - \beta$), Ratio of True to Not-True Relationships (R), and Researcher Bias (u).

$1 - \beta$	R	u	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.

RCT, randomized controlled trial.

DOI: 10.1371/journal.pmed.0020124.t004

Notes: Reproduced from Table 4 of Ioannidis (2005)

Table 2: Examples of Recent Meta-Analyses in Economics

Paper	Topic	Publication Bias?	Papers (Estimates) Used	Notes
Brodeur et al. (2016)	Wide collection of top publications	+	641 (50,078)	Finds that 10-20% of significant results are misplaced, and should not be considered statistically significant.
Vivalt (2015)	Developing country impact evaluation	+	589 (26,170)	Finds publication bias/specification search is more prevalent in non-experimental work.
Viscusi (2015)	Value of a statistical life (VSL)	+	17 (550)	Use of better and more recent fatality data indicates publication bias exists, but that accepted VSL are correct.
Doucouliaagos, Stanley, and Viscusi (2014)	VSL and income elasticity	+	14 (101)	Previous evidence was mixed, but controlling for publication bias shows the income elasticity of VSL is clearly inelastic.
Doucouliaagos and Stanley (2013)	Meta-meta-analysis	+	87/3,599 (19,528)	87 meta analyses with 3,599 original articles and 19,528 estimates show that 60% of research areas feature substantial or severe publication bias.
Havranek and Irsova (2012)	Foreign direct investment spillovers	~	57 (3,626)	Find publication bias only in published papers and only in the estimates authors consider most important
Mookerjee (2006)	Exports and economic growth	+	76 (95)	Relationship between exports and growth remains significant, but is significantly smaller when corrected for publication bias.
Nijkamp and Poot (2005)	Wage curve literature	+	17 (208)	Evidence of publication bias in the wage curve literature (the relationship between wages and local unemployment); adjusting for it gives an elasticity estimate of -0.07 instead of the previous consensus of -0.1
Abreu, de Groot, and Florax (2005)	Growth rate convergence	-	48 (619)	adjusting for publication bias in the growth literature on convergence does not change estimates significantly
Doucouliaagos (2005)	Economic freedom and economic growth	+	52 (148)	Literature is tainted, but relationship persists despite publication bias.
Rose and Stanley (2005)	Trade and currency unions	+	34 (754)	Relationship persists despite publication bias. Currency union increases trade 30-90%.
Longhi, Nijkamp, and Poot (2005)	Immigration and wages	-	18 (348)	Publication bias is not found to be a major factor. The negative effect of immigration is quite small (0.1%) and varies by country.
Knell and Stix (2005)	Income elasticity of money demand	-	50 (381)	Publication bias does not significantly affect the literature. Income elasticities for narrow money range from 0.4 to 0.5 for the US and 1.0 to 1.3 for other countries.
Doucouliaagos and Laroche (2003)	Union productivity effects	+	73 (73)	Publication bias is not considered a major issue. Negative productivity associations are found in the UK, with positive associations in the US.
Gorg and Strobl (2001)	Multi-national corporations and productivity spillovers	+	21 (25)	Study design affects results, with cross-sectional studies reporting higher coefficients than panel data studies. There is also some evidence of publication bias.
Ashenfelter, Harmon, and Oosterbeek (1999)	Returns to Education	+	27 (96)	Publication bias is found, and controlling for it significantly reduces the differences between types of estimates of returns to education.

Notes: Table shows a sample of recent papers conducting meta-analyses and testing for publication bias in certain literatures in economics. Positive evidence for publication bias indicated by '+', evidence for no publication bias with '-', and mixed evidence with '~'. The number of papers and total estimates used in the meta-analysis are also shown.

Table 3: Transparency Policies at Selected Top Economics and Finance Journals

Journal	Data sharing policy?	Notes	Replication/ comment publication?	Notes	Funding/conflict of interest disclosure?	Notes
<i>American Economic Review</i>	Yes	Current policy was announced in 2004, becoming effective in 2005. It is in effect for all AEA journals.	Yes		Yes	Implemented in July 2012 for all AEA journals.
<i>American Economic Journals (Applied Economics; Economic Policy; Macroeconomics)</i>	Yes	Same as <i>AER</i> . Since journal inception in 2009.	Yes	Allow post-publication peer review on website.	Yes	Same as <i>AER</i> .
<i>Econometrica</i>	Yes	Began in 2004. See Dekel et al. (2006).	Yes		Yes	Peer review conflict of interest statement printed January 2009. Current financial disclosure policy adopted May 2014.
<i>Journal of Finance</i>	No		Yes		Yes	Current policy adopted August 2015.
<i>Journal of Financial Economics</i>	No	Some data is available on the journal webpage, but there appears to be no official policy.	No		Yes	Current policy adopted November, 2015.
<i>Journal of Political Economy</i>	Yes	Uses the same policy as the <i>AER</i> . Announced in 2005, effective in 2006.	Yes	Submission instructions state that authors of comments must correspond with original authors.	No	
<i>Quarterly Journal of Economics</i>	Yes	Uses the same policy as the <i>AER</i> , adopted 2016.	Yes		Yes	
<i>Review of Economic Studies</i>	Yes	Start date unclear.	No		No	
<i>Review of Financial Studies</i>	No		Yes		Yes	Adopted August 2006. Updated June 2016.

Notes: These 11 journals are at the top of the Scientific Journal Rankings (SJR), excluding the Journal of Economic Literature, since its publications are generally reviews; see <http://www.scimagojr.com/journalrank.php?area=2000>. The American Economic Journal: Microeconomics has the same policies as the other AEJ journals, but is lower ranked. Data sharing policy indicates whether the journal has a policy requiring authors to submit data that produces final results. Information obtained from journal websites and instructions for authors as well as via email to journal staff through October 2016. Replication/comment publication indicates whether the journal has published a replication, as per Duvendack, Palmer-Jones, and Reed (2015) or The Replication Network list (<http://replicationnetwork.com/replication-studies/>) as well as journal websites. Since “replication” is an imprecise term, this categorization is perhaps subject to some debate.

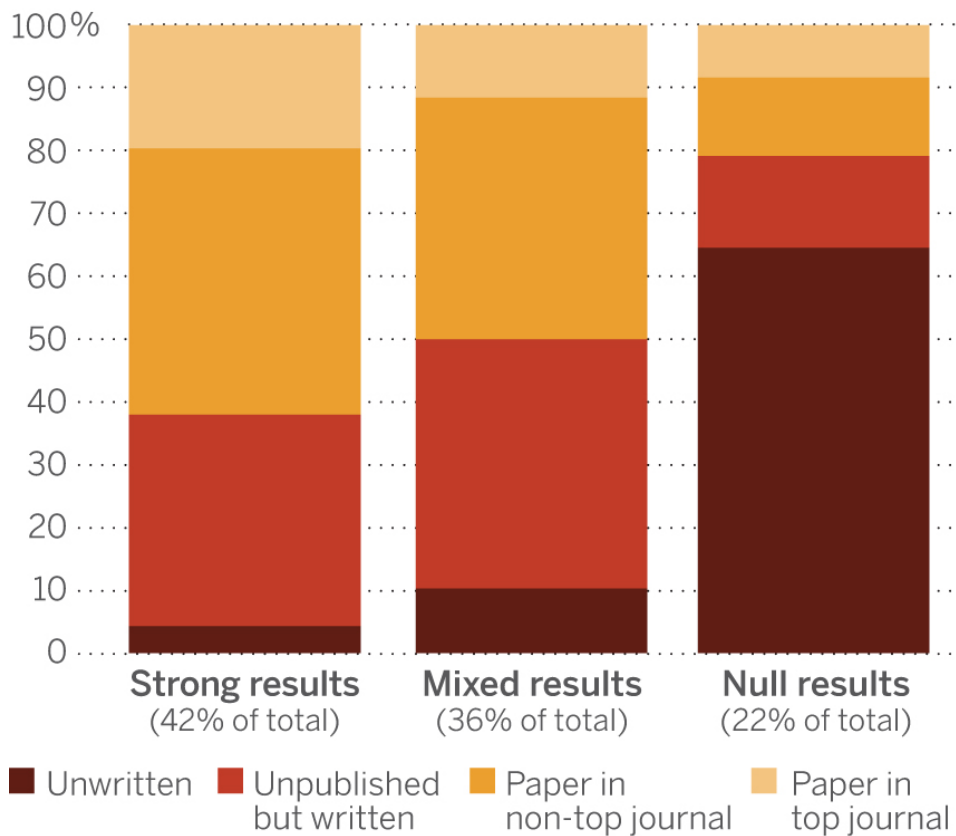
Table 4: Erroneous Interpretations under “Cherry Picking”

Outcome variable:	Mean in control group	Treatment effect	Standard error
Panel A: GoBifo “weakened institutions”			
Attended meeting to decide what to do with the tarp	0.81	-0.04+	(0.02)
Everybody had equal say in deciding how to use the tarp	0.51	-0.11+	(0.06)
Community used the tarp (verified by physical assessment)	0.90	-0.08+	(0.04)
Community can show research team the tarp	0.84	-0.12*	(0.05)
Respondent would like to be a member of the VDC	0.36	-0.04*	(0.02)
Respondent voted in the local government election (2008)	0.85	-0.04*	(0.02)
Panel B: GoBifo “strengthened institutions”			
Community teachers have been trained	0.47	0.12+	(0.07)
Respondent is a member of a women's group	0.24	0.06**	(0.02)
Someone took minutes at the most recent community meeting	0.30	0.14*	(0.06)
Building materials stored in a public place when not in use	0.13	0.25*	(0.10)
Chieftdom official did not have the most influence over tarpaulin use	0.54	0.06*	(0.03)
Respondent agrees with "Responsible young people can be good leaders"	0.76	0.04*	(0.02)
Correctly able to name the year of the next general elections	0.19	0.04*	(0.02)

*Notes: Reproduced from Casey et al (2012, Table VI). i) significance levels (per comparison p-value) indicated by + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$; ii) robust standard errors; iii) treatment effects estimated on follow-up data; and iv) includes fixed effects for the district council wards (the unit of stratification) and the two balancing variables from the randomization (total households and distance to road) as controls.*

Most null results are never written up

The fate of 221 social science experiments



Source: A. Franco *et al.*, *Science* (28 August)

Figure 1: Publication rates and rates of writing-up of results from experiments with strong, mixed, and null results. Figure from Mervis (2014b). Reprinted with permission from AAAS. Experiments represent nearly the complete universe of studies conducted by the Timesharing Experiments for the Social Sciences.

