Manual of Best Practices in Transparent Social Science Research

Garret Christensen[1] and Courtney Soderberg[2]

[1]Berkeley Initiative for Transparency in the Social Sciences

[2]Center for Open Science

June 23, 2015

Comments and suggestions are encouraged. Please send correspondence to garret@berkeley.edu, or find the latest version of the manual on github.

Contents

1 Introduction

Scientific claims should be subject to scrutiny by other researchers and the public at large. An important requirement for such scrutiny is that researchers make their claims transparent in a way that other researchers are able to use easily available resources to form a complete understanding of the methods that were used by the original. In the social sciences, especially 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.

This manual is intended to be a source mainly for empirical social science researchers who desire to make their own research transparent to, and reproducible by, others. The entire process of research, from hypothesis generation to publication, is covered. Although norms differ across disciplines, we attempt to bring a broad view of the empirical social sciences to these recommendations, and hope that students and researchers in any social science field may tailor these recommendations to best fit their field.

The manual is laid out as follows: in section 2 we first discuss the motivation for this document: the desire to do ethical research. A major component of ethical social science research is treating research subjects appropriately. This is mandated by federal law and overseen by Institutional Review Boards (IRBs), and should be taken seriously by researchers. But just as treating subjects fairly is ethical, we believe that transparent, reproducible research is also a major part of ethical research.

In section 3 we discuss study design, including how to power studies appropriately.

In section 4 we discuss one of the major problems in non-transparent research, specifically

publication bias. We 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 But without knowing how many tests were run (the number of unpublished results), it is impossible to know whether these significant results are meaningful, or whether they are the 5% of tests that we would expect to appear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 5 we 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 the social sciences, 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. We do not necessarily place higher intrinsic value on one or the other, but making the distinction clear is key for appropriate interpretation.

In section 6 we 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. We also discuss the issue of reporting standards: a standardized list

of things that authors should report to help make their work reproducible.

Section 7 concludes and presents a vision for moving forward.

2 Ethical Research

Making one's research transparent and reproducible is a key component of ethical research. Not engaging in fraud is an obvious component of this.

2.1 Fraud

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. Although the blog tends to specialize in the life sciences, and there is significantly less money involved in the social sciences than in medical and pharmaceutical research, there is no reason to believe that social science researchers are inherently more benevolent. Uri Simonsohn, Leif Nelson, and Joseph Simmons used statistical methods to detect fraud in the research of a pair of prominent social psychologists [Simonsohn, 2013].

Another source for information on fraud is part of the US Department of Health and Human Services , the Office of Research Integrity (ORI), which works to promote research integrity and document misconduct, especially when it involves federally funded research. The misconduct case summaries of the ORI, and the stories of Diederik Stapel [Carey, 2011, Bhattacharjee, 2013] Hwang Woo-Suk [Cyranoski, 2014] and Marc Hauser [Johnson, 2012] should be sobering warn-

ings to us all.

2.2 Unintentional Bias

Perhaps in addition to the obvious need to avoid deliberate fraud and protect our human subjects is the need to avoid subconsciously biasing our own results.

Nosek et al. [2012] summarize some of the evidence on this subject, concluding that there are many circumstances common to academia and the publishing paradigm that cause researchers to frequently use motivated reasoning:

> Because we have directional goals for success, we are likely to bring to bear motivated reasoning to justify research decisions in the name of accuracy, when they are actually in service of career advancement (Fanelli, 2010a). Motivated reasoning is particularly influential when the situation is complex, the available information is ambiguous, and legitimate reasons can be generated for multiple courses of action (Bersoff, 1999; Boiney, Kennedy, & Nye, 1997; Kunda, 1990).
>
> Motivated reasoning can occur without intention. We are more likely to be convinced that our hypothesis is true, accepting uncritically when it is confirmed and crutinizing heavily when it is not (Bastardi, Uhlmann, & Ross, 2011; Ditto & Lopez, 1992; Lord, Ross, & Lepper, 1979; Pyszczynski & Greenberg, 1987; Trope & Bassok, 1982). With flexible analysis options, we are more likely to find the one that produces a more publishable pattern of results to be more reasonable and defensible than others (Simmons et al., 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

> Once we obtain an unexpected result, we are likely to reconstruct our histories and perceive the outcome as something that we could have, even did, anticipate all along—converting a discovery into a confirmatory result (Fischoff, 1977; Fischoff & Beyth, 1975). And even if we resist those reasoning biases in the moment, after a few months, we might simply forget the details, whether we had hypothesized the moderator, had good justification for one set of exclusion criteria compared with another, and had really thought that the one dependent variable that showed a significant effect was the key outcome. Instead, we might remember the gist of what the study was and what we found (Reyna & Brainerd, 1995). Forgetting the details provides an opportunity for reimagining the study purpose and results to recall and understand them in their best (i.e., most publishable) light. The reader may, as we do, recall personal examples of such motivated decisions—they are entirely ordinary products of human cognition.

2.3 Institutional Review Boards

In addition to fraud, a major ethical concern relates to our human subjects.

2.3.1 History

World history is rife with examples of atrocities conducted in the name of research. Some of these have resulted in major changes in regulations related to research.

