6 Transparency and Reproducibility: Conceptualizing the Problem

Garret Christensen and Edward Miguel*

Openness and transparency have long been considered key pillars of the scientific ethos (Merton 1973b). 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 chapter, 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 for some time (for instance, in economics 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 chapter is to survey this emerging literature on research transparency and reproducibility and synthesize the evidence of the problems. Awareness of these issues has come to the fore in economics (Brodeur et al. 2016), 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). In our next chapter (Chapter 7) we discuss productive avenues for future work and potential solutions.

With the vastly greater computing power of recent decades and the ability to run a nearly infinite number of regressions (Sala-i-Martin 1997), there is renewed concern that null-hypothesis statistical testing is subject to both conscious and unconscious manipulation. At the same time, technological

* Work similar to Chapters 6 and 7 was previously published by the authors as “Transparency, Reproducibility, and the Credibility of Economics Research” in the *Journal of Economic Literature*, 2018, 56(3), 920–980. Any opinions and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the US Census Bureau.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 social science research. Yet, as we discuss below, the progress to date is partial, with some journals and fields in the social sciences adopting new practices to promote transparency and reproducibility and many others not (yet) doing so.

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

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).[1]

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 relationships to no relationships commonly

[1] 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).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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:

$$\mathrm{PPV}_i = (1-\beta)R_i/((1-\beta)R_i + \alpha) \qquad \text{(eqn. 1)}$$

Clearly, the better powered the study, and the stricter the statistical significance level, the closer the PPV is to 1, 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 one-third of all hypotheses are thought to be true ex ante $(R_i = 0.5)$, the PPV is relatively high at 89 percent, a level that would not seem likely to threaten the validity of research in a particular subfield.

However, reality is considerably messier than this best-case scenario and, as Ioannidis describes, this could lead to much higher 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 (2017), 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 83 percent, which is beginning to potentially look like more of a concern. For power = 0.20, fully 33 percent 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 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:

$$\mathrm{PPV}_i = ((1-\beta)R_i + u_i\beta R_i)/((1-\beta)R_i + \alpha + u_i\beta R_i + u_i(1-\alpha)) \qquad \text{(eqn. 2)}$$

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-percent power, $R_i = 0.5$, and the usual 5-percent significance level, but now assume that author bias is low at 10 percent, the PPV falls from 83 to 65 percent. If 30 percent of authors are biased in their presentation of results, the PPV drops dramatically to 49 percent, meaning that nearly half 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 6.1 (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 at 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Table 6.1 Predictive value of research findings

1–β	*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

Note: Positive predictive value (PPV) of research findings for various combinations of power (1 – ß), ratio of true to not-true relationships (*R*), and researcher bias (*u*). The estimated PPVs are derived assuming α = 0.05 for a single study. RCT: randomized controlled trial.
Source: Reproduced from table 4 of Ioannidis (2005).

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 across the social sciences.

First off, literatures characterized by statistically underpowered (i.e., small 1 – β) studies are likely to have many false positives. A study may be underpowered 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 due to both 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 Chapter 7 turn to potential ways to address them.

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-1 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).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Cleverly, the authors are able to look inside the file drawer through their access to the universe of studies that passed peer review and were included in a nationally representative social science survey, namely, the NSF-funded Time-sharing Experiments in the Social Sciences, or TESS.[2] TESS funded studies across research fields, including in economics, e.g., Walsh, Dolfin, 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 of 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 paper form. It implies that the 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.

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 to 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.

[2] See http://tessexperiments.org.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 and 20 percent 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 $\left(1.96 - X\%, 1.96\right]$ to the number in $\left(1.96, 1.96 + X\%\right]$, where X is the size of the caliper, and they examine these differences at 5-, 10-, 15-, and 20-percent critical values.[3]

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 2008b). 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 2008a).

Psychologists have recently developed a related tool called the "p-curve," describing the density of reported p-values in a literature, which 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,

[3] 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).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Nelson, and Simmons 2014a). Intuitively, under the null of no effect, a p-value < 0.08 should occur 8 percent of the time, a *p*-value < 0.07 occurs 7 percent 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 percent 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).[4] 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 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 percent, 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).[5]

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 percent) of trials with positive, i.e., statistically significant, results were published, while only 8 of 24 (33 percent)

[4] 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 < 0.05, and this is the standard used throughout (Simonsohn, Nelson, and Simmons 2014a).