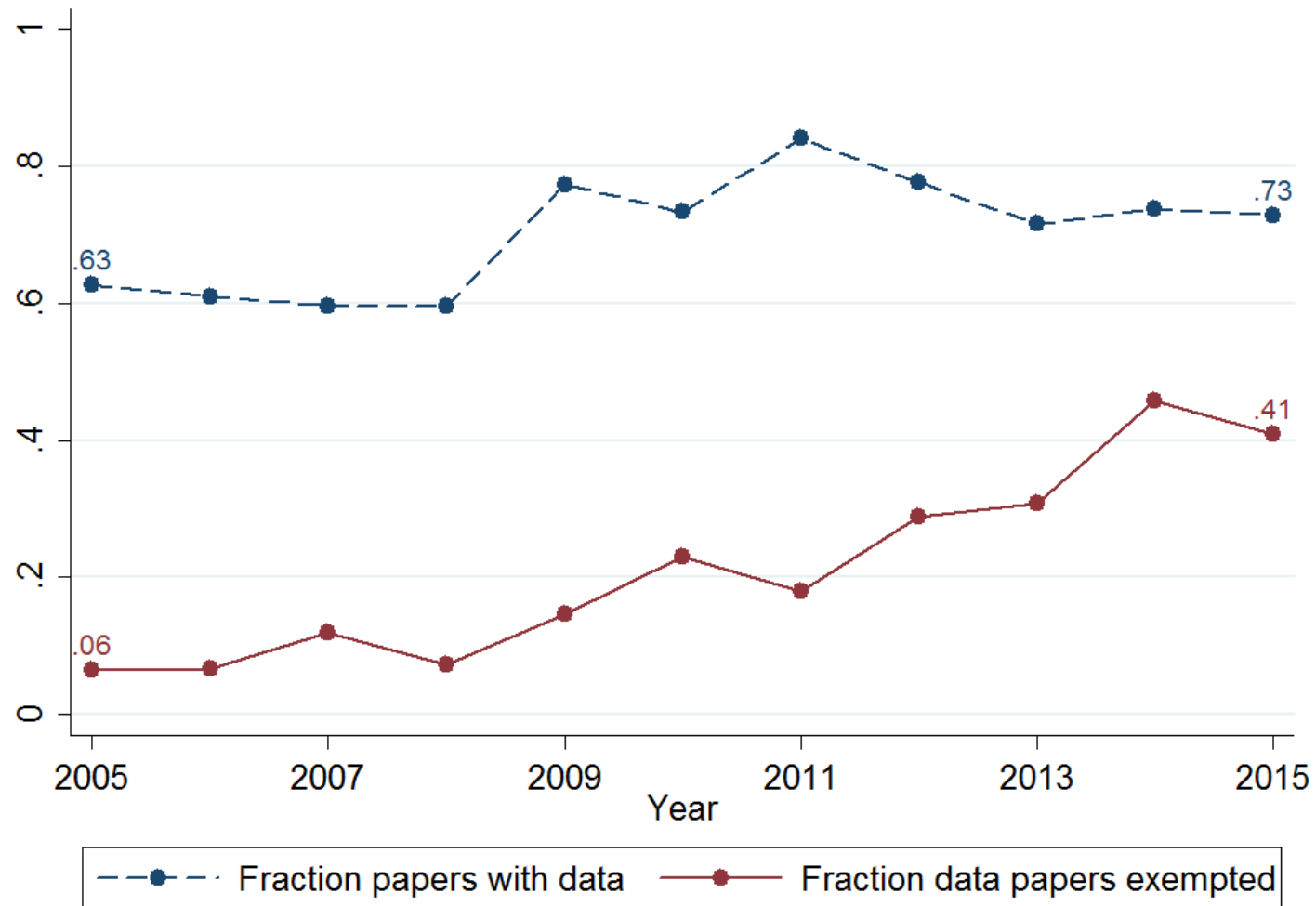


Figure 2: AER Papers with data exempt from the data-sharing requirement. Figure shows annual data on the fraction of American Economic Review papers that use data, and the fraction of those data-using papers that were exempted from the data-sharing policy. Data is taken from the Annual Report of the Editors, which appears annually in the Papers & Proceedings issue of the AER. Figure available in public domain: <http://dx.doi.org/10.7910/DVN/FU07FC>.

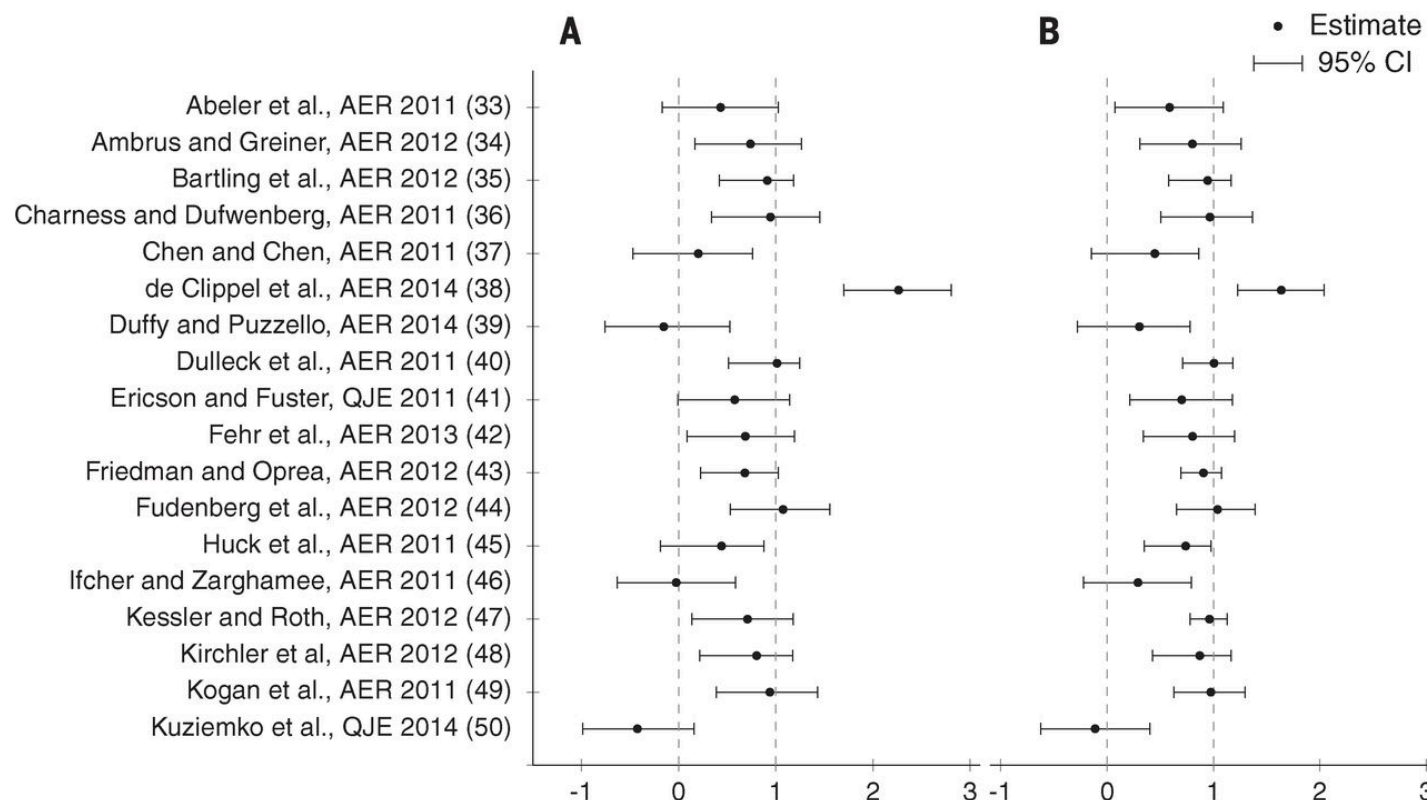


Figure 3: Replicability in Experimental Economics. Figure from Camerer et al. (2016). Reprinted with permission from AAAS. (A) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized so that 1 equals the original effect size (fig. S1 in Camerer et al. 2016 shows a non-normalized version). Eleven replications have a significant effect in the same direction as in the original study [61.1%; 95% CI = (36.2%, 86.1%)]. The 95% CI of the replication effect size includes the original effect size for 12 replications [66.7%; 95% CI = (42.5%, 90.8%)]; if one also includes the study in which the entire 95% CI exceeds the original effect size, this increases to 13 replications [72.2%; 95% CI = (49.3%, 95.1%)]. AER denotes the American Economic Review and QJE denotes the Quarterly Journal of Economics. (B) Meta-analytic estimates of effect sizes, combining the original and replication studies. Plotted are 95% CIs of combined effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized as in (A) (where again fig. S1 shows a non-normalized version). Fourteen studies have a significant effect in the same direction as the original study in the meta-analysis [77.8%; 95% CI = (56.5%, 99.1%)].