Nuremberg Nazi German doctors conducted horrible experiments on subjects during World War II. The "Doctor's Trial" (USA v. Karl Brandt, et al.) tried 23 defendants, and the verdict included

ten principles regarding voluntary consent, societal benefits from the research, minimizing risk, etc. which although never entered as formal regulations in either Germany or the USA, became widely accepted.

Tuskegee and US codification In 1972 whistleblower Peter Buxton revealed to the Associated Press that the US Public Health Service was conducting a 40-year experiment on poor Alabama sharecroppers in which it did not treat those who had syphilis for the disease despite the discovery and verification of penicillin as an effective treatment, and prevented sufferers from obtaining treatment elsewhere. As a result, the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was formed by law in 1974, and released the Belmont Report in 1979. The Belmont Report contains three basic ethical principles, and three applications:

- Ethical Principles
  - Respect for Persons: "Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection."
  - Beneficence: "Two general rules have been formulated as complementary expressions of beneficent actions in this sense: **(1)** do not harm and **(2)** maximize possible benefits and minimize possible harms."
  - Justice: "An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly. Another way of conceiving the principle of justice is that equals ought to be treated equally."

- Applications
  - Informed Consent: “Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.”
  - Assessment of Risks and Benefits: “It is commonly said that benefits and risks must be ‘balanced’ and shown to be ‘in a favorable ratio.’ The metaphorical character of these terms draws attention to the difficulty of making precise judgments. Only on rare occasions will quantitative techniques be available for the scrutiny of research protocols. However, the idea of systematic, nonarbitrary analysis of risks and benefits should be emulated insofar as possible.”
  - Selection of Subjects: “Individual justice in the selection of subjects would require that researchers exhibit fairness: thus, they should not offer potentially beneficial research only to some patients who are in their favor or select only ‘undesirable’ persons for risky research. Social justice requires that distinction be drawn between classes of subjects that ought, and ought not, to participate in any particular kind of research, based on the ability of members of that class to bear burdens and on the appropriateness of placing further burdens on already burdened persons.”

In 1981 the Department of Health and Human Services and the Food and Drug Administration adopted regulations in line with the Belmont report, and 15 federal agencies adopted these regulations (45 CFR part 46) as the “Common Rule” in 1991.

In practice, this means that researchers who receive funding from the US government, or who work at institutions that receive federal funding should have their research approved by an Institutional Review Board (IRB). IRB are a decentralized approval body set up by each research organization itself, consisting of at least five members, a mix of men and women, scientists and non-scientists, and at least one member not affiliated with the institution. Since IRBs and the approval process are decentralized, the exact process varies from institution to institution, but one example can be seen at `http://cphs.berkeley.edu`.

When conducting research internationally, researchers should give their human subjects the same protections as those inside the US. Laws in developing countries may not be as well-defined or enforced, but researchers should still register with their US institution's IRB, and obtain approval from the host country government. A list of laws and regulations that cover research in 107 foreign countries is available from the Office for Human Research Protections.

Another key resource for researchers and research conducted outside the US is the Declaration of Helsinki by the World Medical Association (WMA). Originally adopted by the WMA in 1964, the document has significantly influenced the laws and regulations adopted to govern research worldwide.

Lest one think that ethical concerns are limited to monsters of bygone eras, we refer readers to a dilemma caused by a Montana state election experiment by researchers from Stanford and Dartmouth in 2014, who sent 100,000 people official-looking election flyers bearing the state seal weeks before the election.

2.3.2 Training

A large number of universities participate in the Collaborative Institutional Training Initiative at the University of Miami (CITI). Completing their course on Human Subjects Research is often a requirement of being included on a research proposal. For anyone at an institution not affiliated with CITI, the NIH maintains an online training course that is free and open to the public.

3 Study Design

There are many issues involving study design that are somewhat related to transparent research. For randomized trials, we refer readers to two excellent resources, Duflo et al. [2007] and Glennerster and Takavarasha [2013], which cover many aspects of how to design and implement an excellent field trial. For studies that are not necessarily randomized trials, we recommend Gertler et al. [2011] or Angrist and Pischke [2008] for those with more statistics training. Here we briefly discuss one aspect of study design especially relevant for reproducible research: determining sample size through power calculations.

3.1 Power Analysis

Given that researchers are working within a null hypothesis testing framework, the power of a study, the probability of rejecting the null hypothesis when it is false, is extremely important. Though 80% power is held as a lower bound for acceptable power in many disciplines such as medical research, the actual power of studies can be much lower. For example, research by Button et al. [2013] found that the median power in neuroscience was 21%. This means that if a study

investigating a true effect were run 100 times, 79 studies would fail to reach significance, meaning that the chance of a false negative is extremely high.

Though false negatives are perhaps the most often discussed issue with low powered studies, they are not the only issue. A second issue is that low powered studies actually decrease the likelihood of a true positive [Button et al., 2013]. This means that when a low powered studies does find an effect, the lower the likelihood that this effect is true in the population; it is relatively more likely that it is a false positive. A third issue with low powered studies relates to effect sizes. Small, low powered studies that reach statistical significance will over-estimate the true effect size in the population [Button et al., 2013]. These inflated effect size estimates will make it difficult for others to properly power future studies, and the inaccuracy of the estimate may also be problematic for making decisions based on scientific results.

