Title

Transparency and Reproducibility: Potential Solutions

Permalink

https://escholarship.org/uc/item/8pp0t4ff

ISBN

9781108486774

Authors

Christensen, Garret
Miguel, Edward

Publication Date

2020-03-31

DOI

10.1017/9781108762519.007

7 Transparency and Reproducibility: Potential Solutions

Garret Christensen and Edward Miguel

There is overwhelming evidence that the problems of publication bias, p-hacking, and a lack of reproducibility are real. The previous chapter summarizes this evidence. The published literature in sociology, political science, and economics all suffer from these problems, to varying degrees. In this chapter, we focus on several new methods and tools that have emerged in social science research over the past two decades – and more forcefully over the past ten years – to address these concerns.

These approaches have in common a focus on greater transparency and openness in the research process. They include improved research design (including experimental designs and meta-analysis approaches), study registration and pre-analysis plans, strengthened disclosure and reporting practices, and new norms regarding open data and materials.

It should be clear that these potential solutions are not panaceas, and they have not yet been adopted widely enough in the social sciences to be considered proven. Nonetheless, we strongly believe that experimenting with these new practices is worthwhile.

Improved Analytical Methods: Research Designs and Meta-Analysis

There have been a number of different responses within social sciences to the view that pervasive specification searching and publication bias was affecting the credibility of empirical literatures. As mentioned in the previous chapter, there has been a shift toward a greater focus on prospective research design in several fields of applied economics and political science work. Experimental (Duflo, Glennerster, and Kremer 2007) and quasi-experimental (Angrist and Pischke 2010) research designs arguably place more constraints on researchers relative to earlier empirical approaches, since there are natural ways to present data using these designs that researchers are typically compelled to present by colleagues in seminars and by journal referees and editors. Prospective

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

experimental studies also tend to place greater emphasis on adequately powering an analysis statistically, which may help to reduce the likelihood of publishing only false positives (Duflo, Glennerster, and Kremer 2007).

There is also suggestive evidence that the adoption of experimental and quasi-experimental empirical approaches is beginning to address some concerns about specification search and publication bias. Brodeur et al. (2016) present tentative evidence that the familiar spike in *p*-values just below the 0.05 level is less pronounced in randomized control trial studies than in studies utilizing non-experimental methods. Yet improved research design alone may not solve several other key threats to the credibility of empirical social science research, including the possibility that null or "uninteresting" findings never become known within the research community.

Understanding Statistical Model Uncertainty

In addition to improvements in research design, Leamer (1983) argued for greater disclosure of the decisions made in analysis, in what became known as "extreme bounds analysis" (described in Chapter 6). Research along these lines has dealt with model uncertainty by employing combinations of multiple models and specifications, as well as comparisons between them. Leamer himself has continued to advance this agenda (see Leamer 2016). We describe several related approaches here.

Model averaging. A natural way to deal with statistical model uncertainty is through Bayesian model averaging. In this approach, each model in the space of plausible models is assigned a probability of being true based on researcher priors and goodness of fit criteria. Averaging the resulting estimates generates a statistic incorporating model uncertainty:

$$\hat{\delta}_M = \sum_m \mu(m|D)\hat{\delta}_m, \qquad \text{(eqn. 1)}$$

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

These weights must, of course, be chosen somehow. Cohen-Cole et al. (2009), from whom we borrow the above notation, study the deterrent effect of the death penalty with a model averaging exercise combining evidence from Donohue and Wolfers (2005) and Dezhbakhsh, Rubin, and Shepherd (2003) and use the Bayesian Information Criterion (BIC) (Schwarz 1978). The weighted average they generate implies a large but imprecisely estimated

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

deterrent effect of executions on homicides in the United States. Of course, even without employing explicit probability weights, simply visualizing the distribution of estimates across the entire space of statistical models can also be quite informative on its own.

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

A discussion of model uncertainty from sociology that touches on model averaging is Young (2009), which reanalyzes the question of religiosity and economic growth from McCleary and Barro (2003) and McCleary and Barro (2006). Bayesian model averaging in sociology is also discussed in Raftery (1995) and Western (1996). Young and Holsteen (2017) develop a more formalized conception of model averaging that develops a modeling standard error as well as a measure of the size of the influence of certain covariates on the model space. Applications include estimates of the union wage premium (Hirsch 2004), mortgage lending by gender (Munnell et al. 1996), and tax-induced cross-state migration in the United States (Young and Varner 2011). Bayesian model averaging is applied to political science with examples of comparative political economy and American public opinion and policy in Bartels (1997).

