Methods for Transparent and Reproducible Economics Research

BLINDED*

September 23, 2016

Abstract

Economics and the other social sciences have been in the spotlight recently for failures of reproducibility of prominently published research. This article reviews the problems and discusses recent developments and proposed solutions. The problems of publication bias, specification searches (also called data mining or p-hacking), and failures to replicate published research have been known for decades, but a recent surge of research has documented that the problems are still quite prevalent. This has led to a spate of proposed solutions, including study registration, pre-registration of statistical analysis plans, results-blind reviewing, disclosure checklists, and journal requirements of data and code sharing. New methodological tools to test for publication bias, as well as more established tools such as meta-analysis and

*BLINDED. Funding for this manuscript was provided by an anonymous donor who play no role in writing or reviewing the manuscript, nor in the decision to publish.

multiple hypothesis testing adjustment, have become more widely adopted. Lastly, technological improvements such as version control and dynamic documents have made a reproducible research workflow much simpler.

1 Introduction

The principle that scientific claims should be subject to scrutiny by other researchers and the public at large is well established. Just examine the Royal Society's motto *nullius in verba* ("take nobody's word for it") or seminal works on the sociology of science such as Merton [1973]. An important requirement for such scrutiny is that researchers make their claims transparent in a way that other researchers are easily able to form a complete understanding of the methods that were used by the original. In economics, given the personal computing and Internet revolutions and the wide availability of data and processing power, it is essential that data, code, and analyses be transparent [Marwick, 2015].

This article is intended to be a source for empirical economics researchers who desire to make their own research transparent to, and reproducible by, others. The entire process of research is covered, from hypothesis generation to publication. A longer version of this paper, available online, covers issues more thoroughly and address the social sciences more generally in the hopes that quantitative researchers in political science, sociology, and psychology may benefit as well. [1]

While most of us are likely to presume that we ourselves would not conduct outright fraud, fraud does indeed occur. From making up fake data to creating bogus e-mail addresses so one could do one's own peer review, the Retraction Watch blog documents a distressingly large amount of deliberate fraud in research. This paper will focus on the methodological flexibility and motivated reasoning that is likely far more comon Nosek et al. [2012].

The article is laid out as follows: in section 2 I discuss one of the major problems in non-transparent research, specifically publication bias. I also discuss how this problem can

be resolved through the practice of registration. Publication bias stems from the fact that published results are overwhelmingly statistically significant. But without knowing how many tests were run (the number of unpublished results), it is impossible to know whether these significant results are meaningful, or whether they are the 5% of tests that we would expect to appear significant due to random sampling, even with no true effect. By publicly registering all studies, we can have a better idea of just how many tests have been run.

In section 3 I discuss researcher degrees of freedom and pre-analysis plans. In addition to registering trials, researchers can also specify their outcomes of interest and their exact methods of analysis to bind their hands during the analysis phase by writing a Pre-Analysis Plan (PAP). This is a relatively new idea in economics, so there is not yet a consensus on when a PAP should be required, what the ideal level of detail is, and how much it should constrain a researcher's hands in the actual analysis, but by pre-specifying analyses, researchers can distinguish between confirmatory and exploratory analysis. I do not necessarily place higher intrinsic value on one or the other, but making the distinction clear is key for appropriate interpretation.

In section 4 I discuss workflow and materials sharing, with an eye on making research replicable by others. Researchers should make their code and data publicly available so that others may repeat and verify their analysis. Making data available incentivizes researchers to make their work accurate in the first place, and makes replication easier for others, improving the scientific process, but also raises the concern of differential privacy, since steps should be taken to prevent identification of individuals in the data. I also discuss the issue of reporting standards: a standardized list of things that authors should report to help make their work reproducible.

Section 5 concludes and presents a vision for moving forward.

2 Publication Bias and Registration

2.1 Publication Bias

One of the primary drivers of the recent move towards transparency is increased awareness of publication bias. Numerous papers use collections of published studies to show that the proportion of significant results are extremely unlikely to come from any true population distribution [DeLong and Lang, 1992, Gerber et al., 2001, Ioannidis, 2005]. By examining the publication rates of null results and significant results from a large set of NSF-funded studies, Franco et al. [2014] show that the selective publication of only significant results may stem from the fact that social science researchers largely fail to write up and submit results from studies resulting in null findings, citing lack of interest or fear of rejection. This idea of rejecting, or not even submitting for review, papers with null-results, is commonly referred to as the "file drawer problem" [Rosenthal, 1979]. In fact, the percentage of null findings published in journals appears to have been decreasing over time, across all disciplines [Fanelli, 2012]. It seems unlikely that this would be an accurate reflection of the state of the universe, unless the hypotheses that scientists are testing are systematically changing over time.

If journals only publish statistically significant results, we have no idea how many of those significant results are evidence of real effects, and which are the 5% of random draws that we should expect to show a significant result with a true population effect of zero. One way to combat this problem is to require registration of all studies undertaken. Ideally we could then search the registry for studies of X on Y. If numerous studies show an effect, we have

confidence the effect is real. If 5% of studies show a significant effect, we give these outlier studies less credence.

2.2 Trial Registration

A basic definition of registration is to publicly declare *all* research that one plans on conducting. Ideally this is done in a public registry designed to accept registrations in the given research discipline, and ideally the registration takes place before data collection begins.

Registration of randomized trials has achieved wide adoption in medicine, but is still relatively new to economics. After congress passed a law in 1997 requiring the creation of a registry for FDA-regulated trials, and the NIH created clinicaltrials.gov in 2000, The International Committee of Medical Journal Editors (ICMJE), a collection of editors of top medical journals, instituted a policy of publishing only registered trials in 2005 [De Angelis et al., 2004], and the policy has spread to other journals and been generally accepted by researchers [Laine et al., 2007]. Several other countries have their own national trial registries, and the World Health Organization created the International Clinical Trials Registry Platform (ICTRP) in 2007 to automatically collect all this information in one place.

An example of the benefit of trial registries is detailed in Turner et al. [2008], which details the publication rates of studies related to FDA-approved antidepressants. (See also Ioannidis [2008].) The outcome is unfortunate: essentially all the trials with positive outcomes were published, about half of questionable-outcome studies were published, and a majority of the negative-outcome studies were unpublished a minimum of four years after the study was completed. Figure 1 shows the drastically different rates of publication, and a large amount of publication bias.

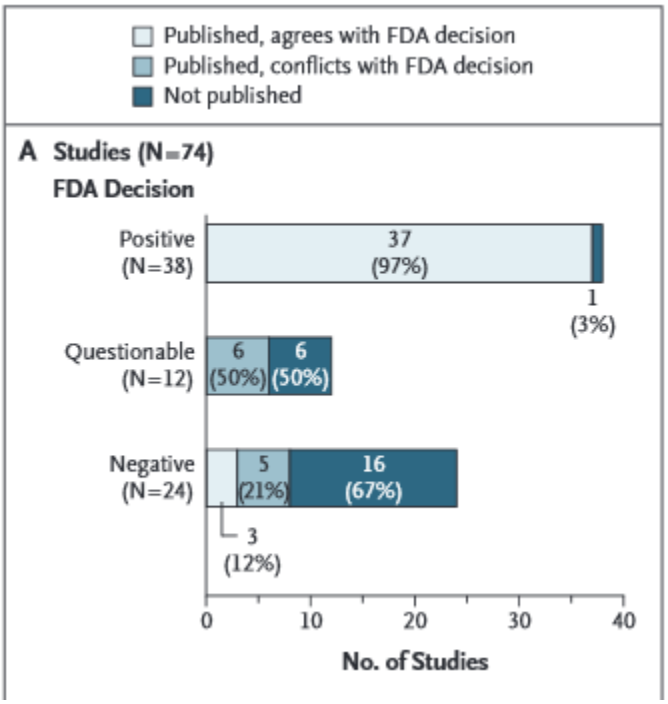

Figure 1: Panel A of Figure 1 from Turner et al. [2008]. Figure shows publication rates of studies based on statistical significance of findings.

Of course for this sort of exercise to be possible, unless a reader merely assumes that a registered trial without an associated published paper produced a null result (as in Rosenthal [1979]), it requires that the registration site itself obtain outcomes of trials. ClinicalTrials.gov is the only publicly available trial registry that requires such reporting of results, and only for certain FDA trials.[2] Hartung et al. [2014] raises concerns about discrepancies between reporting of outcomes in published papers and in the ClinicalTrials.gov database; as many as 20% of studies had discrepancies in primary outcomes and as many as 33% had discrepancies in reporting of adverse events, so there is definitely room for improvement.