An obvious way to help mitigate the problems of underpowered studies is to run studies with higher power. However, this process is not always that straightforward. As previously mentioned, effect sizes from published studies are often inflated, and so may not provide the most accurate estimate. Meta-analyses can provide better information, but they also often suffer from publication bias and thus inflated effect sizes. Additionally, effect sizes may be highly heterogeneous between studies, even studies with the same materials and methodologies [Klein et al., 2014]. Thus, using a single point estimate of an effect size from published literature may lead to inaccurate power calculations.

To combat these problems, alternatives to the standard power analysis have been suggested. For example, Perugini et al. [2014] have suggested a technique called ‘safeguard power’ which takes

into account the uncertainty surrounding effect size estimates when conducting power analyses. Specifically, they suggest basing power calculations off of the effect size corresponding to the lower bound of a 60% confidence interval around the point estimate from published literature. Another approach by McShane and Böckenholt [2014] takes into account the between study variation in effect sizes when conducting power analyses to give a more conservative estimate of the number of participants needed to reach a given level of statistical power. Yet another approach, which is also feasible when an effect size estimate from the previous literature is not available, is to determine the smallest effect size you wish to be able to detect, and power your study to find this effect size [Bloom, 1995].

Additionally, researchers may be able to help one another as well as decrease false-positives by publishing or clearly making available their power calculations. If a reader of a paper saw a clearly stated "this trial was powered to detect an effect size of X," this would help to put the study into the appropriate statistical context. The parametric assumptions involved (such as the intra-cluster correlation) could also be publicly posted in study protocols to help other researchers learn reasonable assumptions for their own studies.

3.2 Practical Considerations

The impact of most of these calculations will in many cases mean a larger sample size than some disciplines are used to working with. There are several ways to mitigate the impact of this.

In psychology, a few of these possible solutions are presented in Collaboration et al. [2014]. Researchers can join crowd-sourced projects where multiple labs share protocols and produce the

same experiment, as was done with 13 psychological effects and 36 samples and settings in Klein et al. [2014]. The Collaborative Replications and Education Project (CREP) tracks findings that can be relatively easily replicated in a teaching setting.

Economists have also discusses how project budgets should play into study design, by maximizing power subject to a budget constraint Duflo et al. [2007]. When data collection is the main cost, the proportion of treatment and control should each be one half, but if treatment is expensive, power can be maximized subject to the project budget using the rule: "the ratio of subjects in the treatment group to those in the comparison should be proportional to the inverse of the square root of their cots." Another important contribution to this are of power maximization subject to a budget constraint is that power depends on the number of rounds of surveying, and also the autocorrelation of measurements over time. McKenzie [2012] shows that baseline data is of most power when autocorrelation is high, and repeated measurements post-treatment of most power when autocorrelation is low.

4 Registration

One of the problems brought into focus recently is publication bias. Publication bias is the selective publication of only significant results. Thankfully, there are tools available for researchers to combat these problems.

4.1 Publication Bias

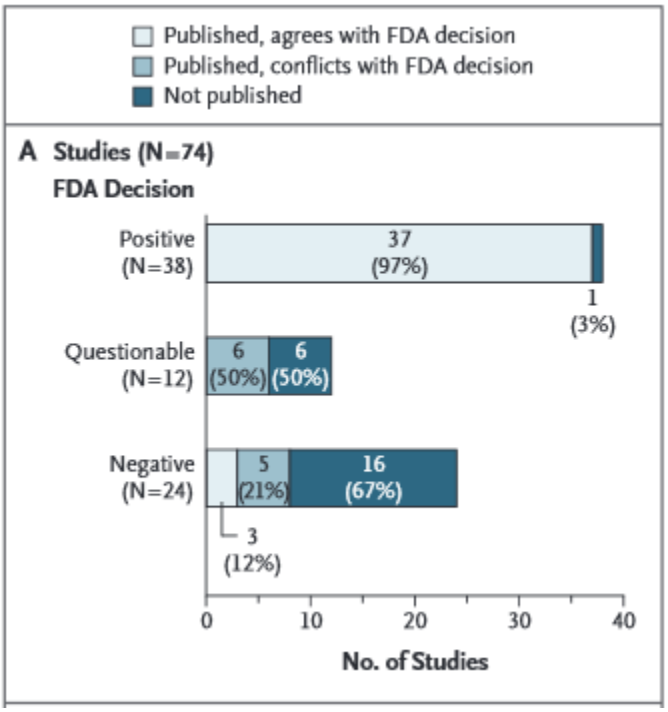
One of the primary drivers of the recent move towards transparency is increased awareness of publication bias. Numerous papers use collections of published papers to show that the proportion of significant results are extremely unlikely to come from any true population distribution [DeLong and Lang, 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 zero effect. One way to combat this problem is to require registration of all studies undertaken. Ideally we 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.

4.2 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 the social sciences. 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].

A profound 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 perhaps what the most hardened cynic would expect: essentially all the trials with positive outcomes were published, a 50/50 mix 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. The figure below shows the drastically different rates of publication, and a large amount of publication bias.



Panel A of Figure 1 from Turner et al. [2008]

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.[1] 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. Not all journals have

[1]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`

adopted the ICMJE policy, and complete enforcement is elusive. 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. We believe that registering all types of analysis should be accepted and encouraged, though there are definitely concerns to 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.