Specification curve. Simonsohn, Simmons, and Nelson (2015b) propose a method, which they call the "specification curve," that is similar in spirit to Leamer's extreme-bounds analysis, but recommends researchers test the exhaustive combination of analytical decisions, not just decisions about which covariates to include in the model. If the full exhaustive set is too large to be practical, a random subset can be used. After plotting the effect size from each of the specifications, researchers can assess how much the estimated effect size varies, and which combinations of decisions lead to which outcomes.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Using permutation tests (for treatment with random assignment) or bootstrapping (for treatment without random assignment), researchers can generate shuffled samples with no true effect by construction, and compare the specification curves from these placebo samples to the specification curve from the actual data. Many comparisons are possible, but the authors suggest comparing the median effect size, the share of results with predicted sign, and share of statistically significant results with predicted sign. A key comparison, which is analogous to the traditional *p*-value, is the percent of the shuffled samples with as many or more extreme results.

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

Improved Publication Bias Tests

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

Statisticians have developed methods to estimate effect sizes in meta-analyses that control for publication bias (Hedges 1992; Hedges and Vevea 1996). The tools most widely used by economists tend to be simpler, including the widely used funnel plot, which is a scatter plot of some measure of

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

statistical precision (typically the inverse of the standard error), versus the estimated effect size. Estimates generated from smaller samples should usually form the wider base of an inverted funnel, which should be symmetric around more precise estimates in the absence of publication bias. The method is illustrated with several economics examples in Stanley and Doucouliagos (2010). In addition to scrutinizing the visual plot, a formal test of the symmetry of this plot can be conducted using data from multiple studies and regressing the relevant t-statistics on inverse standard errors:

$$t_i = \frac{\text{Estimated effect}_i}{SE_i} = \beta_0 + \beta_1\left(\frac{1}{SE_i}\right) + v_i. \qquad \text{(eqn. 2)}$$

The resulting t-test on β_0, referred to as the Funnel Asymmetry Test (FAT) (Stanley 2008), captures the correlation between estimated effect size and precision, and thus tests for publication bias.

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

Multiple Testing Corrections

Other applied econometricians have recently called for increasing the use of multiple testing corrections in order to generate more meaningful inference in study settings with many research hypotheses (Anderson 2008; Fink, McConnell, and Vollmer 2014). The practice of correcting for multiple tests is already widespread in certain scientific fields (e.g., genetics) but has yet to become the norm in the social sciences. Simply put, since we know that p-values fall below traditional significance thresholds (e.g., 0.05) purely by chance a certain proportion of the time, it makes sense to report adjusted p-values that account for the fact that we are running multiple tests, since this

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

makes it more likely that at least one of our test statistics has a significant *p*-value simply by chance.

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

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

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

Controlling the family-wise error rate. The family-wise error rate (FWER) is the probability that at least one true hypothesis in a group is rejected (a Type-1 error, or false positive). This approach is considered most useful when the "damage" from incorrectly claiming *any* hypothesis is false is high. There are several ways to implement this approach, with the simplest method being the Bonferroni correction of simply multiplying every original *p*-value by the number of tests carried out (Bland and Altman 1995), although this is extremely conservative, and improved methods have also been developed.

Holm's sequential method involves ordering *p*-values by class and multiplying the lower *p*-values by higher discount factors (Holm 1979). A related and more efficient recent method is the free step-down resampling method, developed by Westfall and Young (1993), which when implemented by anderson (2008) implies that several highly cited experimental pre-school

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

interventions (namely, the Abecedarian, Perry, and Early Training Project studies) exhibit few positive long-run impacts for males.

Another recent method improves on Holm by incorporating the dependent structure of multiple tests. Lee and Shaikh (2014) apply it to reevaluate the Mexican PROGRESA conditional cash transfer program and find that overall program impacts remain positive and significant, but are statistically significant for fewer subgroups (e.g., by gender, education) when controlling for multiple testing. List, Shaikh, and Xu (2016) propose a method of controlling the FWER for three common situations in experimental economics, namely, testing multiple outcomes, testing for heterogeneous treatment effects in multiple subgroups, and testing with multiple treatment conditions.[1]

Controlling the false discovery rate. In situations where a single Type-1 error is not considered very costly, researchers may be willing to use a somewhat less conservative method than the FWER approached discussed above, and trade off some incorrect hypothesis rejections in exchange for greater statistical power. This is made possible by controlling the false discovery rate (FDR), or the percentage of rejections that are Type-1 errors. Benjamini and Hochberg (1995) detail a simple algorithm to control this rate at a chosen level under the assumption that the p-values from the multiple tests are independent, though the same method was later shown to also be valid under weaker assumptions (Benjamini and Yekutieli 2001). Benjamini, Krieger, and Yekutieli (2006) describes a two-step procedure with greater statistical power, while Romano, Shaikh, and Wolf (2008) propose the first methods to incorporate information about the dependence structure of the test statistics.