Even with dramatic growth in medical trial registration, problems remain. Mathieu S et al. [2009] looked at trials related to three medical conditions and found that only 46% of studies were registered before the end of the trial with primary outcomes clearly specified. Even among those adequately registered, 31% showed some discrepancies between registered and published outcomes, with bias in favor of statistically significant definitions.

Almost all registration efforts have thus far been limited to randomized control trials, as opposed to observational data. Registering all types of analysis could be accepted, though there are definitely concerns about registering observational work—not least of which is the inability to verify that registration preceded analysis. See Dal-Re et al. [2014] for a recent discussion of the pros and cons.

2.3 Social Science Registries

Registries in economics are newer but are growing ever more popular. A brief overview of the major trial registries is shown in Table 1. The Abdul Latif Jameel Poverty Action Lab began hosting a hypothesis registry in 2009, which was superseded by the American Economic

Association's launch of its own registry for randomized trials in May 2013, which had accumulated over 540 studies in 86 countries by January 2016. The International Initiative for Impact Evaluation (3ie) launched its own registry for evaluations of development programs, the Registry for International Development Impact Evaluations (RIDIE) in September 2013, which had approximately 30 evaluations registered in its first year.

In political science, EGAP: Experiments in Governance and Politics has created a registry as "an unsupervised stopgap function to store designs until the creation of a general registry for social science research. The EGAP registry focuses on designs for experiments and observational studies in governance and politics." EGAP's registry had over 260 designs registered as of January 2016.[3]

Another location for registrations is the Open Science Framework (OSF), created by the Center for Open Science. The OSF serves as a broad research management tool that encourages and facilitates transparency (see Nosek et al. [2012].) Registrations are simply unalterable snapshots of research frozen in time, with a persistent URL and timestamp. Researchers can upload their data, code, hypotheses, etc. to the OSF, register it, and then share the resulting URL as proof of registration. OSF registrations can be relatively free-form, but templates exist to conform to standards in different disciplines. Psychology registrations are presently the most numerous on the OSF [4]

2.4 Meta-Analysis Research

Another method of detecting and dealing with publication bias is to conduct meta-analysis. This method of research collects all published findings on a given topic, analyzes the results collectively, and can detect, and attempt to adjust for, publication bias in the literature. A hand-

Table 1: Registries in Medicine and the Social Sciences

Registry	**Sponsor**	**Year Started**	**Study Design**	**Field**	**Studies Registered**
ClinicalTrials.gov	National Institutes of Health	2000	RCT	Medicine	206,000+
International Clinical Trials Registry Platform	World Health Organization	2007	RCT	Medicine	284,000+
AEA RCT Registry (SocialScienceRegistry.org)	American Economic Association	2013	RCT	Economics	540+
Registry for International Development Impact Evaluations	3ie	2013	Any	Developing country impact evaluation	75+
EGAP Design Registry	Experiments in Governance and Politics	2012	Any	Political Science	260+
Open Science Framework	Center for Open Science	2013	Any	Any	3500+

ful of organizations specialize in producing these systematic reviews, including the Cochrane Collaboration for health studies and the Campbell Collaboration for crime & justice, education, international development, and social welfare research, and the International Initiative for Impact Evalution (3ie) for social and economic interventions in low- and middle- income countries. The US government supports this type of analysis: the Department of Education's Institute of Education Sciences maintains the What Works Clearinghouse, and the Department of Labor maintains the Clearinghouse for Labor and Evaluation Research (CLEAR), which serve to collect and grade the evidentiary value of research on education and labor, respectively. (The synthesis method in the government clearinghouses is not quite as formally statistical in nature as the previously mentioned Collaborations.)

Although quite common in medical research, the tool is not widely used in some parts of the social sciences. But even in economics, where many graduate students are unfamiliar with the technqiue, important papers exist that have quantitatively synthesized bodies of literature. The unemployment effects of the minimum wage were meta-analyzed in Card and Krueger [1995], and the returns to education in Ashenfelter et al. [1999]. A meta-analysis of 87 meta-analyses in economics shows that publication bias is widespread, but not universal. A helpful resource, the Meta-Analysis of Economics Research Network (MAER-Net) which includes meta-analysis datasets and guidelines for economics researchers interested in conducting a meta-analysis is available here. Also see Stanley [2005], which helpfully describes the tools of meta-analysis, and is part of a special issue of *The Journal of Economic Surveys* dedicated to meta-analysis. Sol Hsiang, lead author of a prominent meta-analysis of 60 studies measuring the effect of climate on human conflict [Hsiang et al., 2013], has also developed the Distributed Meta-Analysis System, an online tool to crowdsource and simplify meta-analysis.

In psychology Simonsohn et al. [2014] also developed a meta-analysis tool called the "p-curve" that researchers can use to gauge the evidentiary value of a set of studies by analyzing the distribution of p-values and comparing it to what we would observe under true null effects and likely selectively reported results.

3 Researcher Degrees of Freedom

Though registration helps solve the problem of publication bias, it does not solve the problem of fishing for statistical significance within a given study. This problem with research is known as data mining: repeated searching through statistical or regression models unknowingly (or deliberately) until significance is obtained. Simmons et al. [2011] refer to this as exploiting "researcher degrees of freedom," and it has also been referred to as "fishing," "p-hacking," or "specification searching" [Humphreys et al., 2013]. The problem has many names because it can take many shapes.

Using flexibility around when to stop collecting data, excluding certain observations, combining and comparing certain conditions, including certain control variables, and combining or transforming certain measures, 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" is also described in Humphreys et al. [2013] who use simulations to show that multiplicity of outcome measures, multiplicity of heterogeneous treatment effects (sub-group analyses), and multiplicity of cut-points for turning a continuous outcome variable into a binary outcome, can all be used to virtually guarantee a false positive, even with large sample sizes. They also find that selective adding of covariates can produce false positives with small

samples, though they do find little room to produce false positives through arbitrary selection of model for binary outcomes (linear, logit, or probit) regardless of sample size. Gelman and Loken [2013] agree that "[a] dataset can be analyzed in so many different ways (with the choices being not just what statistical test to perform but also decisions on what data to exclude or exclude [sic], 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." What can be done to solve it? Part of the answer lies in detailed pre-analysis plans, described below.

3.1 Pre-Analysis Plans

While registration of studies can help to reduce publication bias or the file drawer problem, a pre-analysis plan (PAP), a detailed outline of the analyses that will be conducted in a study, can be used to reduce researcher degrees of freedom. Registration is now the norm in medical trials, and these often include (or link to) prospective statistical analysis plans as part of the project protocol. Official guidance from the US Food and Drug Administration's Center for Drug Evaluation and Research (CDER) from 1998 describes what should be included in a statistical analysis plan and discusses eight items related to data analysis that should be considered: pre-specification of the analysis; analysis sets; missing values and outliers; data transformation; estimation, confidence intervals, and hypothesis testings; adjustment of significance and confidence levels; subgroups, interactions, and covariates; and integrity of data and computer software validity [Food and Drug Administration, 1998]. This is an excellent start, and in the section below I discuss adapting these ideas for a pre-analysis plan in the social sciences.

A PAP contains a specification of the outcomes of the study (sometimes referred to as

endpoints in the medical literature), as well as a specification of the methods that will be used to analyze the outcomes. By describing the method(s) of analysis ahead of time, and to some degree tying the hands of the researcher, we reduce the ability to data mine. Though one example of this exists in economics from 2001 [Neumark, 2001], the idea is still quite new to the social sciences. The level of detail varies widely, and the research community is still constructing norms for incorporating these documents into final analyses and papers.

What to Include Suggestions have been made for the detailed contents of these documents. David McKenzie of the World Bank Research Group proposed a list of ten items that should be included in a PAP, reproduced below from the World Bank Development Impact Blog.[5]

1. Description of the sample to be used in the study
2. Key data sources
3. Hypotheses to be tested throughout the causal chain
4. Specify how variables will be constructed
5. Specify the treatment effect equation to be estimated
6. What is the plan for how to deal with multiple outcomes and multiple hypothesis testing?
7. Procedures to be used for addressing survey attrition
8. How will the study deal with outcomes with limited variation?
9. If you are going to be testing a model, include the model
10. Remember to archive it

Expecting the Unexpected Glennerster and Takavarasha [2013] also mention the "tension between the benefits of the credibility that comes from tying ones hands versus the benefits of

flexibility to respond to unforeseen events and results." Writing a PAP can lend extra credibility to research by making it confirmatory in nature as opposed to exploratory. Both types of research are absolutely valuable, but knowing the distinction is important. If some sort of restriction on the data is specified ahead of time based on theory or previous research, be it a specific functional form, exclusion of outliers, or an interaction term (subgroup analysis) that turns a null effect for the population into a significant effect for some subgroup, this can be considered confirmatory research.

