Module 3.4: Transparency and Reproducibility[1]

Contents

[1] This module was written by Garret Christensen of the Berkeley Initiative for Transparency in the Social Sciences (BITSS). This module will be appearing in the second edition of the text, *Impact Evaluation in Practice*, published by the World Bank Press. Christensen's full "Manual of Best Practices in Transparent Social Science" is available online (https://github.com/garretchristensen/BestPracticesManual).

1. INTRODUCTION

In the previous modules of this section, we have taught you how to set up and design the survey, how to hire a team that will do a good job, what steps to put in place to keep the data collection running smoothly, and how to prevent/fix data quality concerns.

We will conclude this course with a module on transparency and reproducibility, which are two pivotal pillars in the scientific community. By definition science is supposed to be objective, which means anyone should be able to reproduce the results you find. Akin to the "missing data" problem of causal inference, we can imagine that running the same program on the same sample (but in different parallel universes) *should* get us identical findings. So in this module we will discuss the methods used to avoid introducing unnecessary bias or possible even fraud, albeit intentional or unintentional, into your impact evaluation.

As with all scientific claims, the claims you make in impact evaluations should be subject to scrutiny by other researchers and the public at large. An essential requirement for such scrutiny is that researchers make transparent claims such that other researchers will be able to use available resources to form a complete understanding of the methods that were used in the original research. In impact evaluation, 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. To make research transparent, you can register your impact evaluation prospectively with a study registry and include a pre-analysis plan. Once research is completed, you can share your data and code publicly so that others may replicate your work.

2. FRAUD AND UNINTENTIONAL BIAS

While most of us are likely to presume that we ourselves would not conduct outright fraud in our impact evaluation, fraud does indeed occur. From making up fake data to creating bogus e-mail addresses so that one can do one's own peer review, there is a distressingly large amount of deliberate fraud in research. An office in the US Department of Health and Human Services, the Office of Research Integrity (ORI), 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 warnings to all researchers.

But even more important than the obvious need to avoid blatant fraud is the need to avoid subconsciously biasing our own results. Nosek et al. [2012] concluded that there are many common circumstances in academic research in which researchers tend to use motivated reasoning, which can bias results. Many of these forces are the same for practitioners and impact evaluators. The rest of this chapter describes methods that can help you overcome these powerful subconscious factors and help you produce unbiased research.

3. PUBLICATION BIAS

One common problem in research is publication bias: the selective publication of papers with statistically significant results. Numerous studies use collections of published papers to show that the proportion of significant results is extremely unlikely to come from any true population distribution [DeLong and Lang, 1992, Gerber et al., 2001, Ioannidis, 2005]. By examining these skewed publication rates from a large set of National Science Foundation (NSF)-funded studies, Franco et al. [2014] show that the selective publication of significant results stems from researchers largely failing to write up and submit results from studies with null findings, usually citing lack of interest or fear of journal rejection.