Multiple hypothesis testing adjustments have recently been used in finance (Harvey, Liu, and Zhu 2015) to re-evaluate 316 factors from 313 different papers that explain the cross-section of expected stock returns. The authors employ the Bonferroni; Holm (1979); and Benjamini, Krieger, and Yekutieli (2006) methods to account for multiple testing, and conclude that t-statistics greater than 3.0, and possibly as high as 3.9, should be used instead of the standard 1.96, to actually conclude that a factor explains stock returns with 95-percent confidence. Index tests and both the FWER and FDR multiple testing corrections are also employed in Casey, Glennerster,

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

and Miguel (2012) to estimate the impacts of a community-driven development program in Sierra Leone using a dataset with hundreds of potentially relevant outcome variables.

Study Registration

A leading proposed solution to the problem of publication bias is the registration of empirical studies in a public registry. This would ideally be a centralized database of all attempts to conduct research on a certain question, irrespective of the nature of the results, and such that even null (not statistically significant) findings are not lost to the research community. Top medical journals have adopted a clear standard of publishing only medical trials that are registered (De Angelis et al. 2004). The largest clinical trial registry is clinicaltrials.gov, which helped to inspire the most high-profile study registry within economics, the AEA Randomized Controlled Trial Registry (Katz et al. 2013), which was launched in May 2013.[2]

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

Though it is too soon after the adoption of the AEA's trial registry to measure its impact on research practices and the robustness of empirical results, it is worth noting that the registry is already being used by many empirical researchers – since inception in 2013, over 2,060 studies conducted in over 100 countries have been registered, and the pace of registrations continues to rise rapidly. Panel A of Figure 7.1 presents the total number of registrations over time in the AEA registry (through October 2018), and Panel B shows the number of new registrations per month. A review of the projects currently included in the registry suggests that there are a particularly large number of development economics studies, which is perhaps not surprising given the widespread use of field experimental methods in contemporary development economics.

[2] The registry can be found online at: www.socialscienceregistry.org/.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

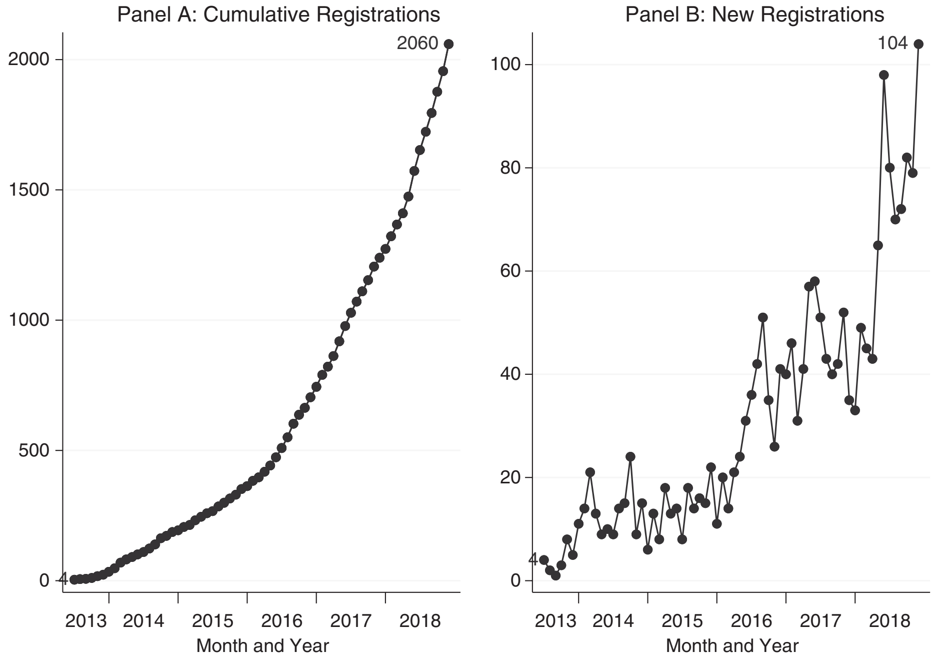

Figure 7.1 Studies in the AEA trial registry, May 2013 to October 2018

Figure shows the cumulative (Panel A) and new (Panel B) trial registrations in the American Economics Association Trial Registry (http://socialscienceregistry.org). Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