Some would say this is of more value than the exploratory research approach of simply running 20 sub-group analyses and finding that one or two are significant. This may be an estimate of a true effect, but should be labeled as exploratory, and future researchers could attempt to confirm this finding by addressing the question of the sub-group specifically. The potential downside to pre-stating hypotheses and analysis plans is that no matter how carefully researchers plan ahead, something truly unexpected can occur. (An example discussed at a recent conference was subjects showing up for an experiment intoxicated. One example from a field experiment involved fatalities from a lightning strike at a school that was part of a competitive girls scholarship program [Kremer et al., 2009].) This is why, even though I may use the phrase "bind our hands," I suggest that researchers not be punished for conducting research outside the analysis plan. I instead recommend that researchers clearly delineate which analysis was included in the analysis plan, and which was not, so that readers can know what is confirmatory and what is exploratory.

When to Write There is some question as to when one should write one's pre-analysis plan. "Before you begin to analyze your data" seems like the obvious answer, but this should

be precisely defined. One could write the PAP before any baseline survey takes place, after any intervention but before endline, or after endline but before analysis has begun. Glennerster and Takavarasha [2013] and Olken [2015] have an informative discussion of the relative values of PAP timing. If one writes the PAP before the baseline, this is in some sense the purest, most free from accusations of p-hacking, but one could also miss valuable information. For example, suppose in baseline one learns that the intended outcome question is phrased poorly and elicits high rates of non-response, or that there is very little variation in the answers to a survey question. If the PAP was written after baseline, one could have accounted for this, but at the same time, researchers would also be free to change the scope of their analysis–for example, in the baseline survey of a field experiment designed to increase wages revealed that few of the subjects worked outside the home, the researcher could change the focus of the analysis. This is not necessarily wrong, but it changes the nature of the analysis somewhat.

PAPs could also be written after endline data has been collected but before the investigators have begun to analyze the data. Some have suggested that one could even look at baseline data from the control group only before writing the PAP. This may be problematic, however, since a researcher could learn that the control group had a particularly low or high value of a certain outcome variable, and then choose to include or not include this variable in the analysis as a result. The original research design could have been intended to analyze the increase in secondary school attendance, but looking at the control group, the researcher sees that the control group had a very high rate of attendance, making a significant difference between control and treatment (the treatment effect) unlikely. Learning this after the experiment has concluded and searching for things that might be easily different between treatment and control is more exploratory than confirmatory.

An alternative proposal discussed in Olken [2015] is to remove the treatment status variable from the dataset before looking at the data, which seems to alleviate some of the concerns. However, one could still search for sub-group analyses at this stage. If you parse the outcome data by gender, and males and females have a similar distribution, to find a differential treatment effect by gender would seem unlikely. If male and female had wildly different outcomes, it would seem like a significant interaction is more likely. This is more exploratory than confirmatory.

3.1.1 Examples

Examples of pre-analysis plans in economics are relatively rare, but several examples of published papers resulting from studies with PAP exist. The items below come from the J-PAL Hypothesis Registry; I highlight those that have publicly available final papers and make reference to the PAP in the paper.

- Casey et al. [2012] includes evidence from a large-scale field experiment on community driven development projects in Sierra Leone. The analysis finds no significant benefits. Given the somewhat vague nature of the development projects that resulted from the funding, and the wide variety of potential outcomes, finding significant results would have been relatively easy. In fact, the paper includes an example of how, if they had the latitude to define outcomes without a pre-analysis plan, the authors could have reported either large and significantly positive or negative outcomes, depending on their preferences. The paper also includes a discussion of the history and purpose of pre-analysis plans. The online appendix contains the PAP.
- Oregon expanded its Medicare enrollment through a random lottery in 2008, providing

researchers with an ideal avenue to evaluate the benefits of enrollment. Finkelstein et al. [2012], Baicker et al. [2013] and Taubman et al. [2014] show that recipients did not improve in physical health measurements, but were more likely to have insurance, had better self-reported health outcomes, utilized emergency rooms more, and had better detection and management of diabetes. Pre-analsyis plans from the project are available at the National Bureau of Economics' site devoted to the project. (See, for example, Taubman et al. [2013], Baicker et al. [2014].)

- The shoe company Toms funded a rigorous evaluation of its in-kind shoe donation program. Researchers wrote a pre-analysis plan before conducting their research, and found no evidence that shoe donations displace local purchasing of shoes. See Wydick et al. [2014], Katz et al. [2013]. The PAP is available in the JPAL Hypothesis Registry. This is one of many projects that has benefited from a pre-analysis plan because of the involvement of a group with a vested interest, such as a government or corporation. Even researchers skeptical of the need for PAPs in general admit the benefit to publicly pre-stating analysis plans when someone involved has such a clear incentive for results to go a certain direction.

- Researchers from UC San Diego and the World Bank evaluated job training programs run by the Turkish government and found only insignificant improvements and a strongly negative return on investment. See Almeida et al. [2012], Hirshleifer et al. [2014]. The PAP is available in the J-PAL Hypothesis Registry as well as the World Bank Development Impact Blog.

Additionally, Alejandro Ganimian developed a template for pre-analysis plans that instruc-

tors may find useful when teaching tranpsarency methods, or researcher themselves may find useful when developing their own pre-analysis plan.

3.1.2 Observational Pre-Analysis Plans

As with registration, part of the concern with pre-analysis plans is whether the "pre" aspect is verifiable. How can we be sure that a researcher didn't look at observational data, run a few hypothesis test, and then write the PAP? One way is to use reliably known availability dates of certain datasets such as government administrative data. Verifiable pre-specification of observational research was undertaken in Neumark [2001], which prospectively detailed analysis of the unemployment effects of the minimum wage, a question that has been debated for at least a century [Neumark et al., 2014].[6]

The federal minimum wage changed in October 1996 and September 1997. Neumark submitted a pre-specified research design consisting of the exact estimating equations, variable definitions, and subgroups that would be used to analyze the effect of the minimum wage on unemployment of younger workers using October, November, and December 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.

Neumark's hope was to eliminate the potential for bias due to "author effects." The obvious time stamp of the federal governments release of data indeed makes this possible, but the situation also benefits from the depth and intensity of the debate prior to this study. Neumark had an extensive extant literature to rely on when picking specific functional forms and

subgroup analyses to be conducted. He tested 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 and Krueger, 1992]) 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 show mostly insignificant results, but that 18 of the 80 specifications result in statistically significant decreases in employment (at the .10 or .05 level), with elasticities ranging from -0.14 to -0.3 for significant estimates and smaller (closer to zero) for insignificant estimates.

How widely adoptable is Neumark's method of pre-specification, and what do we gain from it? Thanks to pre-specification we know exactly how many hypothesis tests Neumark ran (at least 80 coefficients, and more with joint tests), so we know the context in which to appropriately interpret the p-values. This is a significant improvement over much other observational research. Of course, econometricians could still argue about the appropriateness of certain models (whether to include lags or look at contemporaneous effects, for instance).

3.1.3 Project Protocols

A project protocol can be somewhat similar to a PAP, but is distinct. A protocol is a detailed recipe or instruction manual for others to use to reproduce an experiment. Protocols are important both in helping solve researcher degrees of freedom problems by making the exact details of analysis known and help avoid selective reporting, as well as in making one's work reproducible. Protocols are standard in the medical literature, as in areas of lab science, but may be less familiar to those used to working with administrative or observational data. Lab sci-

ences are rife with examples of experiments failing to replicate because of supposedly minor changes such as the brand of bedding in mouse cages, the gender of the laboratory assistant, or the speed at which one stirs a reagent [Sorge et al., 2014, Hines et al., 2014], and the same situation may exist in the social sciences. *Nature* has decided to expand its methods section in order to encourage better reporting.[7]

The social sciences may benefit from more careful documentation of methods. When one uses administrative data this can be accomplished by sharing one's data and code so that analysis is transparent.[8] This is discussed below in Section 4. With original data collection, researchers should provide very detailed descriptions of what exactly they did. A 33-item checklist of suggested items is contained in the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) statement [Chan et al., 2013], including details on the participants, interventions, outcomes, assignment, blinding, data collection, data management, and statistical methods, among other things.

