C2

2

Improving Research Transparency in the Social Sciences

Registration, Preregistration, and Multiple Testing Adjustments

Garret Christensen and Edward Miguel[*]

C2.S1

Introduction

C2.P1

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

C2.P2

The goal of this chapter is to survey this emerging literature on research transparency and reproducibility, and synthesize the insights emerging in economics as well as from other fields—awareness of these issues has also recently come to the fore in political science (Gerber et al., 2001), psychology (Franco et al., 2016; Open Science Collaboration, 2015; Simmons et al., 2011), sociology (Gerber & Malhotra, 2008b), across the social sciences

[*] A similar paper with some related material was published as Christensen and Miguel (2018).

Garret Christensen and Edward Miguel, *Improving Research Transparency in the Social Sciences* In: *Research Integrity*. Edited by: Lee J. Jussim, Jon A. Krosnick, and Sean T. Stevens, Oxford University Press. © Oxford University Press 2022. DOI: 10.1093/oso/9780190938550.003.0002

(Franco et al., 2014), finance (Harvey et al., 2016), and other research disciplines as well, including medicine (Ioannidis, 2005). We also discuss productive avenues for future work.

C2.P3 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 progress has also facilitated various new statistical tools and potential solutions, including improved tests for publication bias, new ways to test the robustness of multiple estimates, and registration and preregistration of studies. Yet, as we discuss here, 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.[1]

C2.P4 The rest of the paper is organized as follows. The first section focuses on documenting the problems, focusing on publication bias specification searching. The second section focuses on possible solutions to these issues: improved analytical methods, study registration, and pre-analysis plans. The final section discusses future directions for research as well as possible approaches to change norms and practices.

C2.S2 Evidence on Problems with the Current Body of Research

C2.P5 Multiple problems have been identified within the body of published research results in the social sciences. We focus on two that have come under greater focus in the recent push for transparency: publication bias and specification searching. Before describing them, it is useful to frame some key issues with a simple model.

[1] In addition to methodological concepts we discuss here, journals have been improving, requiring data (Bernanke, 2004; Wilson, 2012; Wilson, 2010), adopting guidelines such as the Transparency and Openness Promotion (TOP) Guidelines (McNutt, 2016; Nosek et al., 2015), and reviewing and publishing papers based on design rather than results (Chambers, 2013; Findley et al., 2016; Foster et al., 2018; Foster et al., 2019). Organizations such as the Center for Open Science (http://cos.io) and the Berkeley Initiative for Transparency in the Social Sciences (http://bitss.org) have also formed to educate and facilitate adoption. We discuss these in more detail in Christensen et al. (2019).

C2.S3 ## Publication Bias

C2.P6 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.[2] If the research community is unable to track the complete body of statistical tests that have been run, including those that fail to reject the null (and thus are less likely to be published), then we cannot determine the true proportion of tests in a literature that reject the null. Thus it is critically important to understand how many tests have been run. The term "file drawer problem" was coined decades ago (Rosenthal, 1979) to describe this problem of results that are missing from a body of research evidence. The issue was a concern even earlier; see, for example, Sterling (1959), which warned of "embarrassing and unanticipated results" from type I errors if not significant results went unpublished.

C2.P7 Important recent research by Franco et al. (2014, 2016) 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—that is, those that are not statistically significant (Franco et al., 2014).

C2.P8 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.[3] TESS funded studies across research fields, including in economics (e.g., Allcott & Taubinsky, 2015; Walsh et al., 2009) as well as political science, sociology, and other fields. Franco and colleagues 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.

C2.P9 They find a striking empirical pattern: Studies where the main hypothesis tested yielded null results are 40 percentage points less likely to be published

[2] Note that in a general sense "publication bias" could refer to the bias inherent in research publications from fads, topic timeliness, author status, political activism, or numerous other sources, but we mostly refer to the nonpublication of statistical null findings.

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

in a journal than those with 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 positive predictive value (PPV) of research is likely to be lower than it would be otherwise, and also has negative implications for the validity of meta-analyses, if null results are not known to the scholars attempting to draw broader conclusions about a body of evidence. The same TESS database yielded 32 psychology studies that the authors further analyzed, concluding that 40% of studies did not fully report all experimental conditions, and reported effects were twice as large as those unreported (Franco et al., 2016).

C2.P10 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 (*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.

C2.P11 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% 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.

C2.P12 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 (2008b) have used the caliper test, which compares the frequency of test statistics just above and below the key statistical significance cutoff, which is similar in spirit to a regression discontinuity design. Specifically, they compare the number of z-scores lying in the interval [1.96 – *X*%, 1.96] to the number in [1.96, 1.96 + *X*%], where X is the size of the caliper, and they examine these differences at 5%, 10%, 15%, and 20% critical values.[4]

C2.P13 These caliper tests are used to examine reported empirical results in leading sociology journals (*American Sociological Review*, *American Journal of Sociology*, and *Sociological Quarterly*) and reject the hypothesis of no publication bias at the 1-in-10-million level (Gerber & Malhotra, 2008b). Data from two leading political science journals (*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 & Malhotra, 2008a).

C2.P14 Psychologists have recently developed a related tool called the "p-curve," describing the density of reported p-values in a literature, that again takes advantage of the fact that if the null hypothesis were true (i.e., no effect), p-values should be uniformly distributed between 0 and 1 (Simonsohn et al., 2014a). Intuitively, under the null of no effect, a p-value less than 0.08 should occur 8% of the time, a p-value less than 0.07 occurs 7% of the time, etc., meaning a p-value between 0.07 and 0.08, or between any other 0.01-wide interval, should occur 1% of the time. In the case of true non-zero effects, the distribution of p-values should be right-skewed (with a decreasing density), with more low values (0.01) than higher values (0.04) (Hung et al., 1997).[5] In contrast, in bodies of empirical literature suffering from publication bias, or

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

[5] Unlike economics journals, which often use asterisks or other notation to separately indicate p-values (0,.01),[0.01 <.05), and [.05,.1), psychology journals often indicate only whether a p-value is less than 0.05, and this is the standard used throughout (Simonsohn et al., 2014a).