4.3 Social Science Registries

Registries in the social sciences are newer but are growing ever more popular. 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 (www.socialscienceregistry. in May 2013, which had accumulated 260 studies in 59 countries by October 2014. 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 93 designs registered as of October 2014.[2]

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 [3]

4.4 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 handful 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

[2]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).

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

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 methods 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, which includes meta-analysis datasets 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.

Prominent examples in psychology include meta-analyses of work on the Big 5 Personality Test and job performance [Barrick and Mount, 1991] and predictive validity of the Implicit Association Test [Greenwald et al., 2009]. Simonsohn et al. [2014] also developed a meta-analysis tool that

researchers can use to compare the uniform distribution of p-values under the null to the left-skewed distributions observed in research that has been gamed or selectively reported.

5 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: the manipulation or repeated searching through statistical or regression models unknowingly (or deliberately) until significance is obtained. Simmons et al. [2011] refer to this as "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, they "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." However, they also conclude that:

> the term "fishing" was unfortunate, in that it invokes an image of a researcher trying out comparison after comparison, throwing the line into the lake repeatedly until a fish is snagged. We have no reason to think that researchers regularly do that. We think the real story is that researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances.
>
> We regret the spread of the terms "fishing" and "p-hacking" (and even "researcher degrees of freedom") for two reasons: first, because when such terms are used to describe a study, there is the misleading implication that researchers were consciously trying out many different analyses on a single data set; and, second, because it can lead researchers who know they did not tryout many different analyses to mistakenly think they are not so strongly subject to problems of researcher degrees of freedom."

In other words, the problem may be even worse than you think. What can be done to solve it? We believe part of the answer lies in detailed pre-analysis plans, described below.

5.1 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 medicine for randomized 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 we discuss adapting these ideas for a pre-analysis plan in the social sciences.

A pre-analysis plan (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. Glennerster and Takavarasha [2013] suggest including the following:

1. the main outcome measures,
2. which outcome measures are primary and which are secondary,
3. the precise composition of any families that will be used for mean effects analysis,
4. the subgroups that will be analyzed,
5. the direction of expected impact if we want to use a one-sided test, and
6. the primary specification to be used for the analysis.

David McKenzie of the World Bank Research Group proposed a list of ten items that should be included in a PAP, reproduced below. (For more detail see `http://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist`)

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

In their article on researcher degrees of freedom, Simmons, Nelson, and Simonsohn (2011) suggest the following requirements for authors:

1. Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.

2. Authors must collect at least 20 observations per cell or else provide a compelling cost-of-data-collection justification.

3. Authors must list all variables collected in a study.

4. Authors must report all experimental conditions, including failed manipulations.

5. If observations are eliminated, authors must also report what the statistical results are if those observations are included.

6. If an analysis includes a covariate, authors must report the statistical results of the analysis without the covariate.

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 of a confirmatory nature as opposed to an exploratory nature. Both types of research are absolutely valuable, but knowing the distinction is important. If some sort of restriction on the data, 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, is specified ahead of time based on theory or previous research, 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 stoned. 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 we may use the phrase "bind our hands," we advocate that researchers not be punished for conducting research outside the analysis plan. We simply 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 does change 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. We find this 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.

5.1.1 Examples

Examples of pre-analysis plans in the social sciences are relatively rare, but several examples of good published papers resulting from studies with PAP exist. Several of the items below come from the J-PAL Hypothesis Registry; we highlight those that have publicly available final papers.

- 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 signficant 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.

- Teams led by Ben Olken have evaluated multiple randomized interventions in Indonesia. The Generasi program linked community block grants to performance. The PAP are Olken et al. [2009, 2010a] and are available in the J-PAL Hypothesis Registry. The researchers found health improvement, but no education improvement [Olken et al., 2010b, 2014].

- Another project in Indonesia used a field experiment to evaluate different means of poverty targeting for cash transfer programs: proxy-means testing, community-based targeting, and a hybrid of the two. Results show that the proxy-means testing outperformed the other methods by 10%, but that community members were far more satisfied with the community method. The PAP and final paper are available as Olken [2009] and Alatas et al. [2012]

- An example from pyschology is a pre-registered replication of an implicit association test. Existing research showed evidence of stronger racial preferences among fertile women. Hawkins et al. [2015] failed to reproduce this effect in four tries, suggesting the association is weaker than originally found. The resulting manuscript, as well as the time-stamped registration of the analysis plan, can be found on the Open Science Framework at `https://osf.io/g3sca/`.

Additionally, Alejandro Ganimian developed a template for pre-analysis plans that instructors may find useful when teaching tranpsarency methods, or researcher themselves may find useful when developing their own pre-analysis plan.

5.1.2 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 sciences 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 we assume the same situation exists in the social sciences. *Nature* has decided to expand its methods section in order to encourage better reporting.[4]

We believe the social sciences would 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.[5] This is discussed below in section 6. 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.

One area in which social sciences could improve involves details of randomization. Bruhn

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

[5] 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.

and McKenzie [2009] documents the lack of clear explanation pertaining to how randomization was conducted in RCTs published in economics journals. Variables used for stratification are not described, and the decision of whether to control for baseline characteristics was often done after the fact. While there seems to be internal disagreement in both medicine and the social sciences over the appropriateness of including baseline control variables in regression analysis of a randomized trial, having researchers selectively report whatever method gives them the most significant-seeming results is obviously not the optimal outcome.