3.2 Multiple Hypothesis Testing

Several of the PAP and lists of suggestions above include corrections for multiple hypothesis testing. The idea of correcting for multiple tests is widespread in certain fields, but has yet to take hold in the social sciences. Simply put, the idea is that because we are aware of the fact that test statistics and p-values appear significant purely by chance a certain proportion of the time, we can report different, *better* p-values that control for the fact that we are running multiple tests. There are several ways to do this, a few of which are used and explained in a simple and straightforward manner by Anderson [2008]:

- Report index tests—instead of reporting the outcomes of numerous tests, standardize

outcomes and combine them into a smaller number of indexes (e.g. instead of separately reporting whether a long-term health intervention reduced blood pressure, diabetes, obesity, cancer, heart disease, and Alzheimer's, report the results of a single health index.) Kling et al. [2007] implements an index test from the Moving to Opportunity field experiment, using methods developed in biomedicine by O'Brien [1984].

- Control the Family-Wise Error Rate (FWER)—FWER is the probability that at least one true hypothesis in a group is rejected (a type I error), meaning it is advisable when the damage from incorrectly claiming *any* hypotheses are false is important. There are several ways to do this, with the simplest (but very conservative) method being the Bonferroni correction of simply multiplying every original p-value by the number of tests done. Holm's sequential method involves ordering p-values by class and multiplying the lower p-values by higher discount factors [Holm, 1979]. An efficient recent method is the free step-down resampling method, developed by Westfall and Young [1993].
- Control the False Discovery Rate (FDR)—In situations where a single type I error is not catastrophic, researchers may be willing to use a less conservative method and trade off some incorrect rejections in exchange for greater power. This is possible by controlling the FDR, or the percentage of rejections that are type I errors. Benjamini and Hochberg [1995] details a simple algorithm to control this rate at a chosen level, and Benjamini et al. [2006] describes a two-step procedure with greater power.

3.3 Subgroup Analysis

One aspect of researcher degrees of freedom related to multiple hypothesis testing that seems to have taken hold widely in the medical literature is the aversion to sub-group analysis ("in-

teractions" to most economists). Given the ability to test for a differential effect by many different groupings, crossed with each outcome variable, sub-groups analysis can almost always find some sort of supposedly significant effect. An oft-repeated story in the medical literature revolves around the publication of a study on aspirin after heart attacks. When the editors suggested including 40 subgroup analyses, the authors relented on the condition they include some of their own. Gemini and Libras had worse outcomes when taking aspirin after heart attacks, despite the large beneficial effects for everyone else. (Described in Schulz and Grimes [2005], with the original finding in ISIS-2 (SECOND INTERNATIONAL STUDY OF INFARCT SURVIVAL) COLLABORATIVE GROUP [1988]) Whether in a randomized trial or not, social scientists could benefit from reporting the number of interactions tested, possibly adjusting for multiple hypotheses, and ideally specifying beforehand the interactions to be tested, with a justification from theory or previous evidence as to why the test is of interest.

3.4 Results-Blind Reviewing

A new development in research transparency which helps to address both publication bias and researcher degrees of freedom is results-blind reviewing. In results-blind reviewing, authors submit a detailed research plan *before* conducting the research. They submit this plan to a journal, and the journal rejects or gives an in-principle acceptance of the not-yet-written article based on the scientific merit of the question being asked and the methods proposed to answer them, as opposed to whether the results pass an arbitrary threshold of statistical significance. Then the authors conduct the research, and their paper is published as long as they don't deviate too much from what they initially proposed. The editors and reviewers have tied their hands and have no ability to accept only significant results, and the authors have less incentive to

game their statistical analysis in order to find something significant.

This new mode of publication, called "Registered Reports," is championed by Chris Chambers, psychologist at Cardiff University, and has been adopted by over a dozen journals.[9] *Social Pyschology* ran an issue dedicated to this type of article, with an editorial explaining the concept [Nosek and Lakens, 2014]. *AIMS Neuroscience* published an editorial answering 25 frequently asked questions pertaining to registered reports [Chambers et al., 2014]. A forthcoming special issue of *Comparative Political Studies* will also publish articles of this type [Findley et al., 2016].

4 Replication and Reproducibility

> "Economists treat replication the way teenagers treat chastity - as an ideal to be professed but not to be practised."—Daniel Hamermesh, University of Texas at Austin Economics

Replication, in both practice and principle, is a key to social science research. I first define what exactly we mean by replication using the taxonomy developed in Hamermesh [2007] and Hunter [2001]. According to Hamermesh, replication comes in a few different shapes: pure, statistical, and scientific.

- Pure: Using the exact same data and the exact same model to see if the published results are reproduced exactly.
- Scientific: Using a different sample from a different population, and similar, but perhaps not identical model .

- Statistical: Using the same model and underlying population but a different sample. In Hamermesh's view, this is less relevant to certain fields, such as economics, where researchers are likely to already use as large a sample as is available.

One might imagine a fourth type that uses the same data but probes the data using additional robustness and sensitivity checks. Others have described replication in terms of a spectrum from full replication (independent collection of data and re-running analysis) to reproducibility, where the same data or code are re-used by other researchers [Peng, 2011], and another taxonomy is proposed in Clemens [2016]. Whatever the terminology used, replication or reproducibility, transparent research requires making data and code available to other researchers so they can try and get the same results.

4.1 Code and Workflow

Reproducing research often involves using the exact code and statistical programming done by the original researcher. To make this possible, code needs to be both (1) easily available and (2) easily interpretable. Thanks to several free and easy to use websites described below, code can easily be made available by researchers without requiring funding or website hosting. Making code easily interpretable is a somewhat more complicated task, nevertheless, the extra effort spent to make a more manageable code pays off with large dividends.

4.1.1 Publicly Sharing Code

Once analysis is complete (or even before this stage) researchers should share their data and code with the public. GitHub (`http://www.github.com`), The Center for Open Science's Open Science Framework (`http://osf.io`), and Harvard University's Dataverse (`http://`

`thedata.org`) are all free repositories for data and code that include easy to use version control.[10]

Version control is a powerful way to archive versions of files so that old versions are not lost and can be returned to if needed. The most naive way to organize a project would be to just name a file "MyAnalysis.do" and then save over it any time you or your coauthor made changes. This of course would result in a huge loss of information. Most people are probably doing some version of the "date and initials" method. Instead of simply calling one's analysis code "MyAnalysis.do" and repeatedly saving over and losing old versions, when you save a new version, you add the date to the file name: "MyAnalysis.2014.08.13.do", and if a co-author makes changes on the same day, they add their initials:"MyAnalysis.2014.08.13_GC.do" or they change the date: "MyAnalysis.2014.08.14.do". This can work on small projects, but it is not the best way to work. Why does the operating system say the file from August 13 was modified in October? What if one analysis script file is called 14 times in 9 different files—did you remember to go through and update all the calls? Gentzkow and Shapiro [2014] provide strong evidence of the inadequacy of dating and initialing : "*Not one piece of commercial software you have on your PC, your phone, your tablet, your car, or any other modern computing device was written with the "date and initial" method.*" [emphasis original]

The solution is version control (also referred to as revision control). Version control is free software (Git and Mercurial are popular and free implementations) that stores all versions of files you create, and can easily compare versions of the files to highlight changes, revert to earlier versions, accept or reject changes proposed by a collaborator, and reconcile conflicting versions when different people make changes to the same file. Version control creates a repository, and stores all versions of your files in the repository[11] Web services such as GitHub offer

free hosting of these repositories with easy to use web-based interfaces and GUI software, so that you and your collaborators can all access the same files, but you can easily host the repository on a server of your own.

4.1.2 Managing Workflow

Code is just one aspect of a larger structure referred to as "workflow" in Long [2008], by which is meant the combination of data, code, organization, and documentation: everything from file and variable names to folder organization as well as efficient and readable programming, and data storage and documentation. Stata-users should find Long [2008] useful, while R users can refer to Gandrud [2013] for workflow recommendations both general and specific to their respective programming language.

Another valuable, more brief work is Matthew Gentzkow and Jesse Shapiro's manual on code and data [Gentzkow and Shapiro, 2014]. They come from the same background as many economists–they did not take many programming classes, they are interested in tools only to accomplish their applied research goals. Once their data became massive, and their programming problems became massive with it, they figured they should listen to the fulltime programmers and database managers who had spent years and billions of dollars solving these problems. Their manual adapts many of these solutions to data-based social science research.