In addition to the AEA registry, several other social science registries have recently been created, including by the International Initiative for Impact Evaluation's (3ie) Registry for International Development Impact Evaluations (RIDIE, http://ridie.3ieimpact.org), launched in September 2013 (Dahl Rasmussen, Malchow-Møller, and Barnebeck andersen 2011), and the Evidence in Governance and Politics (EGAP) registry (http://egap.org/content/registration), also created in 2013. The Center for Open Science's Open Science Framework (OSF, http://osf.io) accommodates the registration of essentially any study or research document by allowing users to create a frozen time-stamped web URL with associated digital object identifier (DOI) for any materials uploaded to OSF. Several popular data storage options (including Dropbox, Dataverse, and GitHub) can also be synced with the OSF and its storage, creating a flexible way for researchers to register their research and materials. As of December 2018, researchers have posted over 281,000 searchable registrations on the OSF since the service launched in 2013.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Pre-Analysis Plans

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

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

There is ongoing discussion of what one should include in a PAP; detailed discussions include Glennerster and Takavarasha (2013), David McKenzie's World Bank Research Group blog post,[3] and a template for pre-analysis plans by Alejandro Ganimian (2014). Ganimian's template may be particularly useful to researchers themselves when developing their own pre-analysis plans, and instructors may find it useful in their courses. Building on, and modifying, the FDA's 1998 checklist with insights from these other recent treatments of pre-analysis plans, there appears to be a growing consensus that pre-analysis plans in the social sciences should consider discussing at least the following list of ten issues:

1. study design
2. study sample
3. outcome measures
4. mean effects family groupings
5. multiple hypothesis testing adjustments
6. subgroup analyses
7. direction of effect for one-tailed tests

[3] http://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

8. statistical specification and method
9. structural model
10. timestamp for verification

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

For those concerned about the possibility of “scooping” of new research designs and questions based upon a publicly posted pre-analysis plan or project description, several of the social science registries allow temporary embargoing of project details. For instance, the AEA registry allows an embargo until a specific date or project completion. At the time of writing, the OSF allows a four-year embargo until the information is made public.[4]

Examples of Pre-Analysis Plans (PAPs)

Recent examples of social science papers based on experiments with PAPs include Casey, Glennerster, and Miguel (2012) and Finkelstein et al. (2012), among others. Casey, Glennerster, and Miguel (2012) discuss evidence from a large-scale field experiment on community-driven development (CDD) projects in Sierra Leone. The project, called GoBifo, was intended to make local institutions in post-war Sierra Leone more democratic and egalitarian. GoBifo funds were spent on a variety of local public goods infrastructure (e.g., community centers, schools, latrines, roads), agriculture, and business training projects, and were closely monitored to limit leakage. The analysis finds significant short-run benefits in terms of the “hardware” aspects of infrastructure and economic well-being; the latrines were indeed built. However, a larger goal of the project, reshaping local institutions, making them more egalitarian, increasing trust, improving local collective action, and strengthening community groups, which the researchers call the “software effects,” largely failed. There are a large number of plausible outcome measures along these dimensions, hundreds in total, which the authors analyze using a mean effects index approach for nine different families of outcomes (with multiple testing adjustments). The null hypothesis of no impact cannot be rejected at 95-percent confidence for any of the nine families of outcomes.

Yet Casey et al. (2012) go on to show that, given the large numbers of outcomes in their dataset, and the multiplicity of ways to define outcome measures, finding some statistically significant results would have been relatively easy. In

[4] See http://help.osf.io/m/registrations/l/524207-embargoes.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Table 7.1 Erroneous interpretations under "cherry-picking"

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

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

fact, the paper includes an example of how, if they had had the latitude to define outcomes without a pre-analysis plan, as has been standard practice in most empirical economics studies (and in other social science fields), the authors could have reported either statistically significant and positive effects, or significantly negative effects, depending on the nature of the "cherry-picking" of results. We reproduce their results here as Table 7.1, where Panel A presents the statistically significant positive impacts identified in the GoBifo data and Panel B highlights negative effects. This finding begs the question: how many empirical social science papers with statistically significant results are, unbeknownst to us, really just some version of either Panel A or Panel B?

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

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

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

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

Other prominent early examples of economics studies that have employed pre-analysis plans include poverty targeting programs in Indonesia, an evaluation of the Toms shoe company donation program, and a job training program in Turkey, among many others (Olken, Onishi, and Wong 2012; Alatas et al. 2012; Wydick, Katz, and Janet 2014; Hirshleifer et al. 2015). The PAP tool is also spreading to other social sciences beyond economics. For instance, in psychology, a pre-specified replication of an earlier paper that had found a link between female conception risk and racial prejudice failed to find a similar effect (Hawkins, Fitzgerald, and Nosek 2015). In political science the Election Research Preacceptance Competition ran a competition for work with pre-analysis plans based on the 2016 American National

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Election Studies (ANES) data; eligible papers were required to register their analysis plan prior to the public release of the data.[5]

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

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

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

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

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

On the negative side, PAPs are often complex and take valuable time to write. Scientific breakthroughs often come at unexpected times and places,

[5] See www.erpc2016.com/.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

often as a result of exploratory analysis, and the time spent writing PAPs may thus lead less time to spend on less-structured data exploration.