The medical literature also exhibits much greater concern over the blinding and concealment of randomized assignment than some of the social science literature. In some situations, blinding is impossible or irrelevant in a social science field experiment: for example, the recipient of a cash transfer needs to know that they received cash in order for the program to have any effect. Also, a social scientist interested in the potential scaling up of a government program may rightfully be unperturbed by some respondents assigned to the control group somehow gaining access to treatment, since this behavior would undoubtedly occur if the program were scaled up and the researcher still has a valid intention to treat estimate. This is clearly not always the case, especially if one wants an accurate estimate of the efficacy of a program or a treatment on the treated estimate. Tales of trials ruined through carelessness with the original randomization assignment as well as tips on how to avoid the same problem are described in Schulz and Grimes [2002]. In addition to Bruhn and McKenzie [2009], political science has produced guidelines for randomization and related disclosure, avaialable at `http://e-gap.org/resources/guides/randomization/`.

Some medicine and science journals have begun to publish protocols. While the advantages

of publishing a protocol related to the development of a new procedure (e.g. "we have developed a new method of isolating mRNA") should be obvious, the advantages of publishing protocols for randomized trials under way are perhaps less obvious, but still exist. *BioMed Central* and *BMJ Open*, among others, now publish protocols of trials planned or ongoing, with the hopes that this will reduce publication bias, allow patients to see trials in which they might like to enroll, allow funders and researchers to learn of work underway to avoid duplication, and to allow readers to compare what research was originally proposed to what was actually completed. (See `http://www.biomedcentral.com/authors/protocols` and `http://bmjopen.bmj.com/site/about/guidelines.xhtml#studyprotocols`.) *BMJ Open* suggests, but does not require, that its published protocols include the items in the SPIRIT checklist.

Protocols are not a perfects solution, as even in published (or otherwise public) protocols, studies have found important differences between protocols and published results. 60-71% of outcomes described in protocols went unreported in the paper while 62% had major discrepancies between primary outcomes in the protocols and in the published papers, though there was a relatively even mix of these discrepancies favoring significant or insignificant results [Chan A et al., 2004]. Another study found that appropriate level of statistical detail is often lacking in protocols, and there are often discrepancies between protocols and published results [Saquib et al., 2013]. 31% of published papers had some sort of pre-specified plan for their regression adjustments (i.e. specifying which baseline covariates would be controlled for), but only 53% of the plans matched what was published in the ultimate paper.

5.2 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 indeces (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 dicsount 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.

5.3 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 ("interactions" 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, we feel that 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.

5.4 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 signficant.

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. See a full list of journals adopting the procedure at `https://osf.io/8mpji/`. *Social Pyschology* ran an issue dedicated to this type of article, with an editorial explaining the concept [Nosek and Lakens, 2014]. *AIMS Neuroscience*, which uses this format, published an editorial answering 25 frequently asked questions pertaining to registered reports [Chambers et al., 2014].

6 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

> "Reproducibility is just collaboration with people you don't know, including yourself next week"—Philip Stark, University of California Berkeley Statistics

Replication, in both practice and principle, is a key part of social science research. We 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 and code are re-used by other researchers [Peng, 2011]. 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.

6.1 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.

6.1.1 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.[6] Version control is simply archiving previous versions of files so that old versions are not lost and can be returned to if needed. Instead of simply calling one's analysis code "MyAnalysis.do" and repeatedly saving

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

over and losing old versions, and instead of repeatedly changing the file name from "MyAnalysis.2014.8.13.do" to "MyAnalysis.2014.8.14.do" according to the date, version control creates different versions of files and can compare and highlight the differences in version of text files, and restore the used file to previous conditions if desired. Web services such as GitHub have the advantage of being "distributed" (Distributed Version Control System, DVCS) in that several users can have access simultaneously.

6.1.2 Managing Workflow

Code is just one aspect of a larger structure we refer to as "workflow" after Long [2008], by which we mean 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. We strongly recommend that Stata-users read Long [2008] and R users read Gandrud [2013] for workflow recommendations both general and specific to their respective programming language[7]. Our suggestions here borrow heavily from their excellent work. We also refer undergraduate instructors and others who may be interested 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 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).

[7] Also see and Kirchkamp

Software: Although we agree with the movement by many towards open source software such as R and Python, we appreciate that 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 we 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 programming 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 we understand that a fair amount of research has been done using pull down menus in SPSS or Stata, we advise against this. 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 from websites can be done through a script. 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.) Another extremely 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 (set.seed() in R, set seed () in Stata). 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 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.

6.2 General Workflow Suggestions:

Here we offer some specific workflow organization suggestions that should be valid regardless of code or operating system.

- Do not use spaces in directory or file names, as it complicates referring to them in certain software.

- Use "naming directories", .i.e. a directory beginning with "-" (so that it will appear first alphabetically) inside each directory to explain the contents of the above directory.

- Add name, date, and describe contents, as well as updates, to all scripting files.

- Keep a daily research log, i.e. a detailed written diary of what research is done on a given day. You'll be surprised how often this will be useful to answer questions about whether you ran a certain test or not, when you did it, and what you called the file.