"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).

C2.P15 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%, 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 et al., 2014b).[6]

C2.P16 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 et al. (2014). Easterbrook et al. (1991) reviewed the universe of protocols submitted to the Central Oxford Research Ethics Committee, and both Kirsch et al. (2008) and Turner et al. (2008) employ the universe of tests of certain antidepressant drugs submitted to the U.S. Food and Drug Administration (FDA), and all found significantly higher publication rates when tests yield statistically significant results. Turner et al. found that 37 of 38 (97%) of trials with positive (i.e., statistically significant) results were published, while only 8 of 24 (33%) with null (or negative) results were published; for a meta-meta-analysis of the latter two studies, see Ioannidis (2008).

C2.P17 A simple model of publication bias described in McCrary et al. (2015) suggests that, under some relatively strong assumptions regarding the rate of nonpublication of statistically nonsignificant results, readers of research studies could potentially adjust their significance threshold to "undo" the distortion by using a more stringent t-test statistic of 3.02 (rather than 1.96) to infer statistical significance at 95% confidence. They note that approximately

[6] For an online implementation of the p-curve, see http://p-curve.com. Also see a discussion of the robustness of the test in Simonsohn et al. (2015a) and Ulrich & Miller (2015).

30% of published test statistics in the social sciences fall between these two cutoffs. It is also possible that this method would break down and result in a "t-ratio arms race" if all researchers were to use it, so it is mostly intended for illustrative purposes.

C2.P18 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 nonrejection 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 et al. (2015).

C2.P19 Taken together, a growing body of evidence indicates that publication bias is widespread in economics and many other scientific fields. Stepping back, these patterns do not appear to occur by chance, but are likely to indicate some combination of selective editor (and referee) decision-making, the file drawer problem alluded to above, and/or widespread specification searching (the focus of the next subsection), which is closely related to what the Ioannidis (2005) model calls author bias.

C2.S4

Specification Searching

C2.P20 While publication bias implies a distortion of a body of multiple research studies, bias is also possible within any given study. 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 preconceived notions—or were seen as particularly interesting within the research community—and ignoring, whether consciously or not, other specifications that did not.

C2.P21 One the most widely cited articles from this period is Leamer's (1983) "Let's Take the Con Out of Econometrics," which discusses the promise of improved research design (namely, randomized trials) and argues that in observational research, researchers ought to transparently report the entire range of estimates that result from alternative analytical decisions. Leamer's illustrative application employs data from a student's research project, namely U.S. data from 44 states, to test for the existence of a deterrent effect of the death penalty on the murder rate. (These data are also used in McManus

[1985].) Leamer classifies variables in the data as either "important" or "doubtful" determinants of the murder rate, and then runs regressions with all possible combinations of the doubtful variables, producing a range of different estimates. Depending on which set of control variables, or covariates, were included (among state median income, unemployment, percent population nonwhite, percent population 15 to 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.

C2.P22 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]

C2.P23 Since the 1980s, empirical research practices in economics have changed significantly, especially with regards to improvements in research design. Angrist and Pischke (2010) make the point that improved experimental and quasi-experimental research designs have made much econometric inference more credible. Leamer (2010), however, 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.

C2.P24 Related points have been made in other social science fields in recent years. In psychology, Simmons et al. (2011) "prove" that listening to the Beatles' song "When I'm Sixty-Four" made listeners a year and a half younger. The extent and ease of this "fishing" in analysis is also described in political science by Humphreys et al. (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

[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 & Metaxoglou, 2011; Knittel & Metaxoglou, 2013). McCullough and Vinod (2003) contains a more general discussion of robustness and replication failures in nonlinear maximization methods.

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."

C2.P25 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 earlier, the analysis of 641 articles from three top economics journals in recent years presented in Brodeur et al. (2016) still shows a disturbing two-humped distribution of p-values, with relatively few p-values between 0.10 and 0.25 and far more just below 0.05. Their analysis also explores the correlates behind this pattern, and finds that this apparent misallocation of p-values just below the accepted statistical significant level was less pronounced for articles written by tenured authors, and tentatively find it less pronounced among studies based on randomized controlled trials (suggesting that improved research design itself may partially constrain data mining), but they did not detect any differences in the pattern based on whether the authors had publicly posted the study's replication data in the journal's public archive.

C2.S5 ### Subgroup Analysis

C2.P26 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.

C2.P27 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 & Grimes, 2005).

C2.S6

New Research Methods and Tools

C2.P28 This section discusses several new methods and tools that have emerged in social science research over the past two decades—and more forcefully over the past 10 years—to address the concerns we have just discussed. These approaches have in common a focus on greater transparency and openness in the research process. They include improved analytical methods regarding model uncertainty and multiple testing adjustments, and study registration and pre-analysis plans; we discuss each in turn.

C2.S7

Improved Analytical Methods: Model Uncertainty and Multiple Testing Adjustments

C2.P29 There have been a number of different responses within economics to the view that pervasive specification searching and publication bias was affecting the credibility of empirical literatures. As mentioned earlier, 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 et al., 2007) and quasi-experimental (Angrist & Pischke, 2010) research designs arguably place more constraints on researchers relative to earlier empirical approaches, since there are natural ways to present data using these designs that researchers are typically compelled to present by colleagues in seminars and by journal referees and editors. Prospective experimental studies also tend to place greater emphasis on adequately powering an analysis statistically, which may help to reduce the likelihood of publishing only false positives (Duflo et al., 2007).

C2.P30 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 controlled trial studies than in studies utilizing nonexperimental 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.

C2.S8 Understanding Statistical Model Uncertainty

C2.P31 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." 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.