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

Olken (2015) as well as Coffman and Niederle (2015) discuss another potential way to address publication bias and specification search: results-blind review. Scholars in psychology have championed this method; studies that are submitted to such review are often referred to as “registered reports” in that discipline. Authors write a detailed study protocol and pre-analysis plan, and before the experiment is actually run and data are collected, submit the plan to a journal. Journals review the plan for the quality of the design and the scientific value of the research question, and may choose to give “in-principle acceptance.” This can be thought of as a kind of revise and resubmit that is contingent on the data being collected and analyzed as planned. If the author follows through on the proposed design, and the data are of sufficiently high quality (e.g., with sufficiently low sample attrition rates in a longitudinal study, etc.), the results are to be published regardless of whether or not they are statistically significant, and whether they conform to the expectations of the editor or referees, or to the conventional wisdom in the discipline.

Dozens of journals currently have begun using results-blind review, either regularly or in special issues (Chambers 2013; Chambers et al. 2014; Nosek and Lakens 2014).[6] An issue of *Comparative Political Studies* was the first to feature results-blind review in political science (Findley et al. 2016), and it included both experimental and observational research studies.

[6] A list of journals that have adopted registered reports is available at: https://osf.io/8mpji/wiki/home/.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

In our view, it would also be useful to experiment with results-blind review and registered reports in economics journals. *The Journal of Development Economics* announced a pilot of this type of submission in March 2018.[7] The rise in experimental studies and pre-analysis plans in economics, as evidenced by the rapid growth of the AEA registry, is likely to facilitate the eventual acceptance of this approach.

Observational Studies

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

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

A major logical concern with the pre-registration of non-prospective observational studies using pre-existing data is that there is often no credible way to verify that pre-registration took place before analysis was completed, which is different than the case of prospective studies in which the data have not yet been collected or accessed. In our view, proponents of the pre-registration of observational work have not formulated a convincing response to this obvious concern.

The earliest economics study of which we are aware that used a pre-analysis plan on non-experimental data was undertaken in Neumark (2001). Based on conversations with David Levine, Alan Krueger appears to have suggested to Levine, who was the editor of the *Industrial Relations* journal at the time, that multiple researchers could analyze the employment effects of an upcoming change in the federal minimum wage with pre-specified research designs, in a bid to eliminate "author effects," and that this could create a productive "adversarial collaboration" between authors with starkly different prior views on the likely impacts of the policy change (Levine 2001).

[7] See https://blogs.worldbank.org/impactevaluations/registered-reports-piloting-pre-results-review-process-journal-development-economics.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

(The concept of adversarial collaboration – two sets of researchers with opposing theories coming together and agreeing on a way to test hypotheses before observing the data – is often associated with Daniel Kahneman; see, for example Bateman et al. 2005).

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

The verifiable "time stamp" of the federal government's release of data indeed makes this approach possible, but the situation also benefits from the depth and intensity of the minimum wage debate prior to this study. Neumark had an extensive literature to draw upon when choosing specific regression functional forms and subgroup analyses. He tests two definitions of the minimum wage, the ratio of the minimum wage to the average wage (common in Neumark's previous work) as well as the fraction of workers who benefit from the newly raised minimum wage (used in David Card's earlier work, Card 1992a and Card 1992b), and tests both models with and without controls for the employment rate of higher-skilled prime-age adults (as recommended by Deere, Murphy, and Welch 1995). The results mostly fail to reject the null hypothesis of no effect of the minimum wage increase: only 18 of the 80 specifications result in statistically significant decreases in employment (at the 90-percent confidence level), with estimated elasticities ranging from –0.14 to –0.3 for the significant estimates and others closer to zero.

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

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

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

Comparisons to other research fields. Another frontier topic in this realm is the use of pre-specified algorithms, including machine learning approaches, rather than exact pre-analysis plans for prospective studies. For instance, the exact procedure to be used to determine which covariates should be included in order to generate the most statistically precise estimates can be laid out in advance, even if those covariates are unknown (and unknowable) before the data have been collected. This approach has not yet been widely adopted in economics (to our knowledge), but has begun to be used in medical trials and biostatistics (van der Laan et al. 2007; Sinisi et al. 2007).