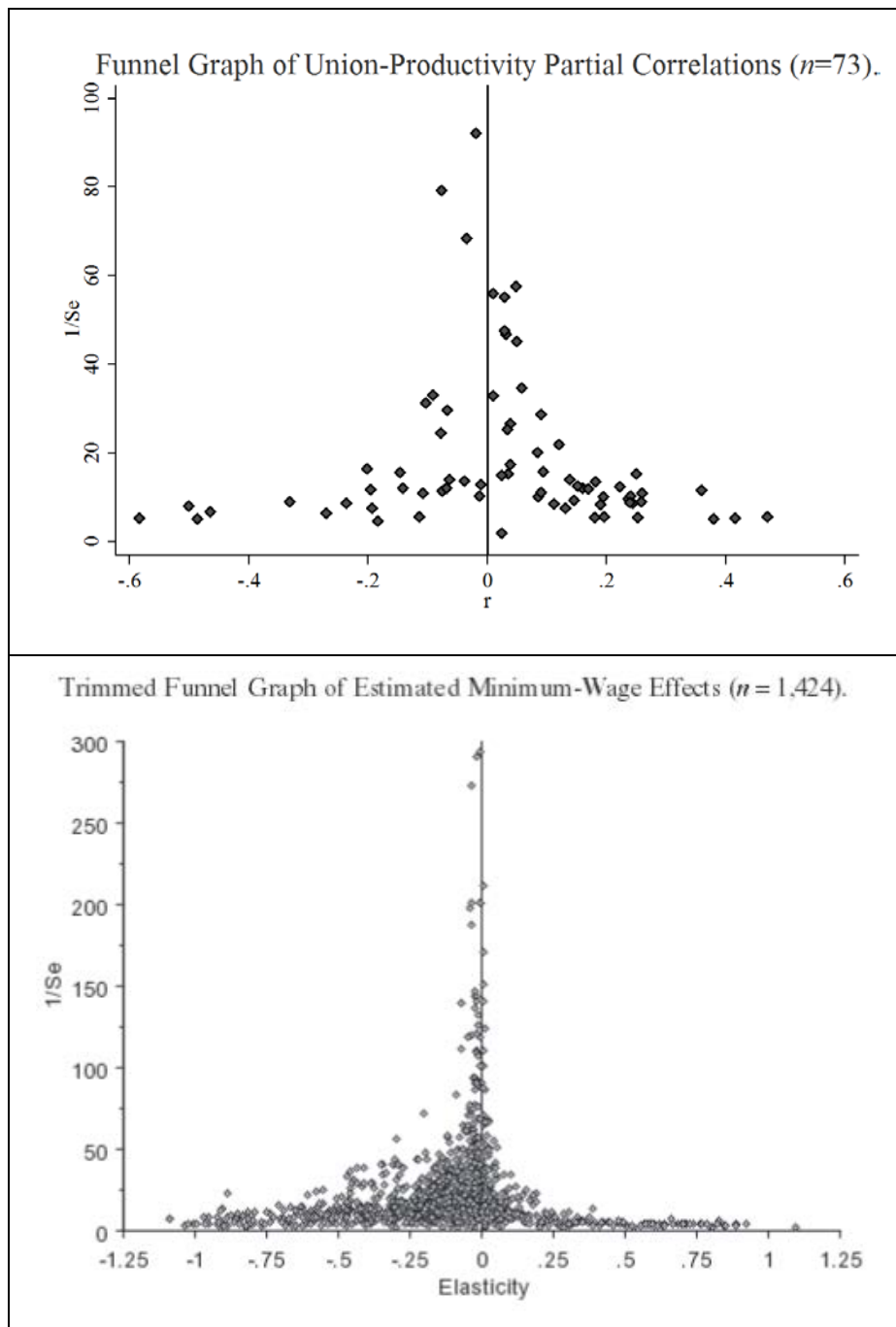


Figure 4: Examples of Funnel Graphs from the Union and Minimum Wage Literature in Labor Economics. Figure shows two funnel graphs with the estimate on the horizontal axis and precision on the vertical axis, with each dot representing an estimate from a study or paper. The top, more symmetric graph is from union-productivity literature, while the bottom, skewed, graph shows estimates from the minimum wage literature. Top figure—constructed by the authors using data from Doucouliagos and Larouche (2003). Bottom figure—Source: Doucouliagos and Stanley (2009). Used by permission. © Blackwell Publishing Ltd/London School of Economics 2009

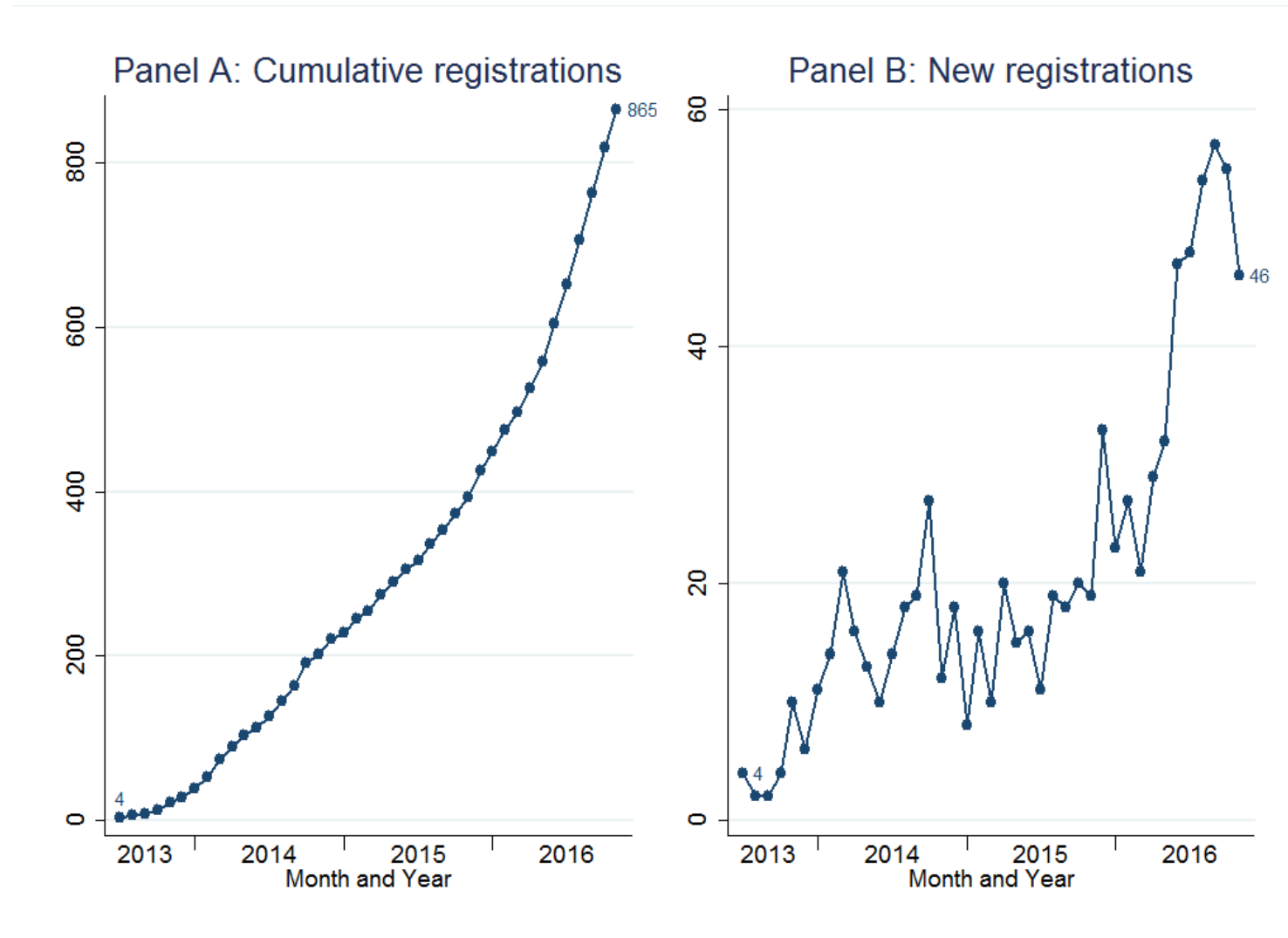


Figure 5: Studies in the AEA trial registry, May 2013 to September 2016. Figure shows the cumulative (Panel A) and new (Panel B) trial registrations in the American Economics Association Trial Registry (<http://socialscienceregistry.org>). Figure available in public domain: <http://dx.doi.org/10.7910/DVN/FUO7FC>.

REFERENCES

- Abreu, Maria, Henri L. F. de Groot, and Raymond J. G. M. Florax. 2005. "A Meta-Analysis of β -Convergence: The Legendary 2%." *Journal of Economic Surveys* 19 (3): 389–420. doi:10.1111/j.0950-0804.2005.00253.x.
- Acemoglu, Daron, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401. doi:10.1257/aer.91.5.1369.
- Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply." *The American Economic Review* 102 (6): 3077–3110. doi:10.1257/aer.102.6.3077.
- Aiken, Alexander M., Calum Davey, James R. Hargreaves, and Richard J. Hayes. 2015. "Re-Analysis of Health and Educational Impacts of a School-Based Deworming Programme in Western Kenya: A Pure Replication." *International Journal of Epidemiology*, July, dyv127. doi:10.1093/ije/dyv127.
- Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102 (4): 1206–40. doi:10.1257/aer.102.4.1206.
- Albouy, David Y. 2012. "The Colonial Origins of Comparative Development: An Empirical Investigation: Comment." *The American Economic Review* 102 (6): 3059–76. doi:10.1257/aer.102.6.3059.
- Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38. doi:10.1257/aer.20131564.
- Anderson, Christopher J., Štěpán Bahník, Michael Barnett-Cowan, Frank A. Bosco, Jesse Chandler, Christopher R. Chartier, Felix Cheung, et al. 2016. "Response to Comment on 'Estimating the Reproducibility of Psychological Science.'" *Science* 351 (6277): 1037–1037. doi:10.1126/science.aad9163.
- Anderson, Melissa S., Brian C. Martinson, and Raymond De Vries. 2007. "Normative Dissonance in Science: Results from a National Survey of U.S. Scientists." *Journal of Empirical Research on Human Research Ethics* 2 (4): 3–14. doi:10.1525/jer.2007.2.4.3.
- Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484): 1481–95. doi:10.1198/016214508000000841.
- Anderson, Richard G., and William G. Dewald. 1994. "Replication and Scientific Standards in Applied Economics a Decade after the Journal of Money, Credit and Banking Project." *Federal Reserve Bank of St. Louis Review*, no. Nov: 79–83.
- Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2008. "The Role of Data/Code Archives in the Future of Economic Research." *Journal of Economic Methodology* 15 (1): 99–119. doi:10.1080/13501780801915574.
- Andreoli-Versbach, Patrick, and Frank Mueller-Langer. 2014. "Open Access to Data: An Ideal Professed but Not Practised." *Research Policy*. Accessed August 29. doi:10.1016/j.respol.2014.04.008.

- Angrist, Joshua D, and Jörn-Steffen Pischke. 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. doi:10.1257/jep.24.2.3.
- Annas, George J. 2003. "HIPAA Regulations — A New Era of Medical-Record Privacy?" *New England Journal of Medicine* 348 (15): 1486–90. doi:10.1056/NEJMLim035027.
- Ashenfelter, Orley, and Michael Greenstone. 2004. "Estimating the Value of a Statistical Life: The Importance of Omitted Variables and Publication Bias." *Princeton University, Department of Economics, Center for Economic Policy Studies, Working Papers: 105, 2004*.
<http://search.proquest.com/econlit/docview/56626726/1A493319C9B3407FPQ/4?accountid=14496>.
- Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias." *Princeton, Department of Economics - Industrial Relations Sections, Princeton, Department of Economics - Industrial Relations Sections, 1999*.
<http://search.proquest.com/econlit/docview/56307618/5A788ADBA14C479EPQ/2?accountid=14496>.
- Backhouse, Roger E., and Mary S. Morgan. 2000. "Introduction: Is Data Mining a Methodological Problem?" *Journal of Economic Methodology* 7 (2): 171–81. doi:10.1080/13501780050045065.
- Bai, Liang, Benjamin Handel, Edward Miguel, and Gautam Rao. 2015. "Self-Control and Chronic Illness: Evidence from Commitment Contracts for Doctor Visits." Unpublished Manuscript. UC Berkeley.
- Baicker, Katherine, Amy Finkelstein, Jae Song, and Sarah Taubman. 2014. "The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment." *American Economic Review* 104 (5): 322–28. doi:10.1257/aer.104.5.322.
- Baicker, Katherine, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. 2013. "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes." *New England Journal of Medicine* 368 (18): 1713–22. doi:10.1056/NEJMsa1212321.
- Bakkensen, Laura A., and William Larson. 2014. "Population Matters When Modeling Hurricane Fatalities." *Proceedings of the National Academy of Sciences of the United States of America* 111 (50): E5331–32. doi:10.1073/pnas.1417030111.
- Bateman, Ian, Daniel Kahneman, Alistair Munro, Chris Starmer, and Robert Sugden. 2005. "Testing Competing Models of Loss Aversion: An Adversarial Collaboration." *Journal of Public Economics, The Experimental Approaches to Public Economics*, 89 (8): 1561–80. doi:10.1016/j.jpubeco.2004.06.013.
- Begg C, Cho M, Eastwood S, and et al. 1996. "Improving the Quality of Reporting of Randomized Controlled Trials: The Consort Statement." *JAMA* 276 (8): 637–39. doi:10.1001/jama.1996.03540080059030.
- Bellavance, François, Georges Dionne, and Martin Lebeau. 2009. "The Value of a Statistical Life: A Meta-Analysis with a Mixed Effects Regression Model." *Journal of Health Economics* 28 (2): 444–64. doi:10.1016/j.jhealeco.2008.10.013.

- Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.
- Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika* 93 (3): 491–507. doi:10.1093/biomet/93.3.491.
- Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics* 29 (4): 1165–88.
- Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, Simon Galle, Edward Miguel, Daniel N. Posner, Bertil Tungodden, and Kelly Zhang. 2015. "How Strong Are Ethnic Preferences?" Working Paper 21715. National Bureau of Economic Research. <http://www.nber.org/papers/w21715>.
- Berger, Ulrich. 2009. "The Convergence of Fictitious Play in Games with Strategic Complementarities: A Comment." <https://mpra.ub.uni-muenchen.de/20241/>.
- Bernanke, Ben S. 2004. "Editorial Statement." *The American Economic Review* 94 (1): 404–404.
- Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94 (4): 991–1013. doi:10.1257/0002828042002561.
- Bhattacharjee, Yudhijit. 2013. "Diederik Stapel's Audacious Academic Fraud." *The New York Times*, April 26, sec. Magazine. <http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html>.
- Bland, J. M., and D. G. Altman. 1995. "Multiple Significance Tests: The Bonferroni Method." *BMJ: British Medical Journal* 310 (6973): 170.
- Bohannon, John. 2016. "About 40% of Economics Experiments Fail Replication Survey." *Science*, March. doi:10.1126/science.aaf4141.
- Bontemps, Christophe, and Grayham E. Mizon. 2008. "Encompassing: Concepts and Implementation*." *Oxford Bulletin of Economics and Statistics* 70 (December): 721–50. doi:10.1111/j.1468-0084.2008.00528.x.
- Boutron, Isabelle, David Moher, Douglas G. Altman, Kenneth F. Schulz, and Philippe Ravaud. 2008. "Extending the CONSORT Statement to Randomized Trials of Nonpharmacologic Treatment: Explanation and Elaboration." *Annals of Internal Medicine* 148 (4): 295–309. doi:10.7326/0003-4819-148-4-200802190-00008.
- Broad, William, and Nicholas Wade. 1983. *Betrayers of the Truth*. Simon and Schuster.
- Brock, William A., Steven N. Durlauf, and Kenneth D. West. 2003. "Policy Evaluation in Uncertain Economic Environments." *Brookings Papers on Economic Activity* 2003 (1): 235–322. doi:10.1353/eca.2003.0013.
- Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8 (1): 1–32.
- Broockman, David, Joshua Kalla, and Peter Aronow. 2015. "Irregularities in LaCour (2014)." UC Berkeley. http://stanford.edu/~dbroock/broockman_kalla_aronow_lg_irregularities.pdf.
- Burnside, Craig, and David Dollar. 2000. "Aid, Policies, and Growth." *The American Economic Review* 90 (4): 847–68.

- . 2004. "Aid, Policies, and Growth: Reply." *The American Economic Review* 94 (3): 781–84.
- Calvert M, Blazeby J, Altman DG, and et al. 2013. "Reporting of Patient-Reported Outcomes in Randomized Trials: The Consort pro Extension." *JAMA* 309 (8): 814–22. doi:10.1001/jama.2013.879.
- Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351 (6280): 1433–36. doi:10.1126/science.aaf0918.
- Camfield, Laura, and Richard Palmer-Jones. 2013. "Three 'Rs' of Econometrics: Repetition, Reproduction and Replication." *Journal of Development Studies* 49 (12): 1607–14. doi:10.1080/00220388.2013.807504.
- Campbell, Marion K., Diana R. Elbourne, and Douglas G. Altman. 2004. "CONSORT Statement: Extension to Cluster Randomised Trials." *BMJ* 328 (7441): 702–8. doi:10.1136/bmj.328.7441.702.
- Campbell, Marion K., Gilda Piaggio, Diana R. Elbourne, and Douglas G. Altman. 2012. "Consort 2010 Statement: Extension to Cluster Randomised Trials." *BMJ* 345 (September): e5661. doi:10.1136/bmj.e5661.
- Campolieti, Michele, Morley Gunderson, and Chris Riddell. 2006. "Minimum Wage Impacts from a Prespecified Research Design: Canada 1981–1997." *Industrial Relations: A Journal of Economy and Society* 45 (2): 195–216. doi:10.1111/j.1468-232X.2006.00424.x.
- Card, David. 1992a. "Do Minimum Wages Reduce Employment? A Case Study of California, 1987–89." *Industrial & Labor Relations Review* 46 (1): 38–54. doi:10.1177/001979399204600104.
- . 1992b. "Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage." *Industrial & Labor Relations Review* 46 (1): 22–37. doi:10.1177/001979399204600103.
- Card, David, Raj Chetty, Martin Feldstein, and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." 112. NSF SBE 2020. http://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Card_David_112.pdf.
- Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review* 84 (4): 772–93.
- . 1995. "Time-Series Minimum-Wage Studies: A Meta-Analysis." *The American Economic Review* 85 (2): 238–43.
- . 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply." *The American Economic Review* 90 (5): 1397–1420.
- Carey, Benedict. 2011. "Noted Dutch Psychologist, Stapel, Accused of Research Fraud." *The New York Times*, November 2, sec. Health / Research. <http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html>.
- Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*." *The Quarterly Journal of Economics* 127 (4): 1755–1812. doi:10.1093/qje/qje027.

- Casselmann, Ben. 2012. "Economists Set Rules on Ethics." *Wall Street Journal*, January 9, sec. Careers.
<http://www.wsj.com/articles/SB10001424052970203436904577148940410667970>.
- Chambers, Christopher. 2013. "Registered Reports: A New Publishing Initiative at Cortex." *Cortex* 49: 609–10. doi:<http://dx.doi.org/10.1016/j.cortex.2012.12.016>.
- Chambers, Christopher D., Eva Feredoes, Suresh D. Muthukumaraswamy, and Peter J. Etchells. 2014. "Instead of 'playing the Game' It Is Time to Change the Rules: Registered Reports at AIMS Neuroscience and beyond." *AIMS Environmental Science* 1 (1): 4–17. doi:10.3934/Neuroscience.2014.1.4.
- Chang, Andrew C., and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not.'" Finance and Economics Discussion Series 2015–83. Board of Governors of the Federal Reserve System (U.S.).
<http://econpapers.repec.org/paper/fipfedgfe/2015-83.htm>.
- Chenault, Larry A. 1984. "A Note on the Stability Limitations in 'A Stable Price Adjustment Process.'" *The Quarterly Journal of Economics* 99 (2): 385. doi:10.2307/1885533.
- Christensen, B., and S. Christensen. 2014. "Are Female Hurricanes Really Deadlier than Male Hurricanes?" *Proceedings of the National Academy of Sciences* 111 (34): E3497–98. doi:10.1073/pnas.1410910111.
- Cicchione, Antonio. 2011. "Economic Shocks and Civil Conflict: A Comment." *American Economic Journal: Applied Economics* 3 (4): 215–27. doi:10.1257/app.3.4.215.
- Clemens, Michael. 2015. "The Meaning of Failed Replications: A Review and Proposal." 399. Center for Global Development Working Paper.
<http://www.cgdev.org/publication/meaning-failed-replications-review-and-proposal-working-paper-399>.
- Coffman, Lucas C., and Muriel Niederle. 2015. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible." *Journal of Economic Perspectives* 29 (3): 81–98.
- Cohen-Cole, Ethan, Steven Durlauf, Jeffrey Fagan, and Daniel Nagin. 2009. "Model Uncertainty and the Deterrent Effect of Capital Punishment." *American Law and Economics Review* 11 (2): 335–69. doi:10.1093/aler/ahn001.
- "Correspondence: David H Autor and Bruno S. Frey." 2011. *Journal of Economic Perspectives* 25 (3): 239–40.
- Cowen, Tyler, and Alex Tabarrok. 2016. "A Skeptical View of the National Science Foundation's Role in Economic Research." *Journal of Economic Perspectives* 30 (3): 235–48. doi:10.1257/jep.30.3.235.
- Dahl Rasmussen, Ole, Nikolaj Malchow-Møller, and Thomas Barnebeck Andersen. 2011. "Walking the Talk: The Need for a Trial Registry for Development Interventions." *Journal of Development Effectiveness* 3 (4): 502–19.
doi:10.1080/19439342.2011.605160.
- Dal-Ré, Rafael, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. 2014. "Making Prospective Registration of Observational Research a Reality." *Science Translational Medicine* 6 (224): 224cm1–224cm1. doi:10.1126/scitranslmed.3007513.

- Davis, Graham A. 2013. "Replicating Sachs and Warner's Working Papers on the Resource Curse." *The Journal of Development Studies* 49 (12): 1615–30. doi:10.1080/00220388.2013.807501.
- De Angelis, Catherine, Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin, et al. 2004. "Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors." *New England Journal of Medicine* 351 (12): 1250–51. doi:10.1056/NEJMe048225.
- Deere, Donald, Kevin M. Murphy, and Finis Welch. 1995. "Employment and the 1990-1991 Minimum-Wage Hike." *The American Economic Review* 85 (2): 232–37.
- Dekel, Eddie, David Levine, Costas Meghir, Whitney Newey, and Andrew Postlewaite. 2006. "Report of the Editors." *Econometrica* 74 (1): 307–10.
- DellaVigna, Stefano, and Devin Pope. 2016. "Predicting Experimental Results: Who Knows What?" Working Paper 22566. National Bureau of Economic Research. <http://www.nber.org/papers/w22566>.
- DeLong, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100 (6): 1257–72.
- Denton, Frank T. 1985. "Data Mining as an Industry." *The Review of Economics and Statistics* 67 (1): 124–27. doi:10.2307/1928442.
- Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review* 76 (4): 587–603.
- Dezhbakhsh, Hashem, Paul H. Rubin, and Joanna M. Shepherd. 2003. "Does Capital Punishment Have a Deterrent Effect? New Evidence from Postmoratorium Panel Data." *American Law and Economics Review* 5 (2): 344–76. doi:10.1093/aler/ahg021.
- Donohue, John J., and Justin Wolfers. 2005. "Uses and Abuses of Empirical Evidence in the Death Penalty Debate." *Stanford Law Review* 58 (3): 791–845.
- Doucouliaos, Chris. 2005. "Publication Bias in the Economic Freedom and Economic Growth Literature." *Journal of Economic Surveys* 19 (3): 367–87. doi:10.1111/j.0950-0804.2005.00252.x.
- Doucouliaos, Chris, T. D. Stanley, and Margaret Giles. 2012. "Are Estimates of the Value of a Statistical Life Exaggerated?" *Journal of Health Economics* 31 (1): 197–206. doi:10.1016/j.jhealeco.2011.10.001.
- Doucouliaos, Chris, and T.d. Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys* 27 (2): 316–39. doi:10.1111/j.1467-6419.2011.00706.x.
- Doucouliaos, Christos, and Patrice Laroche. 2003. "What Do Unions Do to Productivity? A Meta-Analysis." *Industrial Relations: A Journal of Economy and Society* 42 (4): 650–91. doi:10.1111/1468-232X.00310.
- Doucouliaos, Hristos, John P. Ioannidis, and Tom Stanley. 2017. "The Power of Bias in Economics." *The Economic Journal* Forthcoming. https://www.deakin.edu.au/_data/assets/pdf_file/0007/477763/2016_1.pdf.
- Doucouliaos, Hristos, and T. D. Stanley. 2009. "Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis." *British Journal of Industrial Relations* 47 (2): 406–28. doi:10.1111/j.1467-8543.2009.00723.x.