[5] 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) and Simonsohn, Simmons, and Nelson (2015a).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 (2016) 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 higher than 3 (rather than 1.96) to infer statistical significance at 95-percent confidence. They note that approximately 30 percent 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 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 (2016).

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 (discussed in more detail below), which is closely related to what the Ioannidis (2005) model calls author bias.

Publication Bias in Several Empirical Economics Literatures

Scholars in economics have argued that there is considerable publication bias in several specific literatures including labor economics research on minimum-wage impacts and on the value of a statistical life. 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

generally produce more precise estimates (for a given research design).[6] 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 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 the 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-percent confidence level, and other tests, described in Chapter 7, 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 reduces the estimates of VSL by 70–80 percent 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

[6] 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.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

(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 6.2 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 6.2), 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.

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 6.2. One common data visualization approach is the use of funnel graphs; see Stanley and Doucouliagos (2010), Light and Pillemer (1984), and our discussion in Chapter 7.

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 underreporting of effect magnitudes is still a concern. In a review in the *Journal of Economic Literature*, McCloskey and Ziliak (1996) find that 70 percent of full-length *American Economic Review* articles did not distinguish between statistical and practical significance. Follow-up reviews in 2004 and 2008 conclude

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Table 6.2 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 percent 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.
Doucouliagos, 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.
Doucouliagos 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 percent 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.

(*continued*)

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Table 6.2 (*cont.*)

Paper	Topic	Publication Bias?	Papers (Estimates) Used	Notes
Doucouliagos (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 percent.
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 percent) 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.
Doucouliagos 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.

Note: 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 is 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.

that the situation had not meaningfully improved (Ziliak and McCloskey 2004, 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. 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 economic hypotheses are indeed false. They also conclude that the null results that actually are 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 economic hypotheses 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.

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 1980s and 1990s, 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 notions that 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, US 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

included (among state median income, unemployment, percent population non-white, 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.[7]

Since the 1980s, empirical research practices in economics have changed significantly, especially with regards to improvements in research design.

[7] 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, 2013); McCullough and Vinod (2003) contains a more general discussion of robustness and replication failures in nonlinear maximization methods.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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), which 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 $1-\left(1-\hat{\alpha}\right)^{c/k}$. 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 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 (subgroup 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 significance level

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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.

Subgroup Analysis

One area of analytical flexibility that appears particularly important in practice is subgroup 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).

Inability to Replicate Results

Data Availability

There have been longstanding concerns within economics and the social sciences 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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.[8] Despite the adoption of an explicit policy of data sharing by the JMCB during the project, only 78 percent of authors provided data within six months after multiple requests, although this was certainly an improvement over the 34-percent 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, 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.

The call to share data was echoed in sociology (Hauser 1987), but 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

[8] 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 www.nsf.gov/bfa/dias/policy/dmp.jsp.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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.[9]

In addition to all the journals published by the American Economic Association (including the *American Economic Review*, 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*, the *Journal of Political Economy*, the *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 2016.[10]

Table 6.3 summarizes journal policies regarding data sharing, publication of replications or comments, and funding or conflict-of-interest disclosures at 12 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."[11] There is also greater use of disclosure statements. A similar, if dated, review of journal policies in political science is available in Bueno de Mesquita et al. (2003).

[9] www.aeaweb.org/aer/data.php.

[10] www.oxfordjournals.org/our_journals/qje/for_authors/data_policy.html.

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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Table 6.3 Transparency policies at selected top economics and finance journals

Journal	Data Sharing Policy?	Notes	Replication/ Comment Publication?	Notes	Funding/ Conflict-of-Interest Disclosure?	Notes
American Economic Review	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.
American Economic Journals (Applied Economics; Economic Policy; Macroeconomics)	Yes	Same as *AER*. Since journal inception in 2009.	Yes	Allow post-publication peer review on website.	Yes	Same as *AER*.
Econometrica	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.
Journal of Finance	No		Yes		Yes	Current policy adopted August 2015.
Journal of Financial Economics	No	Some data are available on the journal webpage, but there appears to be no official policy.	No		Yes	Current policy adopted November 2015.