C2.S9 *Specification Curve*

C2.P32 Simonsohn et al. (2015b) propose a method, which they call the "specification curve," that is similar in spirit to Leamer's extreme bounds analysis, but they recommend researchers test the exhaustive combination of analytical decisions, not just decisions about which covariates to include in the model. If the full exhaustive set is too large to be practical, a random subset can be used. After plotting the effect size from each of the specifications, researchers can assess how much the estimated effect size varies, and which combinations of decisions lead to which outcomes. Using permutation tests (for treatment with random assignment) or bootstrapping (for treatment without random assignment), researchers can generate shuffled samples with no true effect by construction, and compare the specification curves from these placebo samples to the specification curve from the actual data. Many comparisons are possible, but the authors suggest comparing the median effect size, the share of results with the predicted sign, and the share of statistically significant results with the 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.

C2.P33 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 résumés. Jung et al. (2014) elicited four critical responses taking issue

with the analytical decisions (Bakkensen & Larson, 2014; Christensen & Christensen, 2014; Maley, 2014; Malter, 2014). The specification curve shows that 46% of curves from permuted data show at least as large a median effect size as the original, 16% show at least as many results with the predicted sign, and 85% show at least as many significant results with the predicted sign. This indicates that the results are likely to have been generated by chance. The Bertrand and Mullainathan (2004) specification curve, on the other hand, shows that fewer than 0.2% of the permuted curves generate as large a median effect, 12.5% of permuted curves show at least as many results with the predicted sign, and less than 0.2% of permuted curves show at least as many significant results with the predicted sign, providing evidence that the results are very unlikely to have been generated by chance.

C2.S10 Improved Publication Bias Tests

C2.P34 There have been significant advances in the methodological literature on quantifying the extent of publication bias in a given body of literature. Early methods include the Rosenthal (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.

C2.P35 Statisticians have developed methods to estimate effect sizes in meta-analyses that control for publication bias (Hedges, 1992; Hedges & Vevea, 1996). The tools most widely used by economists tend to be simpler, including the widely used funnel plot, which is a scatterplot of some measure of statistical precision (typically the inverse of the standard error), versus the estimated effect size. Estimates generated from smaller samples should usually form the wider base of an inverted funnel, which should be symmetric around more precise estimates in the absence of publication bias. The method is illustrated with several economics examples in Stanley and Doucouliagos (2010). 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:

C2.P36 (eqn. 4)
$$t_i = \frac{Estimated\ effect_i}{SE_i} = \beta_0 + \beta_1 \left(\frac{1}{SE_i} \right) + v_i .$$

C2.P37 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.

C2.P38 Using the FAT, Doucouliagos and Stanley (2009) find evidence of publication bias in the Card and Krueger (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 & Doucouliagos (2012), while significant debate surrounding the validity, or falsifiability, of this and other meta-analysis techniques can be found in Vosgerau et al. (2019).

C2.S11 ## Multiple Testing Corrections

C2.P39 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 et al., 2014). The practice of correcting for multiple tests is already widespread in certain scientific fields (e.g., genetics) but has yet to become the norm in economics and other social sciences. Simply put, since we know that p-values fall below traditional significance thresholds (e.g., 0.05) purely by chance a certain proportion of the time, it makes sense to report adjusted p-values that account for the fact that we are running multiple tests, since this makes it more likely that at least one of our test statistics has a significant p-value simply by chance.

C2.P40 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.

C2.S12 ### *Reporting Index Tests*

C2.P41 One option for scholars in cases where there are multiple related outcome measures is to forgo 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 et al. (2007) implement an early index test in the "Moving to Opportunity" field experiment using methods developed in biomedicine by O'Brien (1984).

C2.P42 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.

C2.S13 *Controlling the FWER*

C2.P43 The FWER is the probability that at least one true null hypothesis in a group is rejected (a type I error, or false positive). This approach is considered most useful when the "damage" from incorrectly claiming that *any* null 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 & Altman, 1995), although this is extremely conservative, and improved methods have been developed.

C2.P44 Holm's sequential method involves ordering p-values by class and multiplying the lower p-values by higher discount factors (Holm, 1979, p. 1). 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 preschool interventions (namely, the Abecedarian, Perry, and Early Training Project studies) exhibit few positive long-run impacts for males.

C2.P45 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 et al. (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.[8]

C2.S14 *Controlling the FDR*

C2.P46 In situations where a single type I error is not considered very costly, researchers may be willing to use a somewhat less conservative method than the FWER approach, and trade off some incorrect hypothesis rejections in exchange for greater statistical power. This is made possible by controlling the FDR, or the percentage of rejections that are type I 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 & Yekutieli, 2001). Benjamini et al. (2006) describe a two-step procedure with greater statistical power, while Romano et al. (2008) propose the first methods to incorporate information about the dependence structure of the test statistics.

C2.P47 Multiple hypothesis testing adjustments have recently been used in finance (Harvey et al., 2016) to reevaluate 316 factors from 313 different papers that explain the cross-section of expected stock returns. The authors employ the Bonferroni, Holm (1979), and Benjamini et al. (2006) methods to account for multiple testing, and conclude that t-statistics greater than 3.0, and possibly as high as 3.9, should be used instead of the standard 1.96, to actually conclude that a factor explains stock returns with 95% confidence. Index tests and both the FWER and FDR multiple testing corrections are also employed in Casey et al. (2012) to estimate the impacts of a community-driven development program in Sierra Leone using a dataset with hundreds of potentially relevant outcome variables.