- Doucouliaagos, Hristos, T. D. Stanley, and W. Kip Viscusi. 2014. "Publication Selection and the Income Elasticity of the Value of a Statistical Life." *Journal of Health Economics* 33 (January): 67–75. doi:10.1016/j.jhealeco.2013.10.010.
- Drazen, Jeffrey M. 2016. "Data Sharing and the Journal." *New England Journal of Medicine* 374 (19): e24. doi:10.1056/NEJMe1601087.
- Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences*, November, 201516179. doi:10.1073/pnas.1516179112.
- Drukker, David M, and Vince Wiggins. 2004. "Verifying the Solution from a Nonlinear Solver: A Case Study: Comment." *American Economic Review* 94 (1): 397–99. doi:10.1257/000282804322970896.
- Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. "Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties." *Review of Economics and Statistics* 92 (4): 945–64. doi:10.1162/REST_a_00039.
- Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Chapter 61 Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, edited by T. Paul Schultz and John A. Strauss, 4:3895–3962. Elsevier. <http://www.sciencedirect.com/science/article/pii/S1573447107040612>.
- Duvendack, Maren, and Richard Palmer-Jones. 2013. "Replication of Quantitative Work in Development Studies: Experiences and Suggestions." *Progress in Development Studies* 13 (4): 307–22. doi:10.1177/1464993413490480.
- Duvendack, Maren, Richard Palmer-Jones, and Robert W. Reed. 2015. "Replications in Economics: A Progress Report." *Econ Journal Watch* 12 (2): 164–91.
- Dwork, Cynthia, and Adam Smith. 2010. "Differential Privacy for Statistics: What We Know and What We Want to Learn." *Journal of Privacy and Confidentiality* 1 (2). <http://repository.cmu.edu/jpc/vol1/iss2/2>.
- Easterbrook, P. J, R Gopalan, J. A Berlin, and D. R Matthews. 1991. "Publication Bias in Clinical Research." *The Lancet*, Originally published as Volume 1, Issue 8746, 337 (8746): 867–72. doi:10.1016/0140-6736(91)90201-Y.
- Easterly, William, Ross Levine, and David Roodman. 2004. "Aid, Policies, and Growth: Comment." *The American Economic Review* 94 (3): 774–80.
- Eble, Alex, Peter Boone, and Diana Elbourne. 2014. "On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics." SSRN Scholarly Paper ID 2272141. Rochester, NY: Social Science Research Network. <http://papers.ssrn.com/abstract=2272141>.
- Eich, Eric. 2014. "Business Not as Usual." *Psychological Science* 25 (1): 3–6. doi:10.1177/0956797613512465.
- Elm, Erik von, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. 2007. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *Preventive Medicine* 45 (4): 247–51. doi:10.1016/j.ypmed.2007.08.012.
- Enders, Walter, and Gary Hoover. 2006. "Plagiarism in the Economics Profession: A Survey." *Challenge* 49 (5): 92–107. doi:10.2753/0577-5132490506.

- Enders, Walter, and Gary A. Hoover. 2004. "Whose Line Is It? Plagiarism in Economics." *Journal of Economic Literature* 42 (2): 487–93. doi:10.1257/0022051041409066.
- Epidemiology. 2010. "The Registration of Observational Studies—When Metaphors Go Bad." *Epidemiology*, July, 1. doi:10.1097/EDE.0b013e3181eafbcf.
- Fanelli, Daniele. 2009. "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data." *PLoS ONE* 4 (5): e5738. doi:10.1371/journal.pone.0005738.
- . 2013. "Why Growing Retractions Are (Mostly) a Good Sign." *PLOS Med* 10 (12): e1001563. doi:10.1371/journal.pmed.1001563.
- Fang, Ferric C., and Arturo Casadevall. 2011. "Retracted Science and the Retraction Index." *Infection and Immunity* 79 (10): 3855–59. doi:10.1128/IAI.05661-11.
- Feldstein, Martin. 1974. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." *Journal of Political Economy* 82 (5): 905–26.
- Feldstein, Martin S. 1982. "Social Security and Private Saving: Reply." *Journal of Political Economy* 90 (3): 630–42.
- Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel. 2001. "Model Uncertainty in Cross-Country Growth Regressions." *Journal of Applied Econometrics* 16 (5): 563–76. doi:10.1002/jae.623.
- Findley, Michael, Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. 2016. "Introduction: Special Issue on Research Transparency in the Social Science." *Comparative Political Studies*.
- Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures." *Journal of Development Effectiveness* 6 (1): 44–57. doi:10.1080/19439342.2013.875054.
- Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, and Katherine Baicker. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year*." *The Quarterly Journal of Economics* 127 (3): 1057–1106. doi:10.1093/qje/qjs020.
- Fleischmann, Martin, and Stanley Pons. 1989. "Electrochemically Induced Nuclear Fusion of Deuterium." *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry* 261 (2): 301–8. doi:10.1016/0022-0728(89)80006-3.
- Food and Drug Administration. n.d. "E9 Statistical Principles for Clinical Trials." Guidance for Industry. Food and Drug Administration. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf>.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5. doi:10.1126/science.1255484.
- Frey, René L., Bruno S. Frey, and Reiner Eichenberger. 1999. "A Case of Plagiarism." *Kyklos* 52 (3): 311–311. doi:10.1111/j.1467-6435.1999.tb00219.x.
- Fuess, Scott M., Jr. 1996. "On Replication in Business and Economics Research: The QJBE Case." *Quarterly Journal of Business and Economics* 35 (2): 3–13.
- Galbraith, R. F. 1988. "A Note on Graphical Presentation of Estimated Odds Ratios from Several Clinical Trials." *Statistics in Medicine* 7 (8): 889–94. doi:10.1002/sim.4780070807.

- Ganimian, Alejandro. 2014. "Pre-Analysis Plan Template." Harvard University.
http://scholar.harvard.edu/files/alejandro_ganimian/files/pre-analysis_plan_template_0.pdf.
- Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'fishing Expedition' or 'p-Hacking' and the Research Hypothesis Was Posited ahead of Time.," November.
http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.
- Gelman, Andrew, Cristian Pasarica, and Rahul Dodhia. 2002. "Let's Practice What We Preach." *The American Statistician* 56 (2): 121–30.
doi:10.1198/000313002317572790.
- Gerber, Alan, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers, and David J. Hendry. 2014. "Reporting Guidelines for Experimental Research: A Report from the Experimental Research Section Standards Committee." *Journal of Experimental Political Science* 1 (01): 81–98.
doi:10.1017/xps.2014.11.
- Gerber, Alan, and Neil Malhotra. 2008a. "Publication Bias in Empirical Sociological Research Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37 (1): 3–30. doi:10.1177/0049124108318973.
- . 2008b. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3 (3): 313–26. doi:10.1561/100.00008024.
- Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. "Testing for Publication Bias in Political Science." *Political Analysis* 9 (4): 385–92.
- Gerking, Shelby, and William E Morgan. 2007. "Effects of Environmental and Land Use Regulation in the Oil and Gas Industry Using the Wyoming Checkerboard as a Natural Experiment: Retraction." *American Economic Review* 97 (3): 1032–1032.
doi:10.1257/aer.97.3.1032.
- Gherghina, Sergiu, and Alexia Katsanidou. 2013. "Data Availability in Political Science Journals." *European Political Science* 12 (3): 333–49. doi:10.1057/eps.2013.8.
- Gilbert, Christopher L. 1989. "LSE and the British Approach to Time Series Econometrics." *Oxford Economic Papers*, New Series, 41 (1): 108–28.
- Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson. 2016. "Comment on 'Estimating the Reproducibility of Psychological Science.'" *Science* 351 (6277): 1037–1037. doi:10.1126/science.aad7243.
- Glandon, Philip. 2010. "Report on the American Economic Review Data Availability Compliance Project." Vanderbilt University.
https://aeaweb.org/aer/2011_Data_Compliance_Report.pdf.
- Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.
- Gorg, Holger, and Eric Strobl. 2001. "Multinational Companies and Productivity Spillovers: A Meta-Analysis." *The Economic Journal* 111 (475): 723–39. doi:10.1111/1468-0297.00669.
- Grant, Sean, Evan Mayo-Wilson, Sally Hopewell, Geraldine Macdonald, David Moher, and Paul Montgomery. 2013. "Developing a Reporting Guideline for Social and Psychological Intervention Trials." *Journal of Experimental Criminology* 9 (3): 355–67. doi:10.1007/s11292-013-9180-5.