(*continued*)

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Table 6.3 *(cont.)*

Journal	Data Sharing Policy?	Notes	Replication/ Comment Publication?	Notes	Funding/ Conflict-of-Interest Disclosure?	Notes
Journal of Political Economy	Yes	Uses the same policy as the *AER*. Announced in 2005, effective in 2006.	Yes	Submission instructions state that authors of comments must correspond with original authors.	No	
Quarterly Journal of Economics	Yes	Uses the same policy as the *AER*, adopted 2016.	Yes		Yes	
Review of Economic Studies	Yes	Start date unclear.	No		No	
Review of Financial Studies	No		Yes		Yes	Adopted August 2006. Updated June 2016.

Note: 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 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 produce 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.

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, other 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; 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 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 party.[12] 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 30 political science journals.[13] 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.[14] In sociology, Freese (2007a, b) issued a call for American Sociological Association journals to take

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

[13] See www.dartstatement.org/.

[14] More information on badges can be found here: 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: www.psychologicalscience.org/index.php/publications/observer/obsonline/open-practice-badges-in-psychological-science-18-months-out.html.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

advantage of new technology (the Internet) and require data sharing at the time of publication, as well as a defense against objections concerning subject confidentiality and incentives for original data gathering, among others.

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 6.3 – 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 and 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 6.1 presents the percentage of papers in each issue of the *AER* since 2005 (when information became available) through 2017. A few patterns are noteworthy. First, the proportion of papers that include data has risen over time, starting at roughly 60 percent and since increasing into the 70–80 percent range, capturing the shift toward 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 percent over time. Thus, replication data are 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 are US government data. There are currently 29 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).[15] 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

[15] 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.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

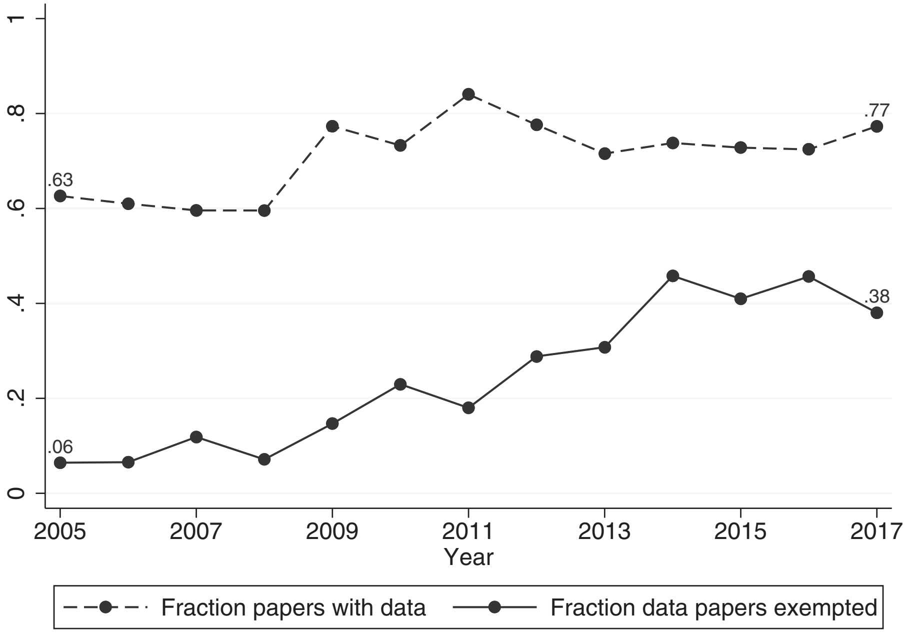

Figure 6.1 AER papers with data exempt from the data-sharing requirement

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 are taken from the Annual Report of the Editors, which appears annually in the Papers and Proceedings issue of the AER. Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

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 should be 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 be

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 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.[16]

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 of the publication of an article (Taichman et al. 2016) an editorial in the leading *New England Journal of Medicine* caused an outcry when the editors 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 its 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

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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

protection."[17] We lack data on how often this exemption is granted, however. Additionally, this journal goes much further than economics 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.

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 PhD 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

[17] See https://ajps.org/ajps-replication-policy/.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 2017) 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.