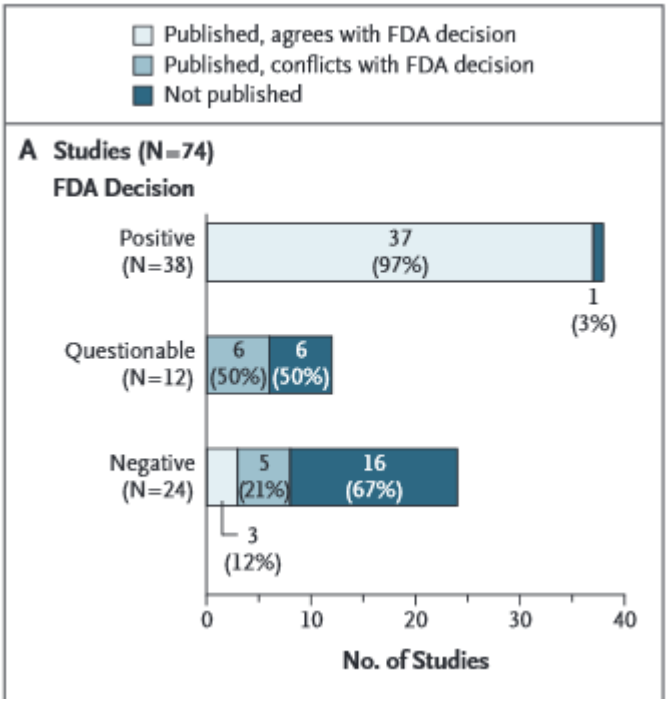
Reviewers rejecting papers with null-results, or authors never even submitting such papers for review, is commonly referred to as the "file drawer problem." In fact, the percentage of null findings published in journals appears to have been decreasing over time across all disciplines [Fanelli, 2012]. This is clearly unlikely to reflect the increasing success of research topic selection. If journals only publish statistically significant results, we have idea poor idea of what fraction of significant results are evidence of real effects, since it will be higher than the five percent expected given the standard power assumption. One way to combat this problem is to require registration of all undertaken studies. Ideally, we could then search the registry for studies of X given a change in Y. If several studies substantiate an effect, we would have confidence the effect is real. If only five percent of studies show a significant effect, we give these outlier studies less credence.

4. TRIAL REGISTRATION

A solution to the problem of publication bias is registration, which involves publicly declaring 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 largely unused 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].

Box 1 Non-Publication of Non-Significant Results

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 strong 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, conducting such an exercise requires that a registration organization itself obtain outcomes of trials. ClinicalTrials.gov is the only publicly available trial registry that requires such reporting of results, and is restricted to certain FDA trials.[2] Almost all registration efforts have been limited to randomized control trials as opposed to observational data. There may be value in registering all types of analysis, but there are concerns about registering retrospective projects—not least of which is the inability to verify that registration preceded analysis. See Dal-Re et al. [2014] for a discussion of the pros and cons of observational study registration.

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

4.1 Social Science Registries

Registries in the social sciences are newer but are growing more popular. The American Economic Association launched a registry for randomized trials (www.socialscienceregistry.org) in May 2013, which had accumulated 483 studies in 81 countries by October 2015.
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 68 evaluations registered in its first two years.

In political science, EGAP (Evidence 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.

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, and hypotheses, to the OSF, register them, and then share the resulting URL as proof of registration. OSF registrations can be relatively free-form, but templates exist to conform to standards in different disciplines. Psychology registrations are presently the most numerous on the OSF.

4.2 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. Although quite common in medical research, meta-analysis is not widely used in some parts of the social sciences. A meta-analysis of 87 meta-analyses in economics shows that publication bias is widespread, but not universal.[3]

[3] 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.

5. DATA MINING

Another problem with impact evaluation is improper data mining: blindly running regression after regression until statistical significance is obtained. Though registration helps solve the problem of publication bias, it does not solve the problem of fishing for statistical significance within a given study. Simmons et al. [2011] refer to this as "researcher degrees of freedom."

Using flexibility around the completion of the data-collection process, excluding certain observations, combining and comparing certain conditions, including or excluding 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" or "p-hacking" is also described in Humphreys et al. [2013]. Gelman and Loken [2013] agree that "[a] dataset can be analyzed in so many different ways (with the choices being not just what statistical test to perform but also decisions on what data to [include] or exclude, what measures to study, what interactions to consider, etc.), that very little information is provided by the statement that a study came up with a $p<.05$ result." 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 perform 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, they misleadingly imply that researchers consciously try many different analyses on a single data set; and, second, because it can lead researchers who know they did not try many different analyses to mistakenly believe that they are not so strongly subject to problems of researcher degrees of freedom.

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

5.1 Pre-Analysis Plans

In addition to using study registration to reduce publication bias, a pre-analysis plan (PAP), a detailed outline of the analyses that will be conducted in a study, can be used to reduce data mining. A pre-analysis plan (PAP) contains a specification of the outcomes of the study, as well as a specification of the methods that will be used to analyze the outcomes (sometimes referred to as endpoints in the medical literature). 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.