- Hamermesh, Daniel S. 2007. "Viewpoint: Replication in Economics." *Canadian Journal of Economics/Revue Canadienne D'économique* 40 (3): 715–33. doi:10.1111/j.1365-2966.2007.00428.x.
- Hansen, Peter Reinhard. 2005. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics* 23 (4): 365–80. doi:10.1198/073500105000000063.
- Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2015. "... and the Cross-Section of Expected Returns." *Review of Financial Studies*, October, hhv059. doi:10.1093/rfs/hhv059.
- Hawkins, Carlee Beth, Cailey E. Fitzgerald, and Brian A. Nosek. 2015. "In Search of an Association Between Conception Risk and Prejudice." *Psychological Science* 26 (2): 249–52. doi:10.1177/0956797614553121.
- Hedges, Larry V. 1992. "Modeling Publication Selection Effects in Meta-Analysis." *Statistical Science* 7 (2): 246–55.
- Hedges, Larry V., and Jack L. Vevea. 1996. "Estimating Effect Size Under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model." *Journal of Educational and Behavioral Statistics* 21 (4): 299–332. doi:10.3102/10769986021004299.
- Heffetz, Ori, and Katrina Ligett. 2014. "Privacy and Data-Based Research." *Journal of Economic Perspectives* 28 (2): 75–98. doi:10.1257/jep.28.2.75.
- Hendry, David F. 1987. "Econometric Methodology: A Personal Perspective." In *Advances in Econometrics: Volume 2: Fifth World Congress*, edited by Truman Bewley. Vol. 2. Cambridge University Press.
- . 1995. *Dynamic Econometrics*. Oxford University Press.
- Henry, Emeric. 2009. "Strategic Disclosure of Research Results: The Cost of Proving Your Honesty*." *The Economic Journal* 119 (539): 1036–64. doi:10.1111/j.1468-0297.2009.02265.x.
- Henry, Emeric, and Marco Ottaviani. 2014. "Research and the Approval Process." <http://www.cepr.org/sites/default/files/Henry-submission%20CEPR.pdf>.
- Hirshleifer, Sarojini, David McKenzie, Rita Almeida, and Cristobal Ridao-Cano. 2015. "The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey." *The Economic Journal*, June, n/a-n/a. doi:10.1111/ecoj.12211.
- Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.
- Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20. doi:10.1093/pan/mps021.
- Hung, H. M. James, Robert T. O'Neill, Peter Bauer, and Karl Kohne. 1997. "The Behavior of the P-Value When the Alternative Hypothesis Is True." *Biometrics* 53 (1): 11–22. doi:10.2307/2533093.
- Hunter, John E. 2001. "The Desperate Need for Replications." *Journal of Consumer Research* 28 (1): 149–58. doi:10.1086/jcr.2001.28.issue-1.
- Husereau, Don, Michael Drummond, Stavros Petrou, Chris Carswell, David Moher, Dan Greenberg, Federico Augustovski, et al. 2013. "Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement." *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research* 16 (2): e1-5. doi:10.1016/j.jval.2013.02.010.

- Ifcher, John, and Homa Zarghamee. 2011. "Happiness and Time Preference: The Effect of Positive Affect in a Random-Assignment Experiment." *American Economic Review* 101 (7): 3109–29. doi:10.1257/aer.101.7.3109.
- Ioannidis, John P. A. 2005a. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8): e124. doi:10.1371/journal.pmed.0020124.
- . 2005b. "Why Most Published Research Findings Are False." *PLoS Med* 2 (8): e124. doi:10.1371/journal.pmed.0020124.
- Ioannidis, John PA. 2008. "Effectiveness of Antidepressants: An Evidence Myth Constructed from a Thousand Randomized Trials?" *Philosophy, Ethics, and Humanities in Medicine* 3 (1): 14. doi:10.1186/1747-5341-3-14.
- ISIS-2 (SECOND INTERNATIONAL STUDY OF INFARCT SURVIVAL) COLLABORATIVE GROUP. 1988. "RANDOMISED TRIAL OF INTRAVENOUS STREPTOKINASE, ORAL ASPIRIN, BOTH, OR NEITHER AMONG 17 187 CASES OF SUSPECTED ACUTE MYOCARDIAL INFARCTION: ISIS-2." *The Lancet*, Originally published as Volume 2, Issue 8607, 332 (8607): 349–60. doi:10.1016/S0140-6736(88)92833-4.
- John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32. doi:10.1177/0956797611430953.
- Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe. 2014. "Female Hurricanes Are Deadlier than Male Hurricanes." *Proceedings of the National Academy of Sciences* 111 (24): 8782–87. doi:10.1073/pnas.1402786111.
- Kane, Edward J. 1984. "WHY JOURNAL EDITORS SHOULD ENCOURAGE THE REPLICATION OF APPLIED ECONOMETRIC RESEARCH." *Quarterly Journal of Business and Economics* 23 (1): 3–8.
- Karabag, Solmaz Filiz, and Christian Berggren. 2012. "Retraction, Dishonesty and Plagiarism: Analysis of a Crucial Issue for Academic Publishing, and the Inadequate Responses from Leading Journals in Economics and Management Disciplines." *Journal of Applied Economics and Business Research* 2 (4): 172–83.
- Katz, Larry, Esther Duflo, Pinelopi Goldberg, and Duncan Thomas. 2013. "AEA E-Mail Announcement." American Economic Association. https://www.aeaweb.org/announcements/20131118_rct_email.php.
- King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28 (03): 444–52. doi:10.2307/420301.
- Kirsch, Irving, Brett J Deacon, Tania B Huedo-Medina, Alan Scoboria, Thomas J Moore, and Blair T Johnson. 2008. "Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration." *PLoS Medicine* 5 (2). doi:10.1371/journal.pmed.0050045.
- Klein, Joshua R., and Aaron Roodman. 2005. "Blind Analysis in Nuclear and Particle Physics." *Annual Review of Nuclear and Particle Science* 55 (1): 141–63. doi:10.1146/annurev.nucl.55.090704.151521.
- Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability." *Social Psychology* 45 (3): 142–52. doi:10.1027/1864-9335/a000178.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2015. "Human Decisions and Machine Problems." Unpublished Manuscript.

- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. doi:10.1257/aer.p20151023.
- Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119. doi:10.1111/j.1468-0262.2007.00733.x.
- Knell, Markus, and Helmut Stix. 2005. "The Income Elasticity of Money Demand: A Meta-Analysis of Empirical Results*." *Journal of Economic Surveys* 19 (3): 513–33. doi:10.1111/j.0950-0804.2005.00257.x.
- Knittel, Christopher R., and Konstantinos Metaxoglou. 2011. "Challenges in Merger Simulation Analysis." *The American Economic Review* 101 (3): 56–59.
- . 2013. "Estimation of Random-Coefficient Demand Models: Two Empiricists' Perspective." *Review of Economics and Statistics* 96 (1): 34–59. doi:10.1162/REST_a_00394.
- Knuth, Donald Ervin. 1992. *Literate Programming*. Center for the Study of Language and Information.
- Koenker, Roger, and Achim Zeileis. 2009. "On Reproducible Econometric Research." *Journal of Applied Econometrics* 24 (5): 833–47. doi:10.1002/jae.1083.
- Kunce, Mitch, Shelby Gerking, and William Morgan. 2002. "Effects of Environmental and Land Use Regulation in the Oil and Gas Industry Using the Wyoming Checkerboard as an Experimental Design." *The American Economic Review* 92 (5): 1588–93.
- Laan, Van Der, Mark J, Eric C. Polley, and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). <http://www.degruyter.com/view/j/sagmb.2007.6.1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml>.
- LaCour, Michael J., and Donald P. Green. 2014. "When Contact Changes Minds: An Experiment on Transmission of Support for Gay Equality." *Science* 346 (6215): 1366–69. doi:10.1126/science.1256151.
- Laine, Christine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug, et al. 2007. "Clinical Trial Registration — Looking Back and Moving Ahead." *New England Journal of Medicine* 356 (26): 2734–36. doi:10.1056/NEJMe078110.
- LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76 (4): 604–20.
- Leamer, E. E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley.
- Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73 (1): 31–43.
- . 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24 (2): 31–46. doi:10.1257/jep.24.2.31.
- . 2016. "S-Values: Conventional Context-Minimal Measures of the Sturdiness of Regression Coefficients." *Journal of Econometrics* 193 (1): 147–61. doi:10.1016/j.jeconom.2015.10.013.
- Leamer, Edward, and Herman Leonard. 1983. "Reporting the Fragility of Regression Estimates." *The Review of Economics and Statistics* 65 (2): 306–17. doi:10.2307/1924497.

- Lee, Soohyung, and Azeem M. Shaikh. 2014. "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of Progreso on School Enrollment." *Journal of Applied Econometrics* 29 (4): 612–26. doi:10.1002/jae.2327.
- Leimer, Dean R., and Selig D. Lesnoy. 1982. "Social Security and Private Saving: New Time-Series Evidence." *Journal of Political Economy* 90 (3): 606–29.
- Levine, David I. 2001. "Editor's Introduction to 'The Unemployment Effects of Minimum Wages: Evidence from a Prespecified Research Design.'" *Industrial Relations: A Journal of Economy and Society* 40 (2): 161–62. doi:10.1111/0019-8676.00204.
- Lewis, Nathan S., C. A. Barnes, M. J. Heben, A. Kumar, S. R. Lunt, G. E. McManis, G. M. Miskelly, et al. 1989. "Searches for Low-Temperature Nuclear Fusion of Deuterium in Palladium." *Nature* 340 (6234). <https://sofi-northwestern.pure.elsevier.com/en/publications/searches-for-low-temperature-nuclear-fusion-of-deuterium-in-palla>.
- Libgober, Jonathan. 2015. "False Positives in Scientific Research." SSRN Scholarly Paper ID 2617130. Rochester, NY: Social Science Research Network. <http://papers.ssrn.com/abstract=2617130>.
- Light, Richard J., and David B. Pillemer. 1984. *Summing Up*. Harvard University Press.
- Lindsay, D. Stephen. 2015. "Replication in Psychological Science." *Psychological Science*, November, 0956797615616374. doi:10.1177/0956797615616374.
- List, Ja, Cd Bailey, Pj Euzent, and Tl Martin. 2001. "Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior." *Economic Inquiry* 39 (1): 162–70. doi:10.1111/j.1465-7295.2001.tb00058.x.
- List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." Working Paper 21875. National Bureau of Economic Research. <http://www.nber.org/papers/w21875>.
- Loder, E., T. Groves, and D. MacAuley. 2010. "Registration of Observational Studies: The next Step towards Research Transparency." *BMJ* 340 (feb18 2): c950–c950. doi:10.1136/bmj.c950.
- Longhi, Simonetta, Peter Nijkamp, and Jacques Poot. 2005. "A Meta-Analytic Assessment of the Effect of Immigration on Wages." *Journal of Economic Surveys* 19 (3): 451–77. doi:10.1111/j.0950-0804.2005.00255.x.
- Longo, Dan L., and Jeffrey M. Drazen. 2016. "Data Sharing." *New England Journal of Medicine* 374 (3): 276–77. doi:10.1056/NEJMe1516564.
- Lovell, Michael C. 1983. "Data Mining." *The Review of Economics and Statistics* 65 (1): 1–12. doi:10.2307/1924403.
- Lundqvist, Heléne, Matz Dahlberg, and Eva Mörk. 2014. "Stimulating Local Public Employment: Do General Grants Work?" *American Economic Journal: Economic Policy* 6 (1): 167–92. doi:10.1257/pol.6.1.167.
- MacCoun, Robert, and Saul Perlmutter. 2015. "Blind Analysis: Hide Results to Seek the Truth." *Nature* 526 (7572): 187–89. doi:10.1038/526187a.
- Maggioni, Aldo P., Bernadette Darne, Dan Atar, Eric Abadie, Bertram Pitt, and Faiez Zannad. 2007. "FDA and CPMP Rulings on Subgroup Analyses." *Cardiology* 107 (2): 97–102. doi:10.1159/000094508.
- Maley, Steve. 2014. "Statistics Show No Evidence of Gender Bias in the Public's Hurricane Preparedness." *Proceedings of the National Academy of Sciences* 111 (37): E3834–E3834. doi:10.1073/pnas.1413079111.