- Make sure that all .do files are self-contained, do not require data in memory, or ideally, certain directory.

- You can never comment too much.

- Indent your code

- Once you post/distribute code or data, any changes at all require a new file name.

- Separate your cleaning and analysis files; don't make any new variables that need saving (or will be used by a different analysis file) in an analysis file— it is better to only create them once so you know they're the same.

- Never name a file "final" because it won't be.

- Name variables "male" instead of "gender."

- Use a prefix such as x_ or temp_ so you know which files can easily be deleted.

- Never change the contents of a variable unless you give it a new name.

- Every variable should have a label.

6.3 Stata-specific Suggestions

- Use the different missing values (“.a”-“.z”, not exclusively “.”) in order to distinguish between “don’t know” and “didn’t ask” or other distinct reasons for missing data.
- Make sure code always produces same result—set seed and sort/merge stable
- Use the ‘version’ command in your .do file to ensure that other researchers who run your code with a newer version of Stata get the same results.
- Don’t use abbreviations for variables (which may become unstable after adding variables) or commands (beyond reason)
- Avoid using global macros. (This is a common piece of programming advice.)
- Use locals for varlists to ensure that long lists of variables include the same variables whenever intended.
- Use ‘return’ command instead of typing in numbers
- If you have a master .do file that calls other .do files, which each have their own .log file, you can run multiple log files at the same time (so you have a master .log file)
- Use the ‘label data’ and ‘notes’ commands to label datasets and help yourself and other researchers easily identify the contents.
- Use the ‘notes’ command for variables as well for identifying information that is too long for the variable label.

- Use the 'datasignature' command to generate a hash and help ensure that data is the same as before.
- Use value labels for all categorical variables, but include the numerical value in the label.
- Even though Stata is case sensitive, don't use capital letters in variable names since not all software packages are case sensitive.
- Make your files as non-proprietary as possible (use the 'saveold' command to enable those with earlier versions to use your data. This is why trusted repositories are so useful–they'll do this for you.)

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 tool to make literate (statistical) programming significantly easier is Knitr (see Xie [2013, 2014]) which is built into R Studio[8]. Knitr uses R Markdown (a very simple plain text markup language, described at `http://rmarkdown.rstudio.com/`) in which one writes both code and

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

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 (`https://rpubs.com`), 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, Weaver, Ketchup, and Synlight) 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.

6.4 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.

6.4.1 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 they have long had an explicit policy of expecting researchers to share their primary data[9]. 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

[9] "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

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.

Shockingly, 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 with papers presented at the St. Louis Federal Reserve Bank conference in 1992, and obtained similar response rates as original JMCB Project. The flagship economics journal, the *American Economic Review* (AER), did not start requesting data until 2003. Finally, after a 2003 article showing 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 2005 [McCullough and Vinod, 2003,

McCullough, 2007].

The current data policy from the *American Economic Review* can be seen here: `https://www.aeaweb.org/aer/data.php`. In addition to all the journals published by the American Economic Association, several top journals, including *Econometrica, The Journal of Applied Econometrics, The Journal of Money Credit and Banking, the Journal of Political Economy, The Review of Economics and Statistics, and the Review of Economic Studies*, now explicitly require data and code to be submitted at the time of publication. The AER conducted a review and found good, but incomplete, compliance [Glandon, 2010].

6.4.2 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 things. 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 (`http://www.re3data.org`) 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 here, in lieu of simply throwing the data up 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.

6.4.3 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 not yet agreed upon, nor widely implemented.

6.5 Reporting Standards

In research, the devil truly is in the details. Whether it is for assessing the validity of a research design or for 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 infinite and easy web storage, researchers can easily post these materials on their website even if journal editors insist on cutting their methods sections for space reasons.

6.5.1 Randomized Trials and CONSORT

The most widely adopted reporting standard guideline is the Consolidated Standards of Reporting Trials (CONSORT), available at `http://www.consort-statement.org`. Parallel to construction of clinicaltrials.gov and registration, reporting standards evolved, 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, as required by the checklist, had significantly increased in journals requiring the standard [Moher D et al., 2001]. The statement was revised in 2001, and simultaneously published in three of the top journals [Moher et al., 2001]). The statement was again revised in 2010 [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.

6.6 Social Science Reporting Standards

Though a standard akin to CONSORT has not been formally adopted by social science or behavioral science 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 in political science [Gerber et al., 2014]. This checklist is

available here.[10]

In economics, one article has highlighted the fact that there is not much 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) was developed in [Montgomery et al., 2013], but has so far not been widely adopted, or required by journals.

6.7 Observational Reporting Standards

Social science has yet to make a serious push for reporting standards in 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), available at `http://www.strobe-statement.org`. 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.[11]

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

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

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

relevant for their research. See http://www.equator-network.org/ for more information.