I also refer undergraduate instructors to Richard Ball and Norm Medeiros' Project TIER (Teaching Integrity in Empirical Research), which is a "protocol for comprehensively documenting all the steps of data management and analysis that go into an empirical research paper." A specific file organization using the Open Science Framework is taught so that teachers can exactly reproduce the work of every student (and so students can reliably get the same

answer every time they conduct their analysis).

Software: The movement by many towards open source software such as R and Python may lead to reproducibility and access gains over time. However, many disciplines have long traditions of using proprietary software such as SAS and Stata, and learning a new programming language may be an undesirable additional task in researchers' busy lives. That said, there are several general coding rules that all researchers should use when organizing and implementing their analysis, and researchers should strive to make their work usable by as many others as possible.

Writing Code: Perhaps the most important rule is to write code instead of working by hand. By that I mean:

- Do not modify data by hand, such as with a spreadsheet. Which is to say, don't use Excel.
- Use neither the command line nor drop-down menus nor point-and-click options in statistical software.
- Instead, do everything with scripts.

The simple reason for this is reproducibility. Modifying data in Excel or any similar spreadsheet program leaves no record of the changes made to the data, nor any explanation of the reasoning or timing behind any changes. Although it may seem easy or quick to do a one-time-only cleaning of data in Excel, or make "minor" changes to get the data into a format readable by a researcher's preferred statistical software, unless these changes are written down in excruciating detail, this is not reproducible by other researchers. It is better to write a pro-

gramming script that imports the raw data, does all necessary changes, with comments in the code that explain changes, and saves any intermediate data sets used in analysis. Then, researchers can share their initial raw data and their code, and other researchers can reproduce their work exactly.

Though a fair amount of research has been done using pull down menus in SPSS or Stata, it generally makes research less reproducible. A bare minimum if one insists on going this route is to use the built-in command-logging features of the software. In Stata this involves the 'cmdlog' command; in SPSS this involves the paste button to add to a syntax. The ideal is to make everything, including changes like rounding and formatting, done with scripts. Even downloading of data can be done through a script.[12]

Another important way to prevent unintentional changes to data is to always set the seed for random number generators whenever any random numbers are to be used.[13] Additionally, information about the exact software version used should be included (include the 'version' command in Stata, or use the session.info() command in R) as well as computer processor and operating system information. The casual programmer may assume that sophisticated software would always produce the exact same answer across multiple versions of software and platforms, but this is not the case. This is also definitely not the case with user-written packages. R users can use the packageVersion() command, and can run old versions of packages since they are archived at CRAN. Stata users can use the viewsource command for any .ado they use, but since the Statistical Software Components (SSC) unfortunately does not archive old versions, reproducibility may be lost, so ideally researchers would include the actual code for the version of the user-written .ado along with their publicly archived data and code files

Finally, two simple organizing principles to consider are:

1. Consider not saving statistical output, and just saving the code and data that generates it. Obviously this would be unrealistically time consuming for large projects, but the idea is that you should be able to reproduce all steps of your analysis such that you could in theory take this approach.

2. What would happen if you, or your laptop hard drive, were hit by a bus? How easily would anyone else be able to reproduce your work? Hopefully the probability is non-zero.

4.1.3 Dynamic Documents

In addition to making code available to the public, the code itself should be written in a reader-friendly format, referred to as "Literate Programming," introduced in Knuth [1984] and Knuth [1992]. The basic idea is that "the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be *works of literature*... Instead of imagining that our main task is to instruct a *computer* what to do, let us concentrate rather on explaining to *human beings* what we want a computer to do." [emphasis original] Simply put, code should be written in as simple and easily understood a way as possible, and should be very well commented, so that researchers other than the original author can more easily understand the goal of the code.

One way to make literate (statistical) programming significantly easier is with dynamic documents, which combine code and output into one automated document. A prominent system is Knitr (see Xie [2013, 2014]), which is built into R Studio[14]. Knitr uses R Markdown[15] in which one writes both code and comments that is automatically spun into an easily read and shareable HTML, PDF, or MS Word document. These can be posted and shared for free

at RPubs, an easy to use hosting service by Rstudio. For Stata users, dynamic documents are slightly less well developed, but E.F. Haghish is actively developing packages (Markdoc and Weaver) that allow users to write their .do files in such a way that the log files output by Stata are formatted and readable in Markdown, HMTL, or LaTeX.

4.2 Sharing Data

In addition to code, researchers should share their data if at all possible. Many journals do not require sharing of data, but the number that do is increasing. Most recently, a consortium of top medical journal editors has proposed a new data-sharing requirement that, if adopted, could drastically improve data availability in medical research [Taichman et al., 2016].

4.2.1 The JMCB Project and Economics

In the field of economics, few, if any journals required sharing of data before "The Journal of Money, Credit, and Banking Project," published in *The American Economic Review* in 1986 [Dewald et al., 1986]. *The Journal of Money, Credit, and Banking* started the *JMCB Data Storage and Evaluation Project* with NSF funding in 1982, which requested data and code from authors who published in the journal. With a great deal of research funded by the NSF, it should be noted that the NSF has long had an explicit policy of expecting researchers to share their primary data[16]. Despite this, and despite the explicit policy of the *Journal* during the project, at most only 78% of authors provided data to the authors within six months after multiple requests. (This is admittedly an improvement over the 34% from the control group—those who published before the *Journal* policy went into effect—who provided data.) Of the papers that were still under review by the *Journal* at the time of the requests for data, one

quarter did not even respond to the request, despite the request coming from the same journal considering their paper. The submitted data was often an unlabeled and undocumented mess. Despite this, the authors attempted to replicate nine papers, and often were completely unable to reproduce published results, despite detailed assistance from the original authors.

Unfortunately, nothing much changed with the publication of this important article. A decade later, in a follow-up piece to the JMCB Project published in the Federal Reserve Bank of St. Louis *Review* [Anderson and Dewald, 1994], the authors note that only two economics journals other than the *Review* itself (*Journal of Applied Econometrics, Journal of Business and Economic Statistics*) requested data from authors, and neither requested code. The JMCB itself discontinued the policy of requesting data in 1993, though it resumed requesting data in 1996. The authors repeated their experiment in 1992, and obtained similar response rates as those from the original JMCB Project. The flagship economics journal, the *American Economic Review* (AER), did not start requesting data until 2003. Then after a 2003 article showed that nonlinear maximization methods often produce wildly different estimates across different software packages, that not a single AER article tested their solution with different software, and that fully half of queried authors from a chosen issue of the AER, including a then editor of the journal, failed to comply with the policy of providing data and code, editor Ben Bernanke made the data and code policy mandatory in 2004 [McCullough and Vinod, 2003, Bernanke, 2004].

The current data policy from the *American Economic Review* can be seen here: `https://www.aeaweb.org/aer/data.php`. The AER conducted a self-review and found good, but incomplete, compliance [Glandon, 2010]; others believe much work remains [Anderson et al., 2008]. In addition to all the journals published by the American Economic Association, several

top journals now explicitly require data and code to be submitted at the time of publication. The last of the "top 5" general interest journals, the *Quarterly Journal of Economics*, adopted a data-sharing requirement in 2016.

4.2.2 General Repositories

The previous section on the *JMCB* describes only a few journals in one field of the social sciences. Even if the journal to which you submit your research does not require you to supply them with your code and data, researchers should still share these materials. Though some repositories, particularly Harvard's Dataverse, seem equipped to handle data from practically any researcher (a free 1 TB of storage is standard, with more possible upon request), many repositories specialize. The Registry of Research Data Repositories has described over 900 data repositories to help you find the right data repository for your data. A key advantage to using a trusted repository such as one listed there, in lieu of simply posting the data on your own website or making your Dropbox folder public, is that many of these repositories will take your data in its proprietary (Stata, SAS, SPSS, etc.) form, and make it accessible in other formats. Storing your data in a repository with other similar datasets also makes it easier for others to find your data, instead of requiring that they already know of its existence, as would likely be the case with personal websites. Your own personal website is also more likely to be taken offline, should a researcher change schools or retire.

4.2.3 Differential Privacy

One important caveat to making data widely available is that despite anonymization, in the age of big data, sometimes individual subjects can easily be identified. Heffetz and Ligett [2014]

recount deliberate data releases by Yahoo! Inc., the Massachusetts state government, and Netflix, that could easily be used to identify individuals in the data, despite the absence of direct identifiers such as names or social security numbers. The problem is that "de-identification does not guarantee anonymization." This problem is well known in computer science, but solutions are still being developed and are not widely implemented.