Leading impact evaluation researchers have made suggestions for the detailed contents of these documents. Glennerster and Takavarasha [2013] suggest including the following:

- ✓ The main outcome measures,
- ✓ Which outcome measures are primary and which are secondary
- ✓ The precise composition of any families that will be used for mean effects analysis
- ✓ The subgroups that will be analyzed
- ✓ The direction of expected impact if we want to use a one-sided test
- ✓ 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 the World Bank Development Impact blog.)

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

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 valuable, but understanding 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. Such analysis may provide 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, many legitimately unexpected events can occur. (An example discussed at a recent conference was subjects showing up for an experiment high on marijuana. Another example is from a field experiment involved fatalities from a lightning strike at a school [Kremer et al., 2009].) This is why, even though we may use the phrase "bind our hands," we believe that researchers should not be punished for conducting research outside their analysis plan.

We simply recommend that researchers clearly delineate which analysis was included in the analysis plan, so that readers understand what is confirmatory and what is exploratory.[4]

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] offers an informative discussion of the relative values of PAP timing. PAPs written before the baseline may be maximally pure, 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, the researcher could have accounted for this, but at the same time, she would also be free to change the scope of their analysis–for example, if 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.

[4] An additional method to explore for pre-analysis plans is the data-adaptive algorithms used in some recent medical literature. See Van der Laan and Petersen [2011].

Box 2 Examples of Analysis Plans in Impact Evaluation

Several examples of good impact evaluation papers resulting from studies with pre-analysis plans have been written in the last few years. 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 significant results would have been relatively easy. In fact, the paper includes an example of how, if they had the latitude to define outcomes without a pre-analysis plan, the authors could have reported large and significantly positive outcomes. 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], 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 Research 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.
- ✓ 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 registry as well as on The World Bank Development Impact Blog.
- ✓ Teams led by Ben Olken from MIT 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 methods 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]

5.2 Project Protocols

Project protocols are similar, but importantly distinct, documents to PAPs. 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 medical research and most 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 or the gender of the laboratory assistant or the speed at which one stirs a reagent [Sorge et al., 2014, Hines et al., 2014], and the same is likely true in impact evaluation.

Impact evaluations could 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. With original data collection, researchers should provide very detailed descriptions of their precise procedures. A 33-item checklist of suggested items is contained in the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) statement (see http://www.spirit-statement.org, 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 impact evaluation could improve involves randomization details. Bruhn and McKenzie [2009] document the lack of a clear explanation pertaining to how randomization was conducted in several RCTs published in economics journals. Variables used for stratification are not described, and the decision of whether to control for baseline characteristics was often made after the completion of data collection.

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, a social scientist interested in the potential scaling up of a government program may rightfully be unperturbed by a respondent assigned to the control group somehow gaining access to treatment, since this behavior would undoubtedly occur if the program were scaled and the researcher still has a valid intention to treat estimate. However, treatment of control group members is highly problematic in general, especially if one wants to accurately estimate the efficacy of a program or the magnitude of the treatment on the treated effect. 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, available at http://e-gap.org/resources/guides/randomization/

5.3 Multiple Hypothesis Testing

Several of the PAP described above, and both of the lists of suggestions above, include corrections for multiple hypothesis testing. Simply put, because we are aware of the fact that test statistics and p-values appear statistically significant 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, all of which are used and explained in a simple and straightforward manner by Anderson [2008]:

- ✓ Report index tests—instead of reporting the outcomes of numerous tests, standardize outcomes and combine them into a smaller number of indexes (e.g. instead of separately reporting whether a long-term health intervention reduced blood pressure, diabetes, obesity, cancer, heart disease, and Alzheimer's, report the results of a single health index.)
- ✓ 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 that 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 too conservative) method being the Bonferroni correction of simply multiplying every original p-value by the number of tests done. Holm's sequential method involves ordering p-values by class and multiplying the lower p-values by higher discount factors [Holm, 1979]. An efficient recent method is the free step-down resampling method developed by Westfall and Young [1993].
- ✓ Control the False Discovery Rate (FDR)—In situations where a single type I error is not catastrophic, researchers may be willing to use a less conservative method and trade off some incorrect rejections in exchange for greater power. This is possible by controlling the FDR, or the percentage of rejections that are type I errors. Benjamini and Hochberg [1995] details a simple algorithm to control this rate at a chosen level, and Benjamini et al. [2006] described a two-step procedure with greater power.

5.4 Subgroup Analysis

One aspect of data mining related to multiple hypothesis testing that is widely avoided in the medical literature is sub-group analysis ("interactions" in economic parlance). 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 journal editors reviewing the article 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 [1988]). Whether in a randomized trial or not, we feel that impact evaluation 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.

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 good research and impact evaluation. Numerous different definitions of the types of replication have been developed (see, for instance Hamermesh [2007]), ranging from trying to run the same code on the same data to try and reproduce the results of a published paper, to testing the model on a new set of data, or attempting to test the robustness of a previous result by testing different regression specifications. Whatever the terminology used, transparent research requires making data and code available to other researchers so that they can try and get the same results.

6.1 Workflow and Code

Reproducing research often involves using the exact code and statistical programming written by the original researcher. To make this possible, code needs to be both (1) easily available and (2) easily interpretable. Thanks to the 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 more complicated task. Nevertheless, the extra effort spent to make a more manageable code pays off with significant dividends.

Perhaps the most important rule is to actually write code in the first place, instead of working by hand. By that we mean:

- ✓ Do not modify data by hand, such as using a spreadsheet program. Which is to say, try not to use Microsoft Excel to clean your data.
- ✓ Use neither the command line nor drop-down menus nor point-and-click options in statistical software.
- ✓ Instead, conduct all analysis using scripts (do files, in Stata).

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 perform one-time-only data cleaning in Excel, or to make "minor" changes to get the data into a format readable by a researcher's preferred statistical software, these changes are not reproducible by other researchers unless they are written down in excruciating detail. It is better to write a code script that imports the raw data, performs all necessary changes (with comments in the code that explain the changes), and saves any intermediate data used in analysis. Then, researchers can share their initial raw data and their code, and other researchers can reproduce their work exactly.

Once analysis is complete (or even before this stage) researchers should share their data and code with the public. See the list of tools below for free resources to facilitate sharing.

In addition to making code available to the public, all code should be written in a reader-friendly format, referred to as "Literate Programming," introduced in Knuth [1984] and Knuth [1992]. According to Knuth, "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." 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.

Box 3 Tools for Transparent Data Analysis

- ✓ Organizing your work so that you can come back to it later, and so that other researchers can understand and replicate your work is difficult. We 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.
- ✓ GitHub(http://www.github.com) and The Center for Open Science's Open Science Framework (http://osf.io) are both free repositories for data and code that include easy-to-use version control. Version control is the archiving previous versions of files so that old versions are not lost and can be returned to if needed. Web services such as GitHub have the advantage of being "distributed" (DVCS), in that several users can have access simultaneously.
- ✓ For data sharing, Harvard University's Dataverse (http://thedata.org) is an excellent free repository, and The Registry of Research Data Repositories has described over 900 data repositories to help you find the right data repository for your data.

6.2 Sharing Data

In addition to code, researchers should share their data for the sake of reproducibility. Many journals do not require sharing of data, but the number that do is increasing. Even if the journal or organization 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 are equipped to handle data from practically any impact evaluation, many repositories specialize. A key advantage to using a trusted repository in lieu of your own resources is that many of these repositories will take your data in its proprietary (Stata, SAS, SPSS, etc.) format 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.