C2.S15 Study Registration

C2.P48 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

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

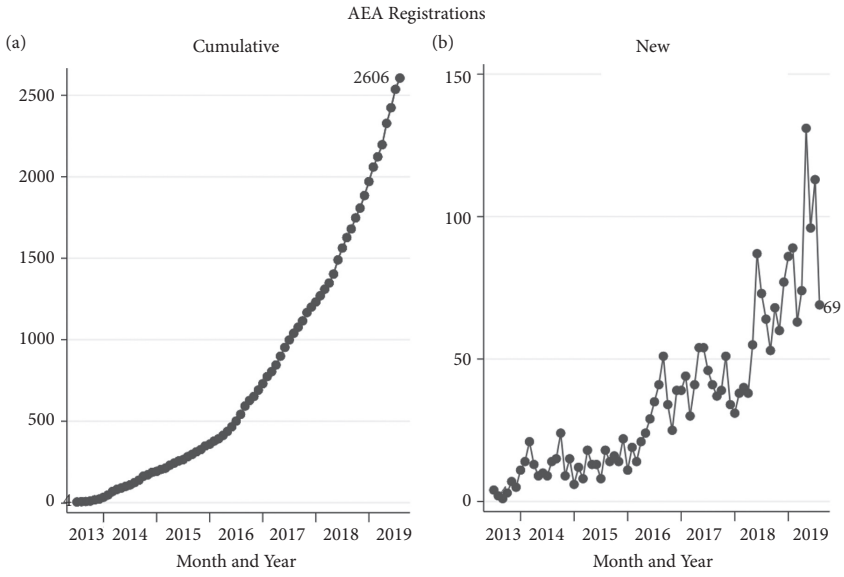
centralized database of all attempts to conduct research on a certain question, irrespective of the nature of the results, and such that even null (not statistically significant) findings are not lost to the research community. Top medical journals have adopted a clear standard of publishing only medical 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.[9]

C2.P49 While recent research in medicine finds that the clinical trial registry has not eliminated all underreporting 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 et al. (2014).

C2.P50 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 its inception in 2013, over 2,600 studies conducted in over 100 countries have been registered, and the pace of registrations continues to rise rapidly. Figure 2.1, Panel A, presents the total number of registrations over time in the AEA registry (through May 2019), 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.

C2.P51 In addition to the AEA registry, several other social science registries have recently been created, including 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 et al., 2011), and the Experiments in Governance and Politics (EGAP) registry (http://egap.org/content/registration), also created in 2013. The Center for Open Science's Open Science Framework (OSF, http://osf.io) accommodates the registration of essentially any study

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



C2.F1 **Figure 2.1.** Studies in the AEA trial registry, May 2013 to May 2019. Figure shows the cumulative (Panel A) and new (Panel B) trial registrations in the American Economic Association Trial Registry (http://socialscienceregistry.org).
Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

or research document by allowing users to create a frozen time-stamped web URL with associated digital object identifier (DOI) for any materials uploaded to OSF. Several popular data storage options (including Dropbox, Dataverse, and GitHub) can also be synced with the OSF and its storage, creating a flexible way for researchers to register their research and materials. As of October 2016, over 7,300 public registrations have been created on OSF since the service launched in 2013.

C2.S16 ## Pre-Analysis Plans

C2.P52 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 preregistration 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 FDA'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: prespecification 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 (U.S. Food and Drug Administration, n.d.).

C2.P53 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 prespecified analysis has grown rapidly in the past few years, mirroring the rise of studies posted on the AEA registry.

C2.P54 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,[10] and a template for PAPs by Ganimian (2014). Ganimian's template may be particularly useful to researchers themselves when developing their own PAPs, and instructors may find it useful in their courses, and additional templates can be found on the OSF.

C2.P55 Building on, and modifying, the FDA's 1998 checklist with insights from these other recent treatments of PAPs, there appears to be a growing consensus that PAPs in the social sciences should consider discussing at least the following list of 10 issues:

C2.P56 1. Study design
C2.P57 2. Study sample
C2.P58 3. Outcome measures
C2.P59 4. Mean effects family groupings
C2.P60 5. Multiple hypothesis testing adjustments
C2.P61 6. Subgroup analyses
C2.P62 7. Direction of effect for one-tailed tests
C2.P63 8. Statistical specification and method
C2.P64 9. Structural model
C2.P65 10. Time stamp for verification

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

C2.P66 PAPs 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.

C2.P67 For those concerned about the possibility of "scooping" of new research designs and questions based upon a publicly posted PAP 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 4-year embargo until the information is made public.[11]

C2.S17 Examples of PAPs

C2.P68 Recent examples of economics papers based on experiments with PAPs include Casey et al. (2012) and Finkelstein et al. (2012), among others. Casey et al. (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 postwar 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% confidence for any of the nine families of outcomes.

C2.P69 Yet Casey et al. (2012) go on to show that, given the large numbers of outcomes in their dataset, and the multiplicity of ways to define outcome measures, finding some statistically significant results would have been relatively easy. In fact, the paper includes an example of how, if they had had the latitude to define outcomes without a PAP, as has been standard practice

[11] See https://help.osf.io/hc/en-us/articles/360019930893-Register-Your-Project. Accessed August 2, 2019.

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 2.1, where Panel A presents the statistically significant positive impacts identified in the GoBifo data and Panel B highlights negative effects. This finding

C2.T1 **Table 2.1** 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 tarpaulin use	0.54	0.06*	(0.03)
Respondent agrees with "Responsible young people can be good leaders"	0.76	0.04*	(0.02)
Correctly able to name the year of the next general elections	0.19	0.04*	(0.02)

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.

prompts us to ask 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?

C2.P70 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 PAP, 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 combination of the PAP and multiple hypothesis testing adjustments gives 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 ERs more, and had better detection and management of diabetes.

C2.P71 Additional studies that have resulted from the experiment have also employed PAPs, and they show that health insurance increased ER use (Taubman et al., 2014), had no effect on measured physical health outcomes after 2 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).

C2.P72 Other prominent early examples of economics studies that have employed PAPs 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 (Alatas et al., 2012; Hirshleifer et al., 2015; Olken et al., 2012; Wydick et al., 2014). The PAP tool is also spreading to other social sciences beyond economics. For instance, in psychology, a prespecified replication of an earlier paper that had found a link between female conception risk and racial prejudice failed to find a similar effect (Hawkins et al.,

2015). In political science, the Election Research Preacceptance Competition ran a competition for work with PAPs based on the 2016 American National Election Studies (ANES) data; eligible papers were required to register their analysis plan prior to the public release of the data.[12]

C2.P73 One issue that arises for studies that did register a PAP 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 prespecified 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.