7 Conclusion

As you may have noticed, many of the activities described in this manual require extra work. Before you run an experiment, we're telling you to write down the hypothesis, carefully explain how you are going to test the hypothesis, write down the very regression analysis you're going to run, write a detailed protocol of the exact experimental setting, and then you have to post all of this publicly on the Internet with some sort of Big Brother organization. Or at least that's one way to look at it. But we strongly belive these steps are (1) not that difficult once you get used to them and (2) well worth the reward. You'll get p-values you can believe in. The next time someone asks you for your data, you just point them to the website, where they'll download the data and code, and the code will produce the exact results in the published paper. The next time you open up a coding file you haven't looked at in months to make a change suggested by a reviewer, your code will be so thoroughly commented, you'll know exactly where to go to make the changes. And the next time you want to extend the analysis of a published paper, you click the link in the paper and have the data on your own computer in seconds. Science moves forward.

8 Glossary of Terms

We feel that it is important to define the terms we use in this document. Although many of the concepts overlap, we suggest that researchers use the following terms:

- Analysis Plan: See pre-analysis plan

- Data citation: The practice of specifically citing datasets, and not just the paper in which a dataset was used. This helps other researchers to find data and rewards researchers who share data, leading to better science. Read more at `http://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp`.

- Data mining: Searching blindly and repeatedly through data to find supposedly statistically significant relationships. While not inherently wrong, if done without a plan or without adjusting for multiple hypothesis testing, test statistics and p-values that result no longer hold their traditional meaning, and can lead to research that cannot be replicated.

- Data sharing: Researchers making the data they use in an analysis available to other researchers, ideally through a trusted public archive.

- Design-based publication: See results-blind review.

- Disclosure: In addition to the widely accepted norm of publicly declaring all potential conflicts of interest, researchers should detail all the ways in which they test a hypothesis (e.g. include the outcome from all regression specifications tested in the appendix.)

- Fishing: See data mining.

- Literate Programming: The idea of writing programming code designed to be read and easily understood by a human. Use of this best practice can make a researchers code far more easily reproducible by others.

- Multiple hypothesis correction: Statistically taking into account the fact that multiple hypotheses have been tested. This tends to decrease the reported statistical significance of any individual test conducted. The oldest method, known as the Bonferroni correction simply divides the significance threshold by the number of tests. This is quite conservative, and more modern methods are helpfully described in (Anderson 2008).

- Open Acces: Journals, or articles that are freely available to the public, as opposed to available only to subscribers.

- Open Data: See data sharing.

- P-hacking: See data mining.

- P-value: The statistic researchers use to make judgments regarding statistical significance, which is quite often misunderstood. It is the probability of obtaining a test statistic at least as extreme as the observed test statistic when the null hypothesis is true.

- Pre-analysis plan: A document that details, ahead of time, the statistical analysis plan that will be conducted for a given research project. Outcomes, control variables, and regression specifications are all written in as much detail as possible. This serves to make research more confirmatory in nature.

- Pre-specification: Detailing the method of analysis before actually beginning data work; the same as writing a pre-analysis plan.

- Protocol: A general term meaning a document that provides a detailed description of a research project, ideally written before the project takes place, and in enough detail that other

researchers may reproduce the project on their own. Often used in the context of human subjects IRB protocols, but increasingly used in connection with pre-analysis plans.

- Publication Bias: The unfortunate fact that research is often only published when it contains the rejection of a null hypothesis test, i.e. a statistically significant relationship. Reviewers or journal editors may consider a null finding to be of less interest, or a researcher may fail to write up a null result, even though the null result may be the truth.

- Registration: Publicly declaring that an investigation of a hypothesis is or will be undertaken.

  - Study Registration: Registering a research project that is not a randomized trial.

  - Trial Registration: Registering a randomized trial.

- Registry: A database of registered studies or trials. For instance, socialscienceregistry.org or clinicaltrials.gov. Some of the largest registries only accept randomized trials, hence the frequent discussion of trial registries.

- Registered Reports: See results-blind review.

- Replicable: See reproducible.

- Replication: The idea of conducting an existing research project over again, with the hope of obtaining the same result. A subtle taxonomy exists, explained in (Hamermesh 2007).

  - Pure Replication: Re-running existing code, with error-checking, on the original dataset and seeing if the published results are obtained.

  - Scientific Replication: Attempting to reproduce the published results with a new sample, with the same code or with slight variations from the original analysis.

- Reproducible: The test of whether research can redone by another researcher and produce the same results as the original.

- Researcher Degrees of Freedom: the flexibility that a researcher has in data analysis, whether consciously abused or not. See data mining.

- Results-blind Review: To help reduce publication bias, peer review can take place before results of a study are determined. Reviewers look evaluate the design of a study to see whether the question is important and whether the study is well-designed. Good studies are given in-principle acceptance, and cannot be discriminated against for a null result, which, after all, may be the truth. See `https://osf.io/8mpji/wiki/home/` for journals practicing this form of publication.

- Specification searching: Testing numerous regression specifications and reporting only the model that produces the desired results. See data mining.

- Trusted Digital Repository: A location for storing data that others can believe will not be manipulated, and will be available into the future. Storing data here is superior to simply posting on an individual website, since it is more easily accessible and less easily changed.

- Version Control: A method of tracking every edit made to a computer file. This is often quite useful for empirical researchers who may edit their programming code hundreds or thousands of times.

Bibliography

Vivi Alatas, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. Targeting the poor: Evidence from a field experiment in indonesia. *American Economic Review*, 102(4):1206–1240, June 2012. ISSN 0002-8282. doi: 10.1257/aer.102.4.1206. URL https://www.aeaweb.org/articles.php?doi=10.1257/aer.102.4.1206&fnd=s.