Box 4 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[5]. Despite this, and despite the explicit policy of the Journal during the project, at most only 78% of authors provided data to the authors within six months after multiple requests. (This is admittedly an improvement over the 34% from the control group—those who published before the Journal policy went into effect—who provided data.) Of the papers that were still under review by the Journal at the time of the requests for data, one quarter did not even respond to the request, despite the request coming from the same journal considering their paper! The submitted data was often unlabeled and poorly formatted. Despite these difficulties, the authors attempted to replicate nine papers, and often were completely unable to reproduce published results, despite detailed assistance from the original authors. Sadly, the publication of this important article changed little regarding publication standards in economics.

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 (the Journal of Applied Econometrics and the 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 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 requires data sharing. 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].

[5] "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

6.3 Reporting Standards

In research, the devil 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 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 often not published with the respective paper, but can be submitted with the original article so that reviewers can check and see 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.

6.3.1 Randomized Trials and CONSORT

The most widely adopted reporting standard is the Consolidated Standards of Reporting Trials (CONSORT). Parallel to the 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 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.

6.3.2 Social Science Reporting Standards

Though a standard akin to CONSORT has not been formally adopted by social science or behavioral science journals, 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]. In psychological and behavioral research, an extension 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.

6.3.3 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.

The field of medicine has come up with too many checklists to describe them all individually. Acknowledging that every field and type of research is different, the Equator Network (Enhancing the Quality of Transparency of Health Research) serves as an umbrella organization that seeks to keep tabs on all the best reporting standards and help researchers find which reporting standard is most relevant for their research. See http://www.equator-network.org/ for more information.

7. CONCLUSION

Many of the activities described in this chapter require extra work. Before you run an experiment or conduct an impact evaluation, it takes extra time 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, and write a detailed protocol of the exact experimental setting, and then it takes additional time to post all of this publicly on the Internet with some sort of monitoring organization. However, we believe these steps are (1) not very difficult once you get used to them and (2) well worth the reward. You'll get statistical results 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 results precisely as they appear 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 that you'll know exactly where to go to make the changes. Importantly, since the results of your impact evaluation are transparent and reproducible, policy makers can be more confident in your results.

8. BIBLIOGRAPHY/FURTHER READING

1. Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review* 102 (4): 1206–40. doi:10.1257/aer.102.4.1206.

2. Almeida Rita, Sarojini Hirshleifer, David McKenzie, Cristobal Ridao-Cano, and Ahmet Levent Yener. 2012. "The impact of vocational training for the unemployed in Turkey: Pre-analysis Plan." *Poverty Action Lab hypothesis registry.*

3. Anderson, Richard G., and William G. Dewald. 1994. "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*, no. Nov: 79–83.

4. Baicker, Katherine, Amy Finkelstein, Jae Song, and Sarah Taubman. 2014. "The Impact of Medicaid on Labor Market Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment." *American Economic Review* 104 (5): 322–28. doi:10.1257/aer.104.5.322.

5. Baicker, Katherine, 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. 2013. "The Oregon Experiment — Effects of Medicaid on Clinical Outcomes." *New England Journal of Medicine* 368 (18): 1713–22. doi:10.1056/NEJMsa1212321.

6. Begg C, Cho M, Eastwood S, and et al. 1996. "Improving the Quality of Reporting of Randomized Controlled Trials: The Consort Statement." *JAMA* 276 (8): 637–39. doi:10.1001/jama.1996.03540080059030.

7. Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

8. Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika* 93 (3): 491–507. doi:10.1093/biomet/93.3.491.

9. Bhattacharjee, Yudhijit. 2013. "Diederik Stapel's Audacious Academic Fraud." *The New York Times*, April 26, sec. Magazine. http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.

10. Bruhn, Miriam, and David J. McKenzie.2009. "In pursuit of balance: Randomization in practice in development field experiments." *American Economic Journal: Applied Economics,* 1(4):200-232. doi: 10.1257/app1.4.200.