- Malin, Bradley, Kathleen Benitez, and Daniel Masys. 2011. "Never Too Old for Anonymity: A Statistical Standard for Demographic Data Sharing via the HIPAA Privacy Rule." *Journal of the American Medical Informatics Association* 18 (1): 3–10. doi:10.1136/jamia.2010.004622.
- Malter, Daniel. 2014. "Female Hurricanes Are Not Deadlier than Male Hurricanes." *Proceedings of the National Academy of Sciences* 111 (34): E3496–E3496. doi:10.1073/pnas.1411428111.
- Mansfield, Edwin, Mark Schwartz, and Samuel Wagner. 1981. "Imitation Costs and Patents: An Empirical Study." *The Economic Journal* 91 (364): 907–18. doi:10.2307/2232499.
- Maringer, Marcus, and Diederik A. Stapel. 2009. "Retracted: Correction or Comparison? The Effects of Prime Awareness on Social Judgments." *European Journal of Social Psychology* 39 (5): 719–33. doi:10.1002/ejsp.569.
- Mathieu S, Boutron I, Moher D, Altman DG, and Ravaud P. 2009. "Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials." *JAMA* 302 (9): 977–84. doi:10.1001/jama.2009.1242.
- McAleer, Michael, Adrian R. Pagan, and Paul A. Volker. 1985. "What Will Take the Con Out of Econometrics?" *The American Economic Review* 75 (3): 293–307.
- McCloskey, Deirdre N., and Stephen T. Ziliak. 1996. "The Standard Error of Regressions." *Journal of Economic Literature* 34 (1): 97–114.
- McCrary, Justin, Garret Christensen, and Daniele Fanelli. 2015. "Conservative Tests under Satisficing Models of Publication Bias."
- McCullough, B. D. 2007. "Got Replicability? The Journal of Money, Credit and Banking Archive." *Econ Journal Watch* 4 (3): 326–37.
- . 2009. "Open Access Economics Journals and the Market for Reproducible Economic Research." March. <http://search.informit.com.au/documentSummary;dn=775889276720076;res=IELBUS>.
- McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison. 2006. "Lessons from the JMCB Archive." *Journal of Money, Credit, and Banking* 38 (4): 1093–1107. doi:10.1353/mcb.2006.0061.
- McCullough, B. D., and H. D. Vinod. 2003. "Verifying the Solution from a Nonlinear Solver: A Case Study." *The American Economic Review* 93 (3): 873–92.
- McCullough, B. D., and H. D. Vinod. 2004a. "Verifying the Solution from a Nonlinear Solver: A Case Study: Reply." *American Economic Review* 94 (1): 391–96. doi:10.1257/000282804322970887.
- . 2004b. "Verifying the Solution from a Nonlinear Solver: A Case Study: Reply." *American Economic Review* 94 (1): 400–403. doi:10.1257/000282804322970904.
- McCullough, B.d., Kerry Anne McGeary, and Teresa D. Harrison. 2008. "Do Economics Journal Archives Promote Replicable Research?" *Canadian Journal of Economics/Revue Canadienne D'économique* 41 (4): 1406–20. doi:10.1111/j.1540-5982.2008.00509.x.
- McManus, Walter S. 1985. "Estimates of the Deterrent Effect of Capital Punishment: The Importance of the Researcher's Prior Beliefs." *Journal of Political Economy* 93 (2): 417–25.

- McNutt, M. 2016. "Taking up TOP." *Science* 352 (6290): 1147–1147.
doi:10.1126/science.aag2359.
- McNutt, Marcia. 2015. "Editorial Retraction." *Science* 348 (6239): 1100–1100.
doi:10.1126/science.aac6638.
- Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.
- Mervis, Jeffrey. 2014a. "How Two Economists Got Direct Access to IRS Tax Records." *Science / AAAS*. May 20. <http://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records>.
- . 2014b. "Why Null Results Rarely See the Light of Day." *Science* 345 (6200): 992–992. doi:10.1126/science.345.6200.992.
- Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31. doi:10.1126/science.1245317.
- Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159–217.
- Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112 (4): 725–53. doi:10.1086/jpe.2004.112.issue-4.
- Mittelstaedt, Robert A., and Thomas S. Zorn. 1984. "ECONOMETRIC REPLICATION: LESSONS FROM THE EXPERIMENTAL SCIENCES." *Quarterly Journal of Business and Economics* 23 (1): 9–15.
- Mizon, Grayham E., and Jean-Francois Richard. 1986. "The Encompassing Principle and Its Application to Testing Non-Nested Hypotheses." *Econometrica* 54 (3): 657–78.
doi:10.2307/1911313.
- Moffitt, Robert A. 2016. "In Defense of the NSF Economics Program." *Journal of Economic Perspectives* 30 (3): 213–34. doi:10.1257/jep.30.3.213.
- Moher D, Jones A, Lepage L, and for the CONSORT Group. 2001. "Use of the Consort Statement and Quality of Reports of Randomized Trials: A Comparative before-and-after Evaluation." *JAMA* 285 (15): 1992–95. doi:10.1001/jama.285.15.1992.
- Moher, David, Kenneth F. Schulz, and Douglas G. Altman. 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel Group Randomized Trials." *BMC Medical Research Methodology* 1 (1): 2. doi:10.1186/1471-2288-1-2.
- Monogan, James E. 2013. "A Case for Registering Studies of Political Outcomes: An Application in the 2010 House Elections." *Political Analysis* 21 (1): 21–37.
doi:10.1093/pan/mps022.
- Montgomery, Paul, Sean Grant, Sally Hopewell, Geraldine Macdonald, David Moher, Susan Michie, and Evan Mayo-Wilson. 2013. "Protocol for CONSORT-SPI: An Extension for Social and Psychological Interventions." *Implementation Science* 8 (1): 99.
doi:10.1186/1748-5908-8-99.
- Mookerjee, Rajen. 2006. "A Meta-Analysis of the Export Growth Hypothesis." *Economics Letters* 91 (3): 395–401. doi:10.1016/j.econlet.2005.12.020.
- Morey, Richard D., Christopher D. Chambers, Peter J. Etchells, Christine R. Harris, Rink Hoekstra, Daniel Lakens, Stephan Lewandowsky, et al. 2015. "The Peer Reviewers'

- Openness Initiative: Incentivising Open Research Practices through Peer Review.” http://opennessinitiative.org/PRO_Initiative_RSOS.pdf.
- Necker, Sarah. 2014. “Scientific Misbehavior in Economics.” *Research Policy* 43 (10): 1747–59. doi:10.1016/j.respol.2014.05.002.
- Neumark, David. 2001. “The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design.” *Industrial Relations: A Journal of Economy and Society* 40 (1): 121–44. doi:10.1111/0019-8676.00199.
- Neumark, David, J. M. Ian Salas, and William Wascher. 2014. “Revisiting the Minimum Wage—Employment Debate: Throwing Out the Baby with the Bathwater?” *Industrial & Labor Relations Review* 67 (3 suppl): 608–48. doi:10.1177/00197939140670S307.
- Neumark, David, and William Wascher. 1998. “Is the Time-Series Evidence on Minimum Wage Effects Contaminated by Publication Bias?” *Economic Inquiry* 36 (3): 458–70.
- . 2000. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Comment.” *The American Economic Review* 90 (5): 1362–96.
- Nijkamp, Peter, and Jacques Poot. 2005. “The Last Word on the Wage Curve?” *Journal of Economic Surveys* 19 (3): 421–50. doi:10.1111/j.0950-0804.2005.00254.x.
- Nofsinger, John R. 2009. “Retraction Notice to ‘Social Mood: The Stock Market and Political Cycles’ [J. Socio-Econ. 36 (2007) 734–744].” *The Journal of Socio-Economics* 38 (3): 547. doi:10.1016/j.socec.2009.01.001.
- Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422–25. doi:10.1126/science.aab2374.
- Nosek, Brian A., and Daniël Lakens. 2014. “Registered Reports.” *Social Psychology* 45 (3): 137–41. doi:10.1027/1864-9335/a000192.
- Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. “Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability.” *Perspectives on Psychological Science* 7 (6): 615–31. doi:10.1177/1745691612459058.
- “Notice to Our Readers.” 1984. *The Quarterly Journal of Economics* 99 (2): 383–84.
- O’Brien, Peter C. 1984. “Procedures for Comparing Samples with Multiple Endpoints.” *Biometrics* 40 (4): 1079–87. doi:10.2307/2531158.
- Olken, Benjamin A. 2015. “Promises and Perils of Pre-Analysis Plans.” *The Journal of Economic Perspectives* 29 (3): 61–80. doi:10.1257/jep.29.3.61.
- Olken, Benjamin A., Junko Onishi, and Susan Wong. 2012. “Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia.” Working Paper 17892. National Bureau of Economic Research. <http://www.nber.org/papers/w17892>.
- Open Science Collaboration. 2012. “An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science.” *Perspectives on Psychological Science* 7 (6): 657–60. doi:10.1177/1745691612462588.
- . 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. doi:10.1126/science.aac4716.
- Pagan, Adrian. 1987. “Three Econometric Methodologies: A Critical Appraisal1.” *Journal of Economic Surveys* 1 (1–2): 3–23. doi:10.1111/j.1467-6419.1987.tb00022.x.

- Patil, Prasad, Roger D. Peng, and Jeffrey T. Leek. 2016. "What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science." *Perspectives on Psychological Science* 11 (4): 539–44. doi:10.1177/1745691616646366.
- Pesaran, Hashem. 2003. "Introducing a Replication Section." *Journal of Applied Econometrics* 18 (1): 111–111. doi:10.1002/jae.709.
- Phillips, P.c.b. 1988. "Reflections on Econometric Methodology*." *Economic Record* 64 (4): 344–59. doi:10.1111/j.1475-4932.1988.tb02075.x.
- "Redundant Publishing – Australasian Journal of Regional Studies." 2011. *Regional Studies* 45 (2): 282–282. doi:10.1080/00343404.2011.554207.
- "Retraction Statement and Authors' Apology." 2009. *Regional Studies* 43 (1): 156–156. doi:10.1080/00343400802667723.
- "Retraction Statement: 'Correction or Comparison? The Effects of Prime Awareness on Social Judgments', by M. Maringer and D. Stapel." 2015. *European Journal of Social Psychology*, January, n/a-n/a. doi:10.1002/ejsp.2173.
- Roberts, Colin J. 2005. "Issues in Meta-Regression Analysis: An Overview." *Journal of Economic Surveys* 19 (3): 295–98. doi:10.1111/j.0950-0804.2005.00248.x.
- Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2008. "Control of the False Discovery Rate under Dependence Using the Bootstrap and Subsampling." *TEST* 17 (3): 417. doi:10.1007/s11749-008-0126-6.
- Rose, Andrew K., and T. D. Stanley. 2005. "A Meta-Analysis of the Effect of Common Currencies on International Trade*." *Journal of Economic Surveys* 19 (3): 347–65. doi:10.1111/j.0950-0804.2005.00251.x.
- Rosenthal, Robert. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86 (3): 638–41. doi:10.1037/0033-2909.86.3.638.
- Sachs, Jeffrey, and Andrew Warner. 1997. "Natural Resource Abundance and Economic Growth." Center for International Development and Harvard Institute for International Development. http://www.cid.harvard.edu/ciddata/warner_files/natresf5.pdf.
- Sala-i-Martin, Xavier, Gernot Doppelhofer, and Ronald I. Miller. 2004. "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach." *The American Economic Review* 94 (4): 813–35.
- Sala-I-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *The American Economic Review* 87 (2): 178–83.
- Schulz, Kenneth F., Douglas G. Altman, David Moher, and \$author firstName \$author.lastName. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMC Medicine* 8 (1): 18. doi:10.1186/1741-7015-8-18.
- Schulz, Kenneth F., and David A Grimes. 2005. "Multiplicity in Randomised Trials II: Subgroup and Interim Analyses." *The Lancet* 365 (9471): 1657–61. doi:10.1016/S0140-6736(05)66516-6.
- Schwabish, Jonathan A. 2014. "An Economist's Guide to Visualizing Data." *The Journal of Economic Perspectives* 28 (1): 209–33. doi:10.1257/jep.28.1.209.
- Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. doi:10.1214/aos/1176344136.