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.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

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.

Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

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.

Yudhijit Bhattacharjee. Diederik stapel audacious academic fraud. *The New York Times*, April 2013. ISSN 0362-4331. URL http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.

Howard S Bloom. Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation review*, 19(5):547–556, 1995.

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.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

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

Benedict Carey. Noted dutch psychologist, stapel, accused of research fraud. *The New York Times*, November 2011. ISSN 0362-4331. URL http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html.

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.

Chan A, Hrbjartsson A, Haahr MT, Gtzsche PC, and Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291(20):2457–2465, May 2004. ISSN 0098-7484. doi: 10.1001/jama.291.20.2457. URL http://dx.doi.org/10.1001/jama.291.20.2457.

Open Science Collaboration et al. *Maximizing the Reproducibility of Your Research*. Wiley, New York, NY, 2014. URL https://osf.io/nte3j/.

David Cyranoski. Cloning comeback. *Nature*, 505(7484):468–471, January 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/505468a. URL http://www.nature.com/news/cloning-comeback-1.14504.

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.

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.

Esther Duflo, Rachel Glennerster, and Michael Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

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.

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 expeditionor 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.

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.

Paul Gertler, Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. *Impact Evaluation in Practice*. World Bank Publications, 2011. ISBN 9780821385418.

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.

Anthony G Greenwald, T Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R Banaji. Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of personality and social psychology*, 97(1):17, 2009.

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.

Carlee Beth Hawkins, Cailey E. Fitzgerald, and Brian A. Nosek. In search of an association between conception risk and prejudice. *Psychological Science*, 26(2):249–252, February 2015. doi: 10.1177/0956797614553121. URL http://pss.sagepub.com/content/26/2/249.

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.peh-med.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.

Carolyn Y. Johnson. Harvard professor who resigned fabricated, manipulated data, US says - the boston globe. *BostonGlobe.com*, September 2012. URL https://www.bostonglobe.com/news/science/2012/09/05/harvard-professor-who-resigned-fabricated-manipulated-data-says/6gDVkzPNxv1ZDkh4wVnKhO/story.html.

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.

Oliver Kirchkamp. *Workflow of Statistical Data Analysis*. URL http://www.kirchkamp.de/oekonometrie/pdf/wf-screen2.pdf.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. Investigating variation in replicability. *Social Psychology*, 45(3):142–152, 2014.

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. Hbert, 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.

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

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. Got replicability? the journal of money, credit and banking archive. *Econ Journal Watch*, 4(3):326–337, September 2007. URL http://econjwatch.org/articles/got-replicability-the-journal-of-money-credit-and-banking-archive?ref=section-archive.

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.

David McKenzie. Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics*, 99(2):210–221, 2012.

Blakeley B McShane and Ulf Böckenholt. You cannot step into the same river twice when power analyses are optimistic. *Perspectives on Psychological Science*, 9(6):612–625, 2014.

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 the employment effects of MinimumWages. *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.

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. Targeting analysis protocol. Technical report, March 2009. URL https://www.povertyactionlab.org/sites/default/files/documents/090318a%20Targeting%20Analysis%20Protocol.pdf.

Benjamin A Olken. Pre-analysisplans in economics. January 2015. URL http://economics.mit.edu/files/10399.

Benjamin A. Olken, Junko Onishi, and Susan Wong. Generasi analysis plan. Technical report, April 2009. URL http://www.povertyactionlab.org/sites/default/files/documents/090408_Generasi_Analysis_Plan_CLEAN.pdf.

Benjamin A. Olken, Junko Onishi, and Susan Wong. Generasi analysis plan: Wave III. Technical report, January 2010a. URL http://www.povertyactionlab.org/sites/default/files/documents/100122_Generasi_AnalysisPlan_Wave_III_CLEAN.pdf.

Benjamin A. Olken, Junko Onishi, and Susan Wong. Indonesia's PNPM generasi program : interim impact evaluation report. Technical Report 59567, The World Bank, January 2010b. URL http://documents.worldbank.org/curated/en/2010/01/13763479/indonesias-pnpm-generasi-program-interim-impact-evaluation-report.

Benjamin A. Olken, Junko Onishi, and Susan Wong. Should aid reward performance? evidence from a field experiment on health and education in indonesia. *American Economic Journal: Applied Economics*, 6(4):1–34, 2014. doi: 10.1257/app.6.4.1. URL http://www.aeaweb.org/articles.php?doi=10.1257/app.6.4.1.

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.

Marco Perugini, Marcello Gallucci, and Giulio Costantini. Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9(3):319–332, 2014.

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.

N. Saquib, J. Saquib, and J. P. A. Ioannidis. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*, 347(jul12 2):f4313–f4313, July 2013. ISSN 1756-1833. doi: 10.1136/bmj.f4313. URL http://www.bmj.com/content/347/bmj.f4313.

Kenneth F Schulz and David A Grimes. Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, 359(9306):614–618, February 2002. ISSN 01406736. doi: 10.1016/S0140-6736(02)07750-4. URL http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(02)07750-4/fulltext.

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. Just post it: the lesson from two cases of fabricated data detected by statistics alone. *Psychological science*, 24(10):1875, 2013.

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.

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.

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.