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

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

This distinction between the types of studies carried out by medical researchers versus economists (including those working on health topics) has a number of important implications for assessing the reliability of evidence. One has to do with the quality standards and perceptions of the risk of bias in a particular design. For medical trialists accustomed to the CONSORT standards or other medical efficacy trial reporting guidelines (described below), studies that do not feature double-blinding, and thus run the risk of endogeneous behavioral responses to the medical intervention, are considered less reliable than those studies that employ double-blinding (for a detailed discussion, see Eble, Boone, and Elbourne 2014). While a few studies conducted by economists do feature double-blinding (e.g., Thomas et al. 2003, 2006), in nearly all settings blinding participants to their status is either logistically difficult (for instance, if government partners are unwilling to distribute placebo treatments to some of their population) or even impossible.

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

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

Disclosure and Reporting Standards

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

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

require authors to include a completed CONSORT checklist at the time of article submission.[8]

Observational research in epidemiology is increasingly subject to its own set of guidelines, the so-called Strengthening the Reporting of Observational Studies in Epidemiology, or STROBE, standards (von Elm et al. 2007). In fact, developing reporting guidelines is a growth industry in medical research: at least 284 sets of guidelines have been developed for different types of health research. To deal with the proliferation of reporting standards, the Equator Network has been established to organize these guidelines and help researchers identify the most appropriate set of guidelines for their research.[9]

There are obviously very strong, and well understood, norms regarding how to report empirical results in economics studies, but there are far fewer formal guidelines or reporting checklists than in medical research. One exception is the AEA policy, announced in January 2012,[10] that its journals would require disclosure statements from authors regarding potential conflicts of interest. The AEA journals enforced the policy in July 2012, and the NBER working paper series has since adopted a similar set of required disclosures.[11] It appears the economics discipline may have been shamed into adopting these conflict-of-interest policies, at least in part, by the scathing Academy Award-winning documentary "Inside Job," which argued that some leading economists with strong (and often undisclosed) ties to the financial services industry were at least somewhat complicit in promoting policy choices that contributed to the 2008 global financial crisis (Casselman 2012).

Despite recent progress on conflict-of-interest disclosure, there has been less change within economics regarding other forms of disclosure or reporting guidelines. The only set of disclosure guideline specific to economics that we are aware of is the Consolidated Health Economic Evaluation Reporting Standards (CHEERS), although these appear to be more widely followed in health than in economics (Husereau et al. 2013). In this regard, there has been less movement within economics than in other social sciences, including political science, where a section of the American Political Science Association

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

[9] Equator: Enhancing the Quality and Transparency of Health Research; see www.equator-network.org/.

[10] See www.aeaweb.org/PDF_files/PR/AEA_Adopts_Extensions_to_Principles_for_Author_Disclosure_01-05-12.pdf.

[11] See www.aeaweb.org/aea_journals/AEA_Disclosure_Policy.pdf and www.nber.org/researchdisclosurepolicy.html.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

has developed guidelines for reporting of experimental research (Gerber et al. 2014). The American Political Science Association has formed committees that resulted in the Data Access and Research Transparency (DART) statement, which APSA adopted in both its Ethics Guide and Journal Editors' Transparency Statement, with 27 journals choosing to enact data sharing, data citation, and analytical methods sharing standards starting January 15, 2016.[12]

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

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.
2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.[13]
3. Authors must list all variables collected in a study.
4. Authors must report all experimental conditions, including failed manipulations.
5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.
6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

These disclosure rules are further simplified into a simple 21-word solution to be used by authors: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (Simmons, Nelson, and Simonsohn 2012). There is a corresponding statement to be used by reviewers: "I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure

[12] See www.dartstatement.org.

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

the statement is accurate. This is the standard reviewer disclosure request endorsed by the Center for Open Science (see http://osf.io/project/hadz3). I include it in every review."[14]

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

In the six months after the guidelines were published in *Science*, 538 journals and 57 organizations across a wide variety of scientific disciplines, including many in the social sciences, expressed their support for the standards and agreed to evaluate them for potential adoption. *Science* has now announced that it will be implementing the standards, effective January 1, 2017 (McNutt 2016). However, none of the leading economics journals have yet chosen to endorse or implement the guidelines; we encourage economics and other social science journal editors to review the guidelines and seriously consider adopting high transparency and reproducibility standards for their journals, keeping in mind that the TOP standards are meant to be modular rather than one-size-fits-all.

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

[14] See http://centerforopenscience.github.io/osc/2013/12/09/reviewer-statement-initiative/.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Fraud and Retractions

Building on the discussion from Chapter 6, it appears that the formulation of explicit social science journal standards for article retraction, and clearer communication on journal websites stating when an article is retracted, could also be beneficial. The RePEC tracking of offenses, mentioned in Chapter 6, is a helpful but only partial start. In political science, Laitin and Reich (2017) published a call to action, arguing for a more proactive approach of strong disciplinary norms and internal policing with improved graduate education, journal practices, and disciplinary practices, in the hopes that this could avoid future situations like the "Inside Job" documentary or the fraud uncovered in (Broockman, Kalla, and Aranow 2015).