4.3 Reporting Standards

In research, the devil is in the details. Whether assessing the validity of a research design or attempting to replicate a study, details of what exactly was done must be recorded and made available to other researchers. The exact details that are relevant will likely differ from field to field, but an increasing number of fields have produced centralized checklists that describe (in excruciating detail) what disclosure is required of published studies. These checklists are not often published with the paper, but can be submitted with the original article so that reviewers can check that it has been completed. With nearly infinite and easy web storage, researchers can easily post these materials in a repository even if journal editors insist on cutting their methods sections for space reasons.

4.3.1 Randomized Trials and CONSORT

The most widely adopted reporting standard guideline is the Consolidated Standards of Reporting Trials (CONSORT). Reporting standards evolved parallel to construction of clinicaltrials.gov and registration, and are now nearly universally adopted for randomized trials published in medical journals, required or requested by reviewers during the review process. This is still in its infancy in the social sciences.

The original CONSORT was developed in the mid 1990's [Begg C et al., 1996]. After five years, research showed that reporting of essential details had significantly increased in journals requiring the standard [Moher D et al., 2001]. The statement was revised in 2001, and again in 2010 [Moher et al., 2001, Schulz et al., 2010]. The statement is a 25-item checklist pertaining to the title, abstract, introduction, methods, results, and discussion of the article in question, and seeks to delineate the minimum requirements of disclosure that may not be sufficiently addressed through other measures.

4.3.2 Social Science Reporting Standards

Though a standard akin to CONSORT has not been formally adopted by social science or economics journals, at least as far as we are aware, there have been attempts to do this: in political science, the Experimental Research Section Standards Committee produced a detailed list of items required for disclosure of experiments [Gerber et al., 2014]. This checklist is available here.[17] In economics, one article has highlighted the fact that there is limited discussion of essential features of randomization (How was randomization stratified, if at all? How were control variables determined?), but no standards have been adopted [Bruhn and McKenzie, 2009]. In psychological and behavioral research, an extention to CONSORT for Social and Psychological Interventions (CONSORT-SPI) has been developed in [Montgomery et al., 2013], but so far has not been widely adopted nor required by journals.

4.3.3 Observational Reporting Standards

Social science has yet to make a serious push for reporting standards in RCTs, let alone observational work, but the medical/epidemiological literature has created standards in this type of

work, though they are not as widely adopted as CONSORT. Perhaps the most well-known is the STROBE Statement (Strengthening the Reporting of Observational Studies in Epidemiology)[Von Elm et al., 2007]. STROBE provides checklists for reporting of cohort, case-control, and cross-sectional studies. These standards have been endorsed by approximately 100 journals in the field.[18]

Medicine has in fact come up with too many checklists to describe them all individually. Acknowledging that every field and type of research is different, the Equator Network (Enhancing the Quality of Transparency of Health Research) serves as an umbrella organization that seeks to keep tabs on all the best reporting standards and help researchers find which reporting standard is most relevant for their research.[19]

5 Conclusion

If science progresses by standing on the shoulders of giants, then good science requires research to be conducted in a transparent and reproducible fashion. This may require extra up-front work by researchers compared to the current state of affairs. Before one runs an experiment, researchers could write down their hypothesis, carefully explain how they are going to test the hypothesis, perhaps going as far as writing down the very regression analysis specification they plan to run, write a detailed protocol of the exact experimental setting, and then post all of this publicly on the Internet in a public repository. In the long run, however, science will reap a reward greater than these costs. Statistical analysis will result in p-values that can be interpreted as intended, and in an appropriate publication bias context. Replicating the work of other researchers will be easy, because their data and code will be in a public repository, and

code could be well documented and understandable, and written to easily recreate the original results, ideally with a single click. Science could potentially move forward at a faster rate.

Notes

[1]See http://www.github.com/BLINDED/bestpracticesmanual.

[2]Prayle et al. [2012] finds that compliance with results reporting even among those required was fairly low (22%). HHS and NIH took steps in November 2014 to expand the amount of results reporting required. See http://www.nih.gov/news/health/nov2014/od-19.htm

[3]Earlier less-widely adopted attempts to create registries in political science are the Political Science Registered Studies Dataverse (PSRSD, http://spia.uga.edu/faculty_pages/monogan/registration.php) and the PAP Registry of the Experimental Research section of the American Political Science Association (http://ps-experiments.ucr.edu/browser).

[4]See https://osf.io/explore/activity/#newPublicRegistrations.

[5]A similar list also appears in Glennerster and Takavarasha [2013] .

[6]An editorial introduction to Neumark [2001] was accidentally left out of the issue, and published the following month, see Levine [2001]. Others followed Neumark's paper in spirit by using the same regression specifications on data from Canada [CAMPOLIETI et al., 2006].

[7]http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852

[8]It should be noted that the need for documentation of survey method is not eliminated by using administrative data, the burden simply falls upon the administration.

[9]See a full list of journals adopting the procedure at https://osf.io/8mpji/.

[10]BitBucket (http://www.bitbucket.org) is another web service that one can use for free version control and archiving of public data and code.

[11]Repositories can take up a significant amount of space, but you can avoid this problem by not storing generated files or binaries (.pdf, .docx, .etc) since every user with access to the repository should be able to recompile and generate files on their own. Version control works best with simple text files.

[12]For example, in R, the download.file() function can be used to save data from a website. Though of course this opens the possibility to the data file changing. When reproducing results from a given

dataset is more important than the data from a specific source, researchers should download their raw dataset once, and never save over it, instead saving all modified intermediate datasets in a separate location.

[13]set.seed() in R, set seed () in Stata.

[14]R Studio is a popular free graphical integrated implementation of R, available at http://www.rstudio.com.

[15]R Markdown is a very simple plain text markup language, described at `http://rmarkdown.rstudio.com/`)

[16]"Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing." See http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[17]`http://www.davidhendry.net/research-supplemental/gerberetal2014-reportingstandards/gerberetal2014-reportingstandards&appendix1.pdf`

[18]`http://www.strobe-statement.org/index.php?id=strobe-endorsement`

[19]See `http://www.equator-network.org/` for more information.

Bibliography

Rita Almeida, Sarojini Hirshleifer, David McKenzie, Cristobal Ridao-Cano, and Ahmed Levent Yener. The impact of vocational training for the unemployed in turkey: Pre-analysis plan. *Poverty Action Lab Hypothesis Registry*, February 2012. URL http://www.povertyactionlab.org/sites/default/files/documents/ISKURIE_AnalysisPlan_v4.pdf.

Michael L. Anderson. 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, 2008. doi: 10.1198/016214508000000841. URL http://dx.doi.org/10.1198/016214508000000841.

Richard G. Anderson and William G. Dewald. Replication and scientific standards in applied economics a decade after the journal of money, credit and banking project. *Federal Reserve Bank of St. Louis Review*, (Nov):79–83, 1994. URL http://econpapers.repec.org/article/fipfedlrv/y_3a1994_3ai_3anov_3ap_3a79-83.htm.

Richard G. Anderson, William H. Greene, B. D. McCullough, and H. D. Vinod. The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, 15(1):99–119, 2008. ISSN 1350-178X. doi: 10.1080/13501780801915574. URL http://dx.doi.org/10.1080/13501780801915574.

Orley Ashenfelter, Colm Harmon, and Hessel Oosterbeek. A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour economics*, 6(4):453–470, 1999.

Katherine Baicker, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. The oregon experiment effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013. doi: 10.1056/NEJMsa1212321. URL http://www.nejm.org/doi/full/10.1056/NEJMsa1212321. PMID: 23635051.

Katherine Baicker, Amy Finkelstein, and Sarah Taubman. The oregon health insurance experiment: Evidence from criminal charges data, analysis plan. April 2014. URL http://www.nber.org/oregon/files/oregon_hie_crime_analysis_plan.pdf.

Begg C, Cho M, Eastwood S, and et al. Improving the quality of reporting of randomized controlled trials: The consort statement. *JAMA*, 276(8):637–639, August 1996. ISSN 0098-7484. doi: 10.1001/jama.1996.03540080059030. URL http://dx.doi.org/10.1001/jama.1996.03540080059030.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

Ben S. Bernanke. Editorial statement. *The American Economic Review*, 94(1):404–404, 2004. ISSN 00028282. URL http://www.jstor.org/stable/3592790.