C2.P74 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 prespecified 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 prespecified are labeled with a "^" character to set them apart and are clearly labeled as such.

C2.S18 ### Strengths, Limitations, and Other Issues Regarding PAPs

C2.P75 There remain many questions about whether, when, and how PAPs 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, PAPs 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).

C2.P76 A further advantage of the use of PAPs 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

[12] See https://www.erpc2016.com/.

reform). This is especially the case if researchers and their institutional partners can agree on the PAP, as a sort of evaluation contract.

C2.P77 On the negative side, PAPs are often complex and take valuable time to write. Scientific breakthroughs often come at unexpected times and places, often as a result of exploratory analysis, and the time spent writing PAPs may thus mean less time to spend on less structured data exploration.

C2.P78 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 PAPs. Yet there does appear to be a growing consensus, endorsed by Coffman and Niederle, that PAPs 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. Berge et al. (2015), for instance, 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 PAP is valuable.

C2.P79 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 PAP 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 (with sufficiently low sample attrition rates in a longitudinal study, etc.), the results are to be published regardless of whether they are statistically significant, and whether they conform to the expectations of the editor or referees, or to the conventional wisdom in the discipline.

C2.P80 Several psychology journals currently have begun using results-blind review, either regularly or in special issues (Chambers, 2013; Chambers et al.,

2014; Nosek & Lakens, 2014).[13] An issue of *Comparative Political Studies* was the first to feature results-blind review in political science (Ansell & Samuels, 2016; Findley et al., 2016), and it included both experimental and observational research studies. The *Journal of Development Economics* announced that it would pilot acceptance of these articles (Foster et al., 2018) and later fully adopted the practice (Foster et al., 2019). The rise in experimental studies and PAPs in economics, as evidenced by the rapid growth of the AEA registry, is likely to facilitate the wider acceptance of this approach.

C2.S19 ### Observational Studies

C2.P81 An important open question is how widely the approach of study registration and hypothesis prespecification could be usefully applied in nonprospective and nonexperimental studies. This issue has been extensively discussed in recent years within medical research but consensus has not yet been reached in that community. It actually appears that some of the most prestigious medical research journals, which typically publish randomized trials, are even more in favor of the registration of observational studies than the editors of journals that publish primarily in nonexperimental research (Dal-Ré et al., 2014; Epidemiology Editors, 2010; Lancet, 2010; Loder et al., 2010).

C2.P82 A major logical concern with the preregistration of nonprospective observational studies using preexisting data is that there is often no credible way to verify that preregistration 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 preregistration of observational work have not formulated a convincing response to this obvious concern.

C2.P83 The only economics study of which we are aware that has used a PAP on nonexperimental data was 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 prespecified research designs, in a bid to eliminate "author effect," and that this could create a productive "adversarial collaboration" between authors with starkly different prior views on the likely impacts of the policy change (Levine, 2001). (The concept of adversarial collaboration—two

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

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).

C2.P84 The U.S. federal minimum wage increased in October 1996 and September 1997. Although Krueger ultimately decided not to participate, Neumark submitted a prespecified 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.

C2.P85 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; Card, 1992b), and tests both models with and without controls for the employment rate of higher-skilled prime-age adults (as recommended by Deere et al., 1995). The results mostly fail to reject the null hypothesis of no effect of the minimum wage increase: Only 18 of the 80 specifications result in statistically significant decreases in employment (at the 90% confidence level), with estimated elasticities ranging from −0.14 to −0.3 for the significant estimates and others closer to zero.

C2.P86 A more recent study bases its analysis on Neumark's exact prespecified tests estimate the effect of minimum wages in Canada and found larger unemployment effects, but the authors had access to the data before estimating their models and did not have an agreement with the journal, so the value of this "prespecification" is perhaps less clear (Campolieti et al., 2006). In political science, a prespecified observational analysis measured the effect of the immigration stances of Republican representatives on their 2010 election outcomes (Monogan, 2013).

C2.P87 It is difficult to see how researchers could reach Neumark's level of prespecified detail with a research question with which they were not already intimately familiar. It seems more likely that in a case where the researchers were less knowledgeable they might either prespecify 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 PAP in cases where the plan omits important details or contains errors, as argued in Casey et al (2012).

C2.P88 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 prespecified fashion, following the example in Neumark (2001). Although prespecification 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 earlier, with similar arguments for future elections), will make it possible for the verifiable prespecification of research analysis to be carried out in many settings.

C2.S20 ## Future Directions and Conclusion

C2.P89 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.

C2.P90 The increased institutionalization of new practices—including through the new AEA randomized controlled trial registry, which has rapidly attracted hundreds of studies, many employing PAPs, something unheard of in economics until a few years ago—is evidence that new norms are emerging. The rise in the use of PAPs has been particularly rapid in certain subfields, especially development economics, pushed forward by policy changes promoting PAPs in the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Center for Effective Global Action. Interest in PAPs, and more broadly in issues of research transparency and openness, appears to be particularly high

among Ph.D. students and younger faculty (at least anecdotally), suggesting that there may be a generational shift at work.

C2.P91 At the same time, we have highlighted many open questions. The role that PAPs and study registration could or should play in observational empirical research—which represents the vast majority of empirical social science work, even a couple of decades into the well-known shift toward experimental designs—as well as in structural econometric work, macroeconomics, economic theory, and other subfields in economics and other social sciences 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 social science research. Will the use of study registries and PAPs 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.