Evidence on replication in economics. The articles in the 1986 *Journal of Money Credit and Banking* project and the 1994 St. Louis Federal Reserve Bank conference 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. 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-percent power to detect the original effect size at the 5-percent significance level. In all, the estimated effects were statistically significant with the same sign in 11 of the 18 replication studies (61.1 percent). 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 percent), estimated effects lie within a 95-percent "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 percent.[18] Overall,

[18] 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 et al. study (Bohannon 2016). One of the replicated studies

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

it is reasonable 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 PhD students and post-doctoral researchers, as well as professors, recruited via email) 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 US data, and have a key result based on US GDP figures. Of the 67 papers, 6 use proprietary data and are thus excluded from consideration; 35 articles are published in journals with data and code sharing requirements, but Chang and Li could obtain data for only 28 of these (80 percent) from the journal archives, suggesting limited enforcement of this requirement in many cases. Web search and emails 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 percent).

With these data in hand, the overall replication success rate is 29 of 67 (43 percent) overall, or 29 of 61 (48 percent) 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,

(Ifcher and Zarghamee 2011) originally showed subjects a clip of comedian Robin Williams 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.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

and this appears to lead to a fairly generous interpretation 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 are available (which they sometimes are 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 (2017) 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.

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 percent. 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

44 percent) (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 (2017) contains a larger selection of examples (see his Table 3).[19] 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 in the first place, especially for these earlier articles published before journal data sharing requirements were established.

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-percent sample of the census, while a different 5-percent 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., US 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 US college undergraduates the population (and do not differentiate by campus), or Amazon MTurk-ers,

[19] 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 were updated, as sometimes happens with US government data. The work of one of the authors of this chapter 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.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 versus 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).

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, 2002), Albouy (2012), on institutions and GDP growth with settler mortality as an instrumental variable. In sociological research related to the evolutionary psychological theory of parental investment (the Trivers-Willard hypothesis), a similar back-and-forth can be seen in Kanazawa (2001) and Freese and Powell (2001). In sociological work on the returns to education in urban China, Jann (2005) reanalyzes Wu and Xie (2003) with what he considers a better statistical test, showing that the earlier conclusions are premature.[20]

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 Chapter 7, all the more important.

[20] Additional examples from sociology include Roth and Kroll (2007), which reanalyzes earlier work by Miller and Stark (2002) on risk preference explanations for gender differences in religiosity.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

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 cyclicality of teenage employment with respect to the business cycle, and shifted teenage employment toward 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 methods,[21] 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 in the next chapter.

[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).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

Fraud and Retractions

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

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 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 percent of respondents admit to having at some time falsified research data, 7–10 percent of respondents admit to having committed one of four relatively minor research infractions, while up to 0.4 percent 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 percent of researchers admit to having falsified data, while 94 percent 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 US 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 percent of respondents admitting to data fabrication, and 34 percent admitting to lesser forms of academic misconduct (Fanelli 2009).

We are not aware of a recent case in economics or sociology 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

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

of research malpractice in economics. The *Journal of Economic Literature* 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 percent claimed that their journal had a formal policy on plagiarism, and 42 cases of plagiarism were discovered in an average year, with nearly 24 percent 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 4 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.

[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: www.oxfordjournals.org/our_journals/qje/for_authors/journal_policies.html.

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press

If one happened to discover the webpage of the original[25] 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 webpage[26] for Maringer and Stapel (2009), which was retracted in 2015,[27] 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 retractions in Karabag and Berggren (2012), as well as other notable recent retractions such as LaCour and Green (2014), which was retracted by Marcia 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.

Conclusion

In conclusion, we believe that the problems of publication bias, specification searching, and an inability to replicate are widespread throughout the social sciences. However, we remain optimistic, as there are numerous potential partial solutions to these problems, including study registration, pre-analysis plans, improved statistical practices such as multiple hypothesis testing adjustments, and better data sharing that we believe can help with these issues. We review these items in Chapter 7.

[25] www.aeaweb.org/articles.php?doi=10.1257/000282802762024656.

[26] http://onlinelibrary.wiley.com/doi/10.1002/ejsp.569/abstract.

[27] See "Retraction Statement: 'Correction or Comparison? The Effects of Prime Awareness on Social Judgments', by M. Maringer and D. Stapel" (2015).

https://doi.org/10.1017/9781108762519.006 Published online by Cambridge University Press