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

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

Researchers have also developed novel statistical tools that one can use to detect fraud, using the fact that humans tend to drastically underestimate how noisy real data are when they are making up fraudulent data. Simonsohn (2013) used this forensic technique after observing summary statistics that were disturbingly similar across treatment arms to successfully combat fraud in psychology, resulting in the retraction of several papers by two prominent scholars.

Another potentially useful tool is post-publication peer review. Formalizing post-publication peer review puts us in relatively uncharted waters. Yet it is worth noting that all four of the AEA's *American Economic Journals* allow

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

for comments to appear on every article's official webpage post-publication (anonymous comments are not allowed). The feature does not appear to be widely used, but in one case, Lundqvist, Dahlberg, and Mörk (2014), comments placed on the website have actually resulted in changes to the article between its initial online pre-publication and the final published version, suggesting that this could be a useful tool for the research community to improve the quality of published work in the future.[15]

Open Data and Materials, and Their Use for Replication

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

This means that in practice, we economists as a discipline are still in a situation in which replication attempts for most empirical studies are still relatively costly in terms of time and effort. Despite improved (if still imperfect) data availability, we also know of no mainstream journal in economics that systematically tests that submitted data and code to actually produce the claimed results as a pre-condition of publication. An interesting new movement hoping to change this is the Peer Reviewer's Openness Initiative, whereby researchers can pledge that after a certain date (January 1, 2017) they

[15] www.aeaweb.org/articles.php?doi=10.1257/pol.6.1.167.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

will begin to require data sharing in the articles they referee (Morey et al. 2015).[16] If journal reviewers demand en masse to have access to the code and data that generated the results, and new norms develop around this expectation, this might lead to rapid changes in data sharing practices, given the central role that journal publication plays in scholars' individual professional success and standing.

As discussed above, the imprecise definition of the term "replication" itself often leads to confusion (Clemens 2017). Clarification of what authors mean when they say a replication "failed" (can the data not even produce the published results, are they not robust to additional specifications, or does a new sample or extended dataset produce different results?) may be an important first step to mainstreaming replication research within the social sciences.

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

In psychology, Nosek, Spies, and Motyl (2012) are also skeptical of creating new journals devoted to replications or null results, and instead suggest crowdsourcing replication efforts. This seems have to been extremely successful, with two large-scale replication efforts in which many researchers worked together to repeat classic experiments in psychology with new samples, the Many Labs project (Klein et al. 2014) and the Replication Project: Psychology (Open Science Collaboration 2012, 2015). Both were published in prominent journals and widely covered in the popular media. A similar project in cancer biology is ongoing.[17]

The Many Labs project sought to reproduce 13 effects found in the literature, testing them in 36 samples with a total sample size of 6344, and determining whether online samples produced different effects than lab samples, and also comparing international to US samples. They find that two types of

[16] For more information, see http://opennessinitiative.org.

[17] http://elifesciences.org/collections/reproducibility-project-cancer-biology.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

interventions failed to replicate entirely, while results for other replications relative to the original studies were more nuanced.

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

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

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

Economists may be interested to know that the researchers behind the RPP also included a prediction market in their project, and the market did a fairly good job of predicting which of the effects studies would ultimately be reproduced (Dreber et al. 2015). Unlike the prediction market in Camerer et al. (2016), the RPP prediction market outperformed a survey of researcher beliefs.[18]

[18] For related research on expert predictions, see DellaVigna and Pope (2016). Other psychology researchers have tried another way to crowdsource replication: instead of bringing different research

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

Despite the inability to replicate so many prominent empirical papers in economics (discussed above), there have been few systematic effort to replicate findings, with one exception in addition to Camerer et al. (2016) being the new 3ie replication program for development economics studies, which has replicated a handful papers to date, including one by an author of this article.[19] Few economics journal editors specifically seek to publish replications, and even fewer are willing to publish "successful" replications, i.e., papers that demonstrate that earlier findings are indeed robust, with the *Journal of Applied Econometrics* being a notable exception (Pesaran 2003). Despite the value to the research enterprise of more systematic evidence on which empirical results are actually reliable, and the fact that many scholars have advocated for changes in this practice over the years with a near-constant stream of editorials (see among others Kane 1984; Mittelstaedt and Zorn 1984; Fuess 1996; Hunter 2001; Camfield and Palmer-Jones 2013; Duvendack and Palmer-Jones 2013; Duvendack, Palmer-Jones, and Reed 2015; Zimmermann 2015), as yet there has been little progress within the economics profession toward actually publishing replication studies on a more general basis (Andreoli-Versbach and Mueller-Langer 2014). In many ways, the patterns in economics are similar to those in the other social sciences, particularly in political science, where prominent voices have long spoken out in favor of replication, but their publication remains rare (King 1995; Gherghina and Katsanidou 2013).