Miriam Bruhn and David McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4): 200–232, October 2009. doi: 10.1257/app.1.4.200.

MICHELE CAMPOLIETI, MORLEY GUNDERSON, and CHRIS RIDDELL. Minimum wage impacts from a prespecified research design: Canada 1981 1997. *Industrial Relations: A Journal of Economy and Society*, 45(2):195–216, 2006. ISSN 1468-232X. doi: 10.1111/j.1468-232X.2006.00424.x. URL http://dx.doi.org/10.1111/j.1468-232X.2006.00424.x.

David Card and Alan B Krueger. Using regional variation in wages to measure the effects of the federal minimum wage. *Industrial & Labor Relations Review*, 46(1):22–37, October 1992.

David Card and Alan B Krueger. Time-series minimum-wage studies: a meta-analysis. *The American Economic Review*, pages 238–243, 1995.

Katherine Casey, Rachel Glennerster, and Edward Miguel. Reshaping institutions: Evidence on aid impacts using a preanalysis plan*. *The Quarterly Journal of Economics*, 127(4): 1755–1812, November 2012. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qje027. URL http://qje.oxfordjournals.org/content/127/4/1755.

Christopher D Chambers, Eva Feredoes, Suresh Daniel Muthukumaraswamy, and Peter Etchells. Instead of" playing the game" it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, 1(1):4–17, 2014.

A.-W. Chan, J. M. Tetzlaff, P. C. Gotzsche, D. G. Altman, H. Mann, J. A. Berlin, K. Dickersin, A. Hrobjartsson, K. F. Schulz, W. R. Parulekar, K. Krleza-Jeric, A. Laupacis, and D. Moher. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*, 346(jan08 15):e7586–e7586, January 2013. ISSN 1756-1833. doi: 10.1136/bmj.e7586. URL http://www.bmj.com/cgi/doi/10.1136/bmj.e7586.

Michael Clemens. The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 2016.

Rafael Dal-Re, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. Making prospective registration of observational research a reality. *Science Translational Medicine*, 6(224):224cm1–224cm1, February 2014. ISSN 1946-6234, 1946-6242. doi: 10.1126/scitranslmed.3007513. URL http://stm.sciencemag.org/content/6/224/224cm1.

Catherine De Angelis, Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, and Sheldon Kotzin. Clinical trial registration: A statement from the international committee of medical journal editors. *New England Journal of Medicine*, 351(12): 1250–1251, 2004. doi:10.1056/NEJMe048225.

Donald Deere, Kevin M. Murphy, and Finis Welch. 1995. Employment and the 1990-1991 minimum-wage hike. *American Economic Review Papers and Proceedings*, 85(2):232–237, May 1995.

J. Bradford DeLong and Kevin Lang. Are all economic hypotheses false? *Journal of Political Economy*, 100(6):1257–1272, December 1992. ISSN 0022-3808. URL http://www.jstor.org/stable/2138833.

William G. Dewald, Jerry G. Thursby, and Richard G. Anderson. Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review*, 76(4):587–603, September 1986. ISSN 0002-8282. URL http://www.jstor.org/stable/1806061.

Daniele Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, March 2012. ISSN 0138-9130. doi: 10.1007/s11192-011-0494-7. WOS:000300325800009.

Michael Findley, Nathan M. Jensen, Edmund J. Malesky, and Thomas B. Pepinsky. Introduction: Special issue on research transparency in the social sciences. *Comparative Political Studies*, 2016.

Amy Finkelstein, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, and Katherine Baicker. The oregon health insurance experiment: Evidence from the first year*. *The Quarterly Journal of Economics*, 127(3):1057–1106, August 2012. ISSN 0033-5533, 1531-4650. doi: 10.1093/qje/qjs020. URL http://qje.oxfordjournals.org/content/127/3/1057.

Food and Drug Administration. Guidance for industry: E9 statistical principles for clinical trials. *Food and Drug Administration: Rockville, Maryland, USA*, 1998. URL http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf.

Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, September 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1255484. URL http://www.sciencemag.org/content/345/6203/1502.

Christopher Gandrud. *Reproducible Research with R and R Studio*. CRC Press, 2013.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition'or p-hacking' and the research hypothesis was posited ahead of time. November 2013. URL http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Matthew Gentzkow and Jesse Shapiro. Code and data for the social sciences: A practioner's guide. March 2014. URL http://faculty.chicagobooth.edu/jesse.shapiro/research/CodeAndData.pdf.

Alan Gerber, Kevin Arceneaux, Cheryl Boudreau, Conor Dowling, Sunshine Hillygus, Thomas Palfrey, Daniel R. Biggers, and David J. Hendry. Reporting guidelines for experimental research: A report from the experimental research section standards committee. Technical report, 2014. URL http://www.davidhendry.net/research-supplemental/gerberetal2014-reportingstandards/gerberetal2014-reportingstandards&appendix1.pdf.

Alan S. Gerber, Donald P. Green, and David Nickerson. Testing for publication bias in political science. *Political Analysis*, 9(4):385–392, January 2001. ISSN 1047-1987, 1476-4989. URL http://pan.oxfordjournals.org/content/9/4/385.

Philip Glandon. Report on the american economic review data availability compliance project. Technical report, Vanderbilt University, November 2010. URL https://aeaweb.org/aer/2011_Data_Compliance_Report.pdf.

Rachel Glennerster and Kudzai Takavarasha. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, November 2013. ISBN 9781400848447.

Daniel S. Hamermesh. Viewpoint: Replication in economics. *Canadian Journal of Economics/Revue canadienne d'economique*, 40(3):715–733, August 2007. ISSN 1540-5982. doi: 10.1111/j.1365-2966.2007.00428.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2966.2007.00428.x/abstract.

Daniel M. Hartung, Deborah A. Zarin, Jeanne-Marie Guise, Marian McDonagh, Robin Paynter, and Mark Helfand. Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed PublicationsDiscrepancies between ClinicalTrials.gov and peer-reviewed publications. *Annals of Internal Medicine*, 160(7):477–483, April 2014. ISSN 0003-4819. doi: 10.7326/M13-0480. URL http://dx.doi.org/10.7326/M13-0480.

Ori Heffetz and Katrina Ligett. Privacy and data-based research. *Journal of Economic Perspectives*, 28(2):75–98, May 2014. ISSN 0895-3309. doi: 10.1257/jep.28.2.75. URL https://www-aeaweb-org.proxy.swarthmore.edu/articles.php?doi=10.1257/jep.28.2.75.

William C Hines, Ying Su, Irene Kuhn, Kornelia Polyak, and Mina J Bissell. Sorting out the facs: A devil in the details. *Cell reports*, 6(5):779–781, 2014.

Sarojini Hirshleifer, David McKenzie, Rita Almeida, and Cristobal Ridao-Cano. The impact of vocational training for the unemployed: Experimental evidence from turkey. *The Economic Journal*, pages n/a–n/a, 2014. ISSN 1468-0297. doi: 10.1111/ecoj.12211. URL http://dx.doi.org/10.1111/ecoj.12211.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):pp. 65–70, 1979. ISSN 03036898. URL http://www.jstor.org/stable/4615733.

Solomon M Hsiang, Marshall Burke, and Edward Miguel. Quantifying the influence of climate on human conflict. *Science*, 341(6151):1235367, 2013.

Macartan Humphreys, Raul Sanchez de la Sierra, and Peter van der Windt. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1):1–20, January 2013. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mps021. URL http://pan.oxfordjournals.org/content/21/1/1.

John Hunter. The desperate need for replications. *Journal of Consumer Research*, 28(1): 149–158, June 2001. ISSN 0093-5301. doi: 10.1086/jcr.2001.28.issue-1. URL http://www.jstor.org/stable/10.1086/321953.

John P. A. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, August 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371/journal.pmed.0020124.

John PA Ioannidis. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? *Philosophy, Ethics, and Humanities in Medicine*, 3(1):14, May 2008. ISSN 1747-5341. doi: 10.1186/1747-5341-3-14. URL http://www.pehmed.com/content/3/1/14/abstract.

ISIS-2 (SECOND INTERNATIONAL STUDY OF INFARCT SURVIVAL) COLLABORATIVE GROUP. RANDOMISED TRIAL OF INTRAVENOUS STREPTOKINASE, ORAL ASPIRIN, BOTH, OR NEITHER AMONG 17 187 CASES OF SUSPECTED ACUTE MYOCARDIAL INFARCTION: ISIS-2. *The Lancet*, 332(8607):349–360, August 1988. ISSN 0140-6736. doi: 10.1016/S0140-6736(88)92833-4. URL http://www.sciencedirect.com/science/article/pii/S0140673688928334.