C2.S21 ## References

C2.P92 Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., & Tobias, J. (2012). Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, *102*(4), 1206–1240. https://doi.org/10.1257/aer.102.4.1206

C2.P93 Allcott, H., & Taubinsky, D. (2015). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, *105*(8), 2501–2538. https://doi.org/10.1257/aer.20131564

C2.P94 Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. https://doi.org/10.1198/016214508000000841

C2.P95 Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

C2.P96 Ansell, B., & Samuels, D. (2016). Journal editors and "results-free" research: A cautionary note. *Comparative Political Studies*, *49*(13), 1809–1815. https://doi.org/10.1177/0010414016669369

C2.P97 Baicker, K., Finkelstein, A., Song, J., & Taubman, S. (2014). The impact of Medicaid on labor market activity and program participation: Evidence from the Oregon health insurance experiment. *American Economic Review*, *104*(5), 322–328. https://doi.org/10.1257/aer.104.5.322

C2.P98 Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M., & Finkelstein, A. N. (2013). The Oregon experiment—Effects of Medicaid on clinical outcomes. *New England Journal of Medicine*, *368*(18), 1713–1722. https://doi.org/10.1056/NEJMsa1212321

C2.P99 Bakkensen, L. A., & Larson, W. (2014). Population matters when modeling hurricane fatalities. *Proceedings of the National Academy of Sciences USA*, *111*(50), E5331–E5332. https://doi.org/10.1073/pnas.1417030111

C2.P100 Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*(8), 1561–1580. https://doi.org/10.1016/j.jpubeco.2004.06.013

C2.P101 Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

C2.P102 Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

C2.P103 Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*(4), 1165–1188.

C2.P104 Berge, L. I. O., Bjorvatn, K., Galle, S., Miguel, E., Posner, D. N., Tungodden, B., & Zhang, K. (2015). *How Strong Are Ethnic Preferences?* (Working Paper No. 21715). National Bureau of Economic Research. http://www.nber.org/papers/w21715

C2.P105 Bernanke, B. S. (2004). Editorial statement. *American Economic Review*, *94*(1), 404–404.

C2.P106 Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

C2.P107 Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ: British Medical Journal*, *310*(6973), 170.

C2.P108 Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2016). Star Wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32.

C2.P109 Campolieti, M., Gunderson, M., & Riddell, C. (2006). Minimum wage impacts from a prespecified research design: Canada 1981–1997. *Industrial Relations: A Journal of Economy and Society*, *45*(2), 195–216. https://doi.org/10.1111/j.1468-232X.2006.00424.x

C2.P110 Card, D. (1992a). Do minimum wages reduce employment? A case study of California, 1987–89. *Industrial & Labor Relations Review*, *46*(1), 38–54. https://doi.org/10.1177/001979399204600104

C2.P111 Card, D. (1992b). Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial & Labor Relations Review*, *46*(1), 22–37. https://doi.org/10.1177/001979399204600103

C2.P112 Card, D., & Krueger, A. B. (1995). Time-series minimum-wage studies: A meta-analysis. *American Economic Review*, *85*(2), 238–243.

C2.P113 Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping institutions: Evidence on aid impacts using a preanalysis plan. *Quarterly Journal of Economics*, *127*(4), 1755–1812. https://doi.org/10.1093/qje/qje027

C2.P114 Chambers, C. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610. http://dx.doi.org/10.1016/j.cortex.2012.12.016

C2.P115 Chambers, C. D., Feredoes, E., D. Muthukumaraswamy, S., & J. Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at AIMS

Neuroscience and beyond. *AIMS Environmental Science*, *1*(1), 4–17. https://doi.org/10.3934/Neuroscience.2014.1.4

C2.P116 Christensen, B., & Christensen, S. (2014). Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences USA*, *111*(34), E3497–E3498. https://doi.org/10.1073/pnas.1410910111

C2.P117 Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and Reproducible Social Science Research: How to Do Open Science*. https://www.ucpress.edu/book/9780520296954/transparent-and-reproducible-social-science-research

C2.P118 Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980. https://doi.org/10.1257/jel.20171350

C2.P119 Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, *29*(3), 81–98.

C2.P120 Dahl Rasmussen, O., Malchow-Møller, N., & Barnebeck Andersen, T. (2011). Walking the talk: The need for a trial registry for development interventions. *Journal of Development Effectiveness*, *3*(4), 502–519. https://doi.org/10.1080/19439342.2011.605160

C2.P121 Dal-Ré, R., Ioannidis, J. P., Bracken, M. B., Buffler, P. A., Chan, A.-W., Franco, E. L., La Vecchia, C., & Weiderpass, E. (2014). Making prospective registration of observational research a reality. *Science Translational Medicine*, *6*(224), 224cm1. https://doi.org/10.1126/scitranslmed.3007513

C2.P122 De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P. M., Schroeder, T. V., & Sox, H. C. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *351*(12), 1250–1251. https://doi.org/10.1056/NEJMe048225

C2.P123 Deere, D., Murphy, K. M., & Welch, F. (1995). Employment and the 1990-1991 minimum-wage hike. *American Economic Review*, *85*(2), 232–237.

C2.P124 DeLong, J. B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, *100*(6), 1257–1272.

C2.P125 Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in empirical economics: The Journal of Money, Credit and Banking project. *American Economic Review*, *76*(4), 587–603.

C2.P126 Doucouliagos, H., & Stanley, T. D. (2009). Publication selection bias in minimum-wage research? A meta-regression analysis. *British Journal of Industrial Relations*, *47*(2), 406–428. https://doi.org/10.1111/j.1467-8543.2009.00723.x

C2.P127 Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61: Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier. http://www.sciencedirect.com/science/article/pii/S1573447107040612

C2.P128 Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

C2.P129 Epidemiology Editors. (2010). The registration of observational studies—When metaphors go bad. *Epidemiology*, 21(5), 607–609. https://doi.org/10.1097/EDE.0b013e3181eafbcf

C2.P130 Findley, M., Jensen, N. M., Malesky, E. J., & Pepinsky, T. B. (2016). Introduction: Special issue on research transparency in the social science. *Comparative Political Studies*.