- Shachar, Ron, and Barry Nalebuff. 2004. "Verifying the Solution from a Nonlinear Solver: A Case Study: Comment." *American Economic Review* 94 (1): 382–90. doi:10.1257/000282804322970878.
- Shen, Helen. 2014. "Interactive Notebooks: Sharing the Code." *Nature* 515 (7525): 151–52. doi:10.1038/515151a.
- Silberzahn, Raphael, and Eric L. Uhlmann. 2015. "Crowdsourced Research: Many Hands Make Tight Work." *Nature* 526 (7572): 189–91. doi:10.1038/526189a.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.
- . 2012. "A 21 Word Solution." *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology* 26 (2): 4–7.
- Simonsohn, Uri. 2013. "Just Post It The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone." *Psychological Science* 24 (10): 1875–88. doi:10.1177/0956797613480366.
- . 2015. "Small Telescopes Detectability and the Evaluation of Replication Results." *Psychological Science* 26 (5): 559–69. doi:10.1177/0956797614567341.
- . 2017. "[58] The Funnel Plot Is Invalid Because of This Crazy Assumption: $r(n,d)=0$." *Data Colada*. March 21. <http://datacolada.org/58>.
- Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014a. "P-Curve: A Key to the File-Drawer." *Journal of Experimental Psychology: General* 143 (2): 534–47. doi:10.1037/a0033242.
- . 2014b. "P-Curve and Effect Size Correcting for Publication Bias Using Only Significant Results." *Perspectives on Psychological Science* 9 (6): 666–81. doi:10.1177/1745691614553988.
- Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2015a. "Better P-Curves: Making P-Curve Analysis More Robust to Errors, Fraud, and Ambitious P-Hacking, a Reply to Ulrich and Miller (2015)." *Journal of Experimental Psychology: General* 144 (6): 1146–52. doi:10.1037/xge0000104.
- . 2015b. "Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications." SSRN Scholarly Paper ID 2694998. Rochester, NY: Social Science Research Network. <http://papers.ssrn.com/abstract=2694998>.
- Sinisi, Sandra E., Eric C Polley, Maya L Petersen, Soo-Yon Rhee, and der Laan Mark J. van. 2007. "Super Learning: An Application to the Prediction of HIV-1 Drug Resistance." *Statistical Applications in Genetics and Molecular Biology* 6 (1). doi:10.2202/1544-6115.1240.
- Siskind, Frederic B. 1977. "Minimum Wage Legislation in the United States: Comment." *Economic Inquiry* 15 (1): 135–38. doi:10.1111/j.1465-7295.1977.tb00457.x.
- Stanley, T. D. 2005a. "Beyond Publication Bias." *Journal of Economic Surveys* 19 (3): 309–45. doi:10.1111/j.0950-0804.2005.00250.x.
- . 2005b. "Beyond Publication Bias." *Journal of Economic Surveys* 19 (3): 309–45. doi:10.1111/j.0950-0804.2005.00250.x.
- . 2008. "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection*." *Oxford Bulletin of Economics and Statistics* 70 (1): 103–27. doi:10.1111/j.1468-0084.2007.00487.x.

- Stanley, T. D., and Hristos Doucouliagos. 2012. *Meta-Regression Analysis in Economics and Business*. Routledge.
- Stanley, T.d., and Hristos Doucouliagos. 2010. "Picture This: A Simple Graph That Reveals Much Ado About Research." *Journal of Economic Surveys* 24 (1): 170–91. doi:10.1111/j.1467-6419.2009.00593.x.
- Stanley, Tom, Hristos Doucouliagos, Margaret Giles, Jost H. Heckemeyer, Robert J. Johnston, Patrice Laroche, Jon P. Nelson, et al. 2013. "Meta-Analysis of Economics Research Reporting Guidelines." *Journal of Economic Surveys* 27 (2): 390–94. doi:10.1111/joes.12008.
- "Statement of Retraction." 2010. *Journal of Economic Policy Reform* 13 (4): 387–387. doi:10.1080/17487870.2010.533897.
- . "Statement of Retraction." 2012. *Applied Economics Letters* 19 (16): 1649–1649. doi:10.1080/13504851.2011.639572.
- Steen, R. Grant. 2010. "Retractions in the Scientific Literature: Is the Incidence of Research Fraud Increasing?" *Journal of Medical Ethics*, December, jme.2010.040923. doi:10.1136/jme.2010.040923.
- Steen, R. Grant, Arturo Casadevall, and Ferric C. Fang. 2013. "Why Has the Number of Scientific Retractions Increased?" *PLOS ONE* 8 (7): e68397. doi:10.1371/journal.pone.0068397.
- Sterling, Theodore D. 1959. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa." *Journal of the American Statistical Association* 54 (285): 30–34. doi:10.1080/01621459.1959.10501497.
- Stodden, Victoria, Friedrich Leisch, and Roger D Peng. 2014. *Implementing Reproducible Research*. CRC Press.
- Sullivan, Ryan, Allan Timmermann, and Halbert White. 1999. "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap." *The Journal of Finance* 54 (5): 1647–91. doi:10.1111/0022-1082.00163.
- Sweeney, Latanya. 2002. "K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05): 557–70. doi:10.1142/S0218488502001648.
- Taichman DB, Backus J, Baethge C, and et al. 2016. "Sharing Clinical Trial Data: A Proposal from the International Committee of Medical Journal Editors." *JAMA* 315 (5): 467–68. doi:10.1001/jama.2015.18164.
- Taubman, Sarah L., Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. 2014. "Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment." *Science* 343 (6168): 263–68. doi:10.1126/science.1246183.
- The International Consortium of Investigators for Fairness in Trial Data Sharing. 2016. "Toward Fairness in Data Sharing." *New England Journal of Medicine* 375 (5): 405–7. doi:10.1056/NEJMp1605654.
- The Lancet. 2010. "Should Protocols for Observational Research Be Registered?" *The Lancet* 375 (9712): 348. doi:10.1016/S0140-6736(10)60148-1.
- Thomas, Duncan, Elizabeth Frankenberg, Jed Friedman, Jean-Pierre Habicht, Mohammed Hakimi, Nicholas Ingwersen, Nathan Jones, et al. 2006. "Causal Effect of Health on Labor Market Outcomes: Experimental Evidence." *California Center for Population Research*.

- Thomas, Duncan, Elizabeth Frankenberg, Jed Friedman, Jean-Pierre Habicht, Mohammed Hakimi, Nathan Jones Jaswadi, Gretel Pelto, et al. 2003. "Iron Deficiency and the Well-Being of Older Adults: Early Results from a Randomized Nutrition Intervention." In *Population Association of America Annual Meetings, Minneapolis*.
- Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Graphics Press.
- Turner, Erick H., Annette M. Matthews, Eftihia Linardatos, Robert A. Tell, and Robert Rosenthal. 2008. "Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy." *New England Journal of Medicine* 358 (3): 252–60. doi:10.1056/NEJMsa065779.
- Ulrich, Rolf, and Jeff Miller. 2015. "P-Hacking by Post Hoc Selection with Multiple Opportunities: Detectability by Skewness Test?: Comment on Simonsohn, Nelson, and Simmons (2014)." *Journal of Experimental Psychology: General* 144 (6): 1137–45. doi:10.1037/xge0000086.
- Viscusi, W. Kip. 2015. "The Role of Publication Selection Bias in Estimates of the Value of a Statistical Life." *American Journal of Health Economics* 1 (1): 27–52. doi:10.1162/AJHE_a_00002.
- Vivalt, Eva. 2015. "The Trajectory of Specification Searching Across Methods and Disciplines." <http://evavivalt.com/wp-content/uploads/2015/09/Trajectory-of-Specification-Searching.pdf>.
- Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. 2004. "Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies." *Journal of the National Cancer Institute* 96 (6): 434–42. doi:10.1093/jnci/djh075.
- Walsh, Elias, Sarah Dolfin, and John DiNardo. 2009. "Lies, Damn Lies, and Pre-Election Polling." *American Economic Review* 99 (2): 316–22. doi:10.1257/aer.99.2.316.
- Warren, Elizabeth. 2016. "Strengthening Research through Data Sharing." *New England Journal of Medicine* 375 (5): 401–3. doi:10.1056/NEJMp1607282.
- Welch, Finis. 1974. "Minimum Wage Legislation in the United States*." *Economic Inquiry* 12 (3): 285–318. doi:10.1111/j.1465-7295.1974.tb00401.x.
- . 1977. "Minimum Wage Legislation in the United States: Reply." *Economic Inquiry* 15 (1): 139–42. doi:10.1111/j.1465-7295.1977.tb00458.x.
- Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons.
- White, Halbert. 2000. "A Reality Check for Data Snooping." *Econometrica* 68 (5): 1097–1126. doi:10.1111/1468-0262.00152.
- Wydick, Bruce, Elizabeth Katz, and Brendan Janet. 2014. "Do in-Kind Transfers Damage Local Markets? The Case of TOMS Shoe Donations in El Salvador." *Journal of Development Effectiveness* 6 (3): 249–67. doi:10.1080/19439342.2014.919012.
- Xie, Yihui. 2013. *Dynamic Documents with R and Knitr*. CRC Press.
- . 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Research*, 3–32. CRC Press.
- Ziliak, Stephen T., and Deirdre N. McCloskey. 2004. "Size Matters: The Standard Error of Regressions in the American Economic Review." *The Journal of Socio-Economics, Statistical Significance*, 33 (5): 527–46. doi:10.1016/j.socec.2004.09.024.
- Ziliak, Stephen Thomas, and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press.

Zimmermann, Christian. 2015. "On the Need for a Replication Journal." *Working Paper Series*, no. No. 2015-016 (August). <http://research.stlouisfed.org/wp/more/2015-016/>.