Elizabeth Katz, Brendan Janet, Bruce Wydick, and Felipe Gutierrez. Pre-analysis plan: TOMS shoes impact study. Technical report, February 2013. URL http://www.povertyactionlab.org/sites/default/files/documents/Pre-Analysis%20Plan_Wydick_2-12-13.pdf.

Jeffrey R Kling, Jeffrey B Liebman, and Lawrence F Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

D. E. Knuth. Literate programming. *The Computer Journal*, 27(2):97–111, January 1984. ISSN 0010-4620, 1460-2067. doi: 10.1093/comjnl/27.2.97. URL http://comjnl.oxfordjournals.org/content/27/2/97.

Donald Ervin Knuth. *Literate Programming*. Center for the Study of Language and Information, January 1992. ISBN 9780937073810.

Michael Kremer, Edward Miguel, and Rebecca Thornton. Incentives to learn. *The Review of Economics and Statistics*, 91(3):437–456, 2009.

Christine Laine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug, Paul C. H bert, Sheldon Kotzin, Ana Marusic, Peush Sahni, Torben V. Schroeder, Harold C. Sox, Martin B. Van Der Weyden, and Freek W.A. Verheugt. Clinical trial registration looking back and moving ahead. *New England Journal of Medicine*, 356(26):2734–2736, June 2007. ISSN 0028-4793. doi: 10.1056/NEJMe078110. URL http://www.nejm.org/doi/full/10.1056/NEJMe078110.

David I. Levine. 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, 2001. ISSN 1468-232X. doi: 10.1111/0019-8676.00204. URL http://dx.doi.org/10.1111/0019-8676.00204.

J. Scott Long. *The Workflow of Data Analysis Using Stata*. Stata Press, December 2008. ISBN 9781597180474.

Ben Marwick. How computers broke science - and what we can do to fix it. *The Conversation*, November 2015. URL http://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938.

Mathieu S, Boutron I, Moher D, Altman DG, and Ravaud P. COmparison of registered and published primary outcomes in randomized controlled trials. *JAMA*, 302(9):977–984, September 2009. ISSN 0098-7484. doi: 10.1001/jama.2009.1242. URL http://dx.doi.org/10.1001/jama.2009.1242.

B. D. McCullough and H. D. Vinod. Verifying the solution from a nonlinear solver: A case study. *The American Economic Review*, 93(3):873–892, June 2003. ISSN 0002-8282. URL http://www.jstor.org/stable/3132121.

Robert K Merton. *The sociology of science: Theoretical and empirical investigations*. University of Chicago press, 1973.

David Moher, Kenneth F. Schulz, and Douglas G. Altman. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology*, 1(1):2, April 2001. ISSN 1471-2288. doi: 10.1186/1471-2288-1-2. URL http://www.biomedcentral.com/1471-2288/1/2/abstract.

Moher D, Jones A, Lepage L, and for the CONSORT Group. Use of the consort statement and quality of reports of randomized trials: A comparative before-and-after evaluation. *JAMA*, 285(15):1992–1995, April 2001. ISSN 0098-7484. doi: 10.1001/jama.285.15.1992. URL http://dx.doi.org/10.1001/jama.285.15.1992.

Paul Montgomery, Sean Grant, Sally Hopewell, Geraldine Macdonald, David Moher, Susan Michie, and Evan Mayo-Wilson. Protocol for consort-spi: an extension for social and psychological interventions. *Implement Sci*, 8(1):99, 2013.

David Neumark. The employment effects of minimum wages: Evidence from a prespecified research design. *Industrial Relations: A Journal of Economy and Society*, 40(1):121–144, January 2001. ISSN 1468-232X. doi: 10.1111/0019-8676.00199. URL http://onlinelibrary.wiley.com/doi/10.1111/0019-8676.00199/abstract.

David Neumark, JM Ian Salas, and William Wascher. Revisiting the minimum wage employment debate: Throwing out the baby with the bathwater? *Industrial & Labor Relations Review*, 67(3):608–648, 2014.

Brian A Nosek and Daniël Lakens. Registered reports. *Social Psychology*, 45(3):137–141, 2014.

Brian A. Nosek, Jeffrey R. Spies, and Matt Motyl. Scientific utopia II. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, November 2012. ISSN 1745-6916, 1745-6924. doi: 10.1177/1745691612459058. URL http://pps.sagepub.com/content/7/6/615.

Peter C O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087, 1984.

Benjamin A Olken. Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives*, 29(3):61–80, 2015.

Roger D. Peng. Reproducible research in computational. *Science*, 334(6060):1226–1227, December 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1213847. URL http://www.sciencemag.org/content/334/6060/1226.

Andrew P Prayle, Matthew N Hurley, and Alan R Smyth. Compliance with mandatory reporting of clinical trial results on clinicaltrials.gov: cross sectional study. *BMJ*, 344, 2012. ISSN 0959-8138. doi: 10.1136/bmj.d7373.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

Kenneth F Schulz and David A Grimes. Multiplicity in randomised trials II: subgroup and interim analyses. *The Lancet*, 365(9471):1657–1661, May 2005. ISSN 0140-6736. doi: 10.1016/S0140-6736(05)66516-6. URL http://www.sciencedirect.com/science/article/pii/S0140673605665166.

Kenneth F. Schulz, Douglas G. Altman, David Moher, and $author firstName $author.lastName. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(1):18, March 2010. ISSN 1741-7015. doi: 10.1186/1741-7015-8-18. URL http://www.biomedcentral.com/1741-7015/8/18/abstract.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, November 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797611417632. URL http://pss.sagepub.com/content/22/11/1359.

Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534, 2014.

Robert E Sorge, Loren J Martin, Kelsey A Isbester, Susana G Sotocinal, Sarah Rosen, Alexander H Tuttle, Jeffrey S Wieskopf, Erinn L Acland, Anastassia Dokova, Basil Kadoura, et al. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature methods*, 2014.

Tom D Stanley. Beyond publication bias. *Journal of Economic Surveys*, 19(3):309–345, 2005.

Darren B. Taichman, Joyce Backus, Christopher Baethge, Howard Bauchner, Peter W. de Leeuw, Jeffrey M. Drazen, John Fletcher, Frank A. Frizelle, Trish Groves, Abraham Haileamlak, Astrid James, Christine Laine, Larry Peiperl, Anja Pinborg, Peush Sahni, and Sinan Wu. Sharing clinical trial data: A proposal from the international committee of medical journal editorssharing clinical trial data. *Annals of Internal Medicine*, 2016. doi: 10.7326/M15-2928. URL +http://dx.doi.org/10.7326/M15-2928.

Sarah Taubman, Heidi Allen, Katherine Baicker, Bill Wright, and Amy Finkelstein. THE OREGON HEALTH INSURANCE EXPERIMENT: EVIDENCE FROM EMERGENCY DEPARTMENT DATA analysis plan. March 2013. URL http://www.nber.org/oregon/files/ED%20Analysis%20Plan.pdf.

Sarah L. Taubman, Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. Medicaid increases emergency-department use: Evidence from oregon's health insurance experiment. *Science*, 343(6168):263–268, 2014. doi: 10.1126/science.1246183. URL http://www.sciencemag.org/content/343/6168/263.abstract.

Erick H. Turner, Annette M. Matthews, Eftihia Linardatos, Robert A. Tell, and Robert Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3):252–260, January 2008. ISSN 0028-4793. doi: 10.1056/NEJMsa065779. URL http://www.nejm.org/doi/full/10.1056/NEJMsa065779.

Erik Von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, Jan P Vandenbroucke, Strobe Initiative, et al. The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *Preventive medicine*, 45(4):247–251, 2007.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing*. Wiley, 1993.

Bruce Wydick, Elizabeth Katz, and Brendan Janet. Do in-kind transfers damage local markets? the case of TOMS shoe donations in el salvador. *Journal of Development Effectiveness*, 6 (3):249–267, May 2014. ISSN 1943-9342. doi: 10.1080/19439342.2014.919012. URL http://dx.doi.org/10.1080/19439342.2014.919012.

Yihui Xie. *Dynamic Documents with R and knitr*. CRC Press, July 2013. ISBN 9781482203530.

Yihui Xie. knitr: A comprehensive tool for reproducible research in r. In *Implementing Reproducible Research*, pages 3–32. CRC Press, April 2014. ISBN 9781466561595.