C2.P131 Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness*, *6*(1), 44–57. https://doi.org/10.1080/19439342.2013.875054

C2.P132 Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., & Baicker, K. (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics*, *127*(3), 1057–1106. https://doi.org/10.1093/qje/qjs020

C2.P133 Foster, A., Karlan, D., & Miguel, E. (2018, March 9). Registered reports: Piloting a pre-results review process at the *Journal of Development Economics*. https://blogs.worldbank.org/impactevaluations/registered-reports-piloting-pre-results-review-process-journal-development-economics

C2.P134 Foster, A., Karlan, D., Miguel, E., & Bogdanoski, A. (2019, July 15). Pre-results review at the *Journal of Development Economics*: Lessons learned so far. https://blogs.worldbank.org/impactevaluations/pre-results-review-journal-development-economics-lessons-learned-so-far

C2.P135 Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. https://doi.org/10.1126/science.1255484

C2.P136 Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*, *7*(1), 8–12. https://doi.org/10.1177/1948550615598377

C2.P137 Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine*, *7*(8), 889–894. https://doi.org/10.1002/sim.4780070807

C2.P138 Ganimian, A. (2014). *Pre-analysis plan template*. http://scholar.harvard.edu/files/alejandro_ganimian/files/pre-analysis_plan_template_0.pdf

C2.P139 Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

C2.P140 Gerber, A., & Malhotra, N. (2008a). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*(3), 313–326. https://doi.org/10.1561/100.00008024

C2.P141 Gerber, A., & Malhotra, N. (2008b). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*(1), 3–30. https://doi.org/10.1177/0049124108318973

C2.P142 Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis*, *9*(4), 385–392.

C2.P143 Glennerster, R., & Takavarasha, K. (2013). *Running Randomized Evaluations: A Practical Guide*. Princeton University Press.

C2.P144 Harvey, C. R., Liu, Y., & Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies*, *29*(1), 5–68. https://doi.org/10.1093/rfs/hhv059

C2.P145 Hawkins, C. B., Fitzgerald, C. E., & Nosek, B. A. (2015). In search of an association between conception risk and prejudice. *Psychological Science*, *26*(2), 249–252. https://doi.org/10.1177/0956797614553121

C2.P146 Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, *7*(2), 246–255.

C2.P147 Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*(4), 299–332. https://doi.org/10.3102/10769986021004299

C2.P148 Hirshleifer, S., McKenzie, D., Almeida, R., & Ridao-Cano, C. (2016). The impact of vocational training for the unemployed: Experimental evidence from Turkey. *Economic Journal*, 126(597), 2115–2146. https://doi.org/10.1111/ecoj.12211

C2.P149 Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

C2.P150 Humphreys, M., Sierra, R. S. de la, & Windt, P. van der. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, *21*(1), 1–20. https://doi.org/10.1093/pan/mps021

C2.P151 Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, *53*(1), 11–22. https://doi.org/10.2307/2533093

C2.P152 Ioannidis, J. P. (2008). Effectiveness of antidepressants: An evidence myth constructed from a thousand randomized trials? *Philosophy, Ethics, and Humanities in Medicine*, *3*(1), 14. https://doi.org/10.1186/1747-5341-3-14

C2.P153 Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

C2.P154 ISIS-2 Collaborative Group. (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*, *332*(8607), 349–360. https://doi.org/10.1016/S0140-6736(88)92833-4

C2.P155 Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences USA*, *111*(24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

C2.P156 Katz, L., Duflo, E., Goldberg, P., & Thomas, D. (2013, November 18). *AEA e-mail announcement*. https://www.aeaweb.org/announcements/20131118_rct_email.php

C2.P157 Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS Medicine*, *5*(2), e45. https://doi.org/10.1371/journal.pmed.0050045

C2.P158 Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, *75*(1), 83–119. https://doi.org/10.1111/j.1468-0262.2007.00733.x

C2.P159 Knittel, C. R., & Metaxoglou, K. (2011). Challenges in merger simulation analysis. *American Economic Review*, *101*(3), 56–59.

C2.P160 Knittel, C. R., & Metaxoglou, K. (2013). Estimation of random-coefficient demand models: Two empiricists' perspective. *Review of Economics and Statistics*, *96*(1), 34–59. https://doi.org/10.1162/REST_a_00394

C2.P161 Laine, C., Horton, R., DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Godlee, F., Haug, C., Hébert, P. C., Kotzin, S., Marusic, A., Sahni, P., & Schroeder, T. V. (2007). Clinical trial registration—Looking back and moving ahead. *New England Journal of Medicine*, *356*(26), 2734–2736. https://doi.org/10.1056/NEJMe078110

C2.P162 LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*(4), 604–620.

C2.P163 Lancet. (2010). Should protocols for observational research be registered? *Lancet*, *375*(9712), 348. https://doi.org/10.1016/S0140-6736(10)60148-1

C2.P164 Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, *73*(1), 31–43.

C2.P165 Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives*, *24*(2), 31–46. https://doi.org/10.1257/jep.24.2.31

C2.P166 Leamer, E. E. (2016). S-values: Conventional context-minimal measures of the sturdiness of regression coefficients. *Journal of Econometrics*, *193*(1), 147–161. https://doi.org/10.1016/j.jeconom.2015.10.013

C2.P167 Lee, S., & Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics*, *29*(4), 612–626. https://doi.org/10.1002/jae.2327

C2.P168 Levine, D. I. (2001). Editor's introduction to "The unemployment effects of minimum wages: Evidence from a prespecified research design." *Industrial Relations: A Journal of Economy and Society*, *40*(2), 161–162. https://doi.org/10.1111/0019-8676.00204

C2.P169 List, J. A., Shaikh, A. M., & Xu, Y. (2016). *Multiple Hypothesis Testing in Experimental Economics* (Working Paper No. 21875). National Bureau of Economic Research. http://www.nber.org/papers/w21875