11. Carey, Benedict. 2011. "Noted Dutch Psychologist, Stapel, Accused of Research Fraud." *The New York Times*, November 2, sec. Health / Research.

http://www.nytimes.com/2011/11/03/health/research/noted-dutch-psychologist-stapel-accused-of-research-fraud.html.

12. Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. “Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*.” *The Quarterly Journal of Economics* 127 (4): 1755–1812. doi:10.1093/qje/qje027.

13. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin JA, Doré CJ. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ 2013; 346 (jan08 15), January 2013. ISSN 1756-1833. doi: http://dx.doi.org/10.1136/bmj.e7586..

14. Christensen, Garret, and Courtney Soderberg. "Manual of best practices in transparent social science research." Berkeley, CA: University of California. Retrieved April 23 (2015): 2016.

15. Dal-Ré, Rafael, John P. Ioannidis, Michael B. Bracken, Patricia A. Buffler, An-Wen Chan, Eduardo L. Franco, Carlo La Vecchia, and Elisabete Weiderpass. 2014. “Making Prospective Registration of Observational Research a Reality.” *Science Translational Medicine* 6 (224): 224cm1–224cm1. doi:10.1126/scitranslmed.3007513.

16. De Angelis, Catherine, Jeffrey M. Drazen, Frank A. Frizelle, Charlotte Haug, John Hoey, Richard Horton, Sheldon Kotzin, et al. 2004. “Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors.” *New England Journal of Medicine* 351 (12): 1250–51. doi:10.1056/NEJMe048225.

17. Cyranoski, David. 2014. “Cloning comeback”. *Nature, 505(7484):468-71.* ISSN 0028-0836, 1476-4687. doi: 10.1038/505468a.

18. DeLong, J. Bradford, and Kevin Lang. 1992. “Are All Economic Hypotheses False?” *Journal of Political Economy* 100 (6): 1257–72.

19. Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. “Replication in Empirical Economics: The Journal of Money, Credit and Banking Project.” *The American Economic Review* 76 (4): 587–603.

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

21. Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, and Katherine Baicker. 2012. “The Oregon Health Insurance Experiment: Evidence from the First Year*.” *The Quarterly Journal of Economics* 127 (3): 1057–1106. doi:10.1093/qje/qjs020.

22. Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345 (6203): 1502–5. doi:10.1126/science.1255484.

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

24. Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No ‘fishing Expedition’or ‘p-Hacking’ and the Research Hypothesis Was Posited ahead of Time.,” November. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

25. Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. “Testing for Publication Bias in Political Science.” *Political Analysis* 9 (4): 385–92.

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

27. Hamermesh, Daniel S. 2007. “Viewpoint: Replication in Economics.” *Canadian Journal of Economics/Revue Canadienne D’économique* 40 (3): 715–33. doi:10.1111/j.1365-2966.2007.00428.x.

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

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

30. Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. “Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration.” *Political Analysis* 21 (1): 1–20. doi:10.1093/pan/mps021.

31. Holm, Sture. 1979. “A Simple Sequentially Rejective Multiple Test Procedure.” *Scandinavian Journal of Statistics* 6 (2): 65–70.

32. Ioannidis, John P. A. 2005a. “Why Most Published Research Findings Are False.” *PLoS Med* 2 (8): e124. doi:10.1371/journal.pmed.0020124.———. 2005b. “Why Most Published Research Findings Are False.” *PLoS Med* 2 (8): e124. doi:10.1371/journal.pmed.0020124.

33. Ioannidis, John PA. 2008. “Effectiveness of Antidepressants: An Evidence Myth Constructed from a Thousand Randomized Trials?” *Philosophy, Ethics, and Humanities in Medicine* 3 (1): 14. doi:10.1186/1747-5341-3-14.