Computational Issues

Scholars' ability to carry out replications and share data has been facilitated by new software and computational improvements. Some of these advances are described in Koenker and Zeileis (2009). They discuss what has come to be called Claerbout's principle: "An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures." Koenker and Zeileis recommend version control, using open-source

groups together to all independently run the same classic experiment, other researchers have independently analyzed the same observational dataset and attempted to answer the same question, in this case, the question of whether darker skin-toned soccer players receive more red cards as a result of their race, conditional on other factors (Silberzahn and Uhlmann 2015).

[19] www.3ieimpact.org/evaluation/impact-evaluation-replication-programme/.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

programming environments when possible (including for document preparation), and literate programming, which is defined below.

Version control software makes it easier to maintain detailed record-keeping of changes to statistical code even among multiple collaborators. Koenker and Zeileis discuss one such centralized system, Subversion (SVN, http://subversion.apache.org), but in recent years distributed forms of version control such as Git have become more widely used, and are well supported by a user community.[20]

For document preparation, Koenker and Zeleis discuss LaTeX, which has a steep learning curve but has the advantage of being open-source, and has the ability to intermix, or "weave" text, code, and output. Even more recently, dynamic documents (which Koenker and Zeileis refer to as literate programming; see also Knuth 1992) can be used to write statistical analysis code and the final paper all in a single master document, making it less likely that copying and pasting between programs will lead to errors, and making it possible in some cases to reproduce an entire project with a single mouse click. The knitr package for R, incorporated into R Studio, makes this relatively easy to implement (Xie 2013, 2014). Jupyter notebooks (http://jupyter.org) also simplifies interactive sharing of computational code with over 40 popular open-source programming languages (Shen 2014). Many programs that accommodate these approaches, including R, Python, and Julia, are open-source, making it easier for members of the research community to look under the hood and possibly reduce the risk of the software computational errors documented in McCullough and Vinod (2003).[21] Computational aspects of reproducibility are discussed at length in Stodden, Leisch, and Peng (2014).

The Limits of Open Data

While we believe that the social sciences as a whole would benefit from stronger data sharing requirements and more widespread publication of

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

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

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

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

Future Directions and Conclusion

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

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

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

plans, and more broadly in issues of research transparency and openness, appears to be particularly high among PhD students and younger faculty (at least anecdotally), suggesting that there may be a generational shift at work.

The Berkeley Initiative for Transparency in the Social Sciences (BITSS) is another institution that has emerged in recent years to promote dialogue and build consensus around transparency practices. BITSS has established an active training program for the next generation of economists and other social scientists, as well as an award to recognize emerging leaders in this area, the Leamer-Rosenthal Prize for Open Social Science.[22] Other specialized organizations have also emerged in economics: the Replication Network aims promote the publication of replication studies, Project TIER has developed a curriculum to teach computational reproducibility to economics undergraduates, and MAER-NET has developed guidelines for meta-analysis (Stanley et al. 2013). Similar organizations play analogous roles in other disciplines, including the Center for Open Science (COS), which is most active within psychology (although it spans other fields), and the Evidence in Governance and Politics (EGAP) group.[23]

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

There are many potential avenues for promoting the adoption of new and arguably preferable practices, such as the data sharing, disclosure, and

[22] www.bitss.org. In the interest of full disclosure, Miguel is one of the founders of BITSS and currently its faculty director, and Christensen worked as a post-doctoral research fellow at BITSS. BITSS is an initiative of the Center for Effective Global Action at UC Berkeley.

[23] http://cos.io; www.egap.org.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press

pre-registration approaches described at length in this chapter. One issue that this chapter does not directly address is how to most effectively – and rapidly – shift professional norms and practices within the empirical social science research community. Shifts in graduate training curricula,[24] journal standards (such as the Transparency and Openness Promotion Guidelines), and research funder policies might also contribute to the faster adoption of new practices, but their relative importance remains an open question. The study of how social norms among economists have shifted, and continue to evolve, in this area is an exciting social science research topic in its own right, and one that we hope is also the object of greater scholarly inquiry in the coming years.

[24] See http://emiguel.econ.berkeley.edu/teaching/12 for an example of a recent PhD-level course on research transparency methods at UC Berkeley taught by the authors.

https://doi.org/10.1017/9781108762519.007 Published online by Cambridge University Press