C2.P170 Loder, E., Groves, T., & MacAuley, D. (2010). Registration of observational studies: The next step towards research transparency. *BMJ: British Medical Journal*, *340*, c950. https://doi.org/10.1136/bmj.c950

C2.P171 Maggioni, A. P., Darne, B., Atar, D., Abadie, E., Pitt, B., & Zannad, F. (2007). FDA and CPMP rulings on subgroup analyses. *Cardiology*, *107*(2), 97–102. https://doi.org/10.1159/000094508

C2.P172 Maley, S. (2014). Statistics show no evidence of gender bias in the public's hurricane preparedness. *Proceedings of the National Academy of Sciences USA*, *111*(37), E3834. https://doi.org/10.1073/pnas.1413079111

C2.P173 Malter, D. (2014). Female hurricanes are not deadlier than male hurricanes. *Proceedings of the National Academy of Sciences USA*, *111*(34), E3496. https://doi.org/10.1073/pnas.1411428111

C2.P174 Mathieu, S., Boutron, I., Moher, D., Altman, D. G., & Ravaud, P. (2009). Comparison of registered and published primary outcomes in randomized controlled trials. *Journal of the American Medical Association*, *302*(9), 977–984. https://doi.org/10.1001/jama.2009.1242

C2.P175 McCrary, J., Christensen, G., & Fanelli, D. (2015). *Conservative Tests under Satisficing Models of Publication Bias* .

C2.P176 McCullough, B. D., & Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, *93*(3), 873–892.

C2.P177 McManus, W. S. (1985). Estimates of the deterrent effect of capital punishment: The importance of the researcher's prior beliefs. *Journal of Political Economy*, *93*(2), 417–425.

C2.P178 McNutt, M. (2016). Taking up TOP. *Science*, *352*(6290), 1147–1147. https://doi.org/10.1126/science.aag2359

C2.P179 Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

C2.P180 Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Laan, M. V. der. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

C2.P181 Monogan, J. E. (2013). A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis*, *21*(1), 21–37. https://doi.org/10.1093/pan/mps022

C2.P182 Neumark, D. (2001). The employment effects of minimum wages: Evidence from a prespecified research design. *Industrial Relations: A Journal of Economy and Society*, *40*(1), 121–144. https://doi.org/10.1111/0019-8676.00199

C2.P183 Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

C2.P184 Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

C2.P185 O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, *40*(4), 1079–1087. https://doi.org/10.2307/2531158

C2.P186 Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61–80. https://doi.org/10.1257/jep.29.3.61

C2.P187 Olken, B. A., Onishi, J., & Wong, S. (2012). *Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia* (Working Paper No. 17892). National Bureau of Economic Research. http://www.nber.org/papers/w17892

C2.P188 Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

C2.P189 Romano, J. P., Shaikh, A. M., & Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST*, *17*(3), 417. https://doi.org/10.1007/s11749-008-0126-6

C2.P190 Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

C2.P191 Sala-I-Martin, X. X. (1997). I just ran two million regressions. *American Economic Review*, *87*(2), 178–183.

C2.P192 Schulz, K. F., & Grimes, D. A. (2005). Multiplicity in randomised trials II: Subgroup and interim analyses. *Lancet*, *365*(9471), 1657–1661. https://doi.org/10.1016/S0140-6736(05)66516-6

C2.P193 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

C2.P194 Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

C2.P195 Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

C2.P196 Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015a). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

C2.P197 Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015b). *Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications* (SSRN Scholarly Paper No. ID 2694998). Social Science Research Network. http://papers.ssrn.com/abstract=2694998

C2.P198 Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics*, *70*(1), 103–127. https://doi.org/10.1111/j.1468-0084.2007.00487.x

C2.P199 Stanley, T. D., & Doucouliagos, H. (2010). Picture this: A simple graph that reveals much ado about research. *Journal of Economic Surveys*, *24*(1), 170–191. https://doi.org/10.1111/j.1467-6419.2009.00593.x

C2.P200 Stanley, T. D., & Doucouliagos, H. (2012). *Meta-regression Analysis in Economics and Business*. Routledge.

C2.P201 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

C2.P202 Taubman, S. L., Allen, H. L., Wright, B. J., Baicker, K., & Finkelstein, A. N. (2014). Medicaid increases emergency-department use: Evidence from Oregon's health insurance experiment. *Science*, *343*(6168), 263–268. https://doi.org/10.1126/science.1246183

C2.P203 Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*(3), 252–260. https://doi.org/10.1056/NEJMsa065779

C2.P204 Ulrich, R., & Miller, J. (2015). P-hacking by post hoc selection with multiple opportunities: Detectability by skewness test? Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*(6), 1137–1145. https://doi.org/10.1037/xge0000086

C2.P205 U.S. Food and Drug Administration. (n.d.). *E9 statistical principles for clinical trials*. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf

C2.P206 Vosgerau, J., Nelson, L. D., Simonsohn, U., & Simmons, J. P. (2019). *99% Impossible: A Valid, or Falsifiable, Internal Meta-Analysis* (SSRN Scholarly Paper No. ID 3271372). Social Science Research Network. https://papers.ssrn.com/abstract=3271372

C2.P207 Walsh, E., Dolfin, S., & DiNardo, J. (2009). Lies, damn lies, and pre-election polling. *American Economic Review*, *99*(2), 316–322. https://doi.org/10.1257/aer.99.2.316

C2.P208 Westfall, P. H., & Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons.

C2.P209 Wilson, R. (2012). Note from the editor. *American Journal of Political Science*, *56*(3), 519.

C2.P210 Wilson, R. K. (2010). Editorial. *American Journal of Political Science*, *54*(4), 837–838.

C2.P211 Wydick, B., Katz, E., & Janet, B. (2014). Do in-kind transfers damage local markets? The case of TOMS shoe donations in El Salvador. *Journal of Development Effectiveness*, *6*(3), 249–267. https://doi.org/10.1080/19439342.2014.919012