34. Johnson, Carolyn Y. 2012. Harvard professor who resigned fabricated, manipulated data, Us says – the boston globe. *BostonGlobe.com.*

35. Katz, Larry, Esther Duflo, Pinelopi Goldberg, and Duncan Thomas. 2013. "AEA E-Mail Announcement." American Economic Association. https://www.aeaweb.org/announcements/20131118_rct_email.php.

36. Knuth, Donald E. 1984. "Literate programming". *The Computer Journal*, *27*(2), 97-111. ISNN 0010-4620, 1460-2067. doi: 10.1093/comjnl/27.2.97.
37. Knuth, Donald E. "Literate programming." *Center for the Study of Language and Information*. ISBN 9780937073810.
38. Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to learn." *The Review of Economics and Statistics,* 91 (3): 437-456.

39. Laine, Christine, Richard Horton, Catherine D. DeAngelis, Jeffrey M. Drazen, Frank A. Frizelle, Fiona Godlee, Charlotte Haug, et al. 2007. "Clinical Trial Registration — Looking Back and Moving Ahead." *New England Journal of Medicine* 356 (26): 2734–36. doi:10.1056/NEJMe078110.

40. Long, J. Scott. 2009. "The workflow of data analysis using Stata". *Stata Press*. ISBN 9781597180474.

41. McCullough, B. D. 2007. "Got Replicability? The Journal of Money, Credit and Banking Archive." *Econ Journal Watch* 4 (3): 326–37.

42. McCullough, B. D., and H. D. Vinod. 2003. "Verifying the Solution from a Nonlinear Solver: A Case Study." *The American Economic Review* 93 (3): 873–92.

43. Moher, D, Kenneth F. Schulz, and Douglas G. Altman. 2001. "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel Group Randomized Trials." *BMC Medical Research Methodology* 1 (1): 2. doi:10.1186/1471-2288-1-2.

44. Moher D, Jones A, Lepage L, and for the CONSORT Group. 2001. "Use of the Consort Statement and Quality of Reports of Randomized Trials: A Comparative before-and-after Evaluation." *JAMA* 285 (15): 1992–95. doi:10.1001/jama.285.15.1992.

45. Neumark, David. 2001. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design." *Industrial Relations: A Journal of Economy and Society* 40 (1): 121–44. doi:10.1111/0019-8676.00199.

46. Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–31. doi:10.1177/1745691612459058.

47. Olken, Benjamin A. 2009. "Targeting analysis protocol." Technical report.
48. Olken, Benjamin A, Junko Onishi, and Susan Wong. 2010a. "Generasi Analysis Plan: WAVE III." *Technical report*.
49. Olken, Benjamin A, Junko Onishi, and Susan Wong. 2010b. "Indonesia's PNPM Generasi program : interim impact evaluation report." *Technical report* 59567. The World Bank.

50. Olken, Benjamin A., Junko Onishi, and Susan Wong. 2014. “Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia.” *American Economic Journal*: Applied Economics, 6(4):1-34. doi: 10.1257/app.6.4.1.

51. Schulz, Kenneth F., and David A. Grimes. 2012. "Allocation concealment in randomised trials: defending against deciphering." *The Lancet* 359(9306): 614-618. ISSN 01406736. doi: 10.1016/S0140-6736(02)07750-4.

52. Schulz, Kenneth F, and David A Grimes. 2005. “Multiplicity in Randomised Trials II: Subgroup and Interim Analyses.” *The Lancet* 365 (9471): 1657–61. doi:10.1016/S0140-6736(05)66516-6.

53. Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

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

55. Stanley, T. D. 2005. “Beyond Publication Bias.” *Journal of Economic Surveys* 19 (3): 309–45. doi:10.1111/j.0950-0804.2005.00250.x.

56. Taubman, S., Allen, H., Baicker, K., Wright, B. and Finkelstein, A., 2013. THE OREGON HEALTH INSURANCE EXPERIMENT: EVIDENCE FROM EMERGENCY DEPARTMENT DATA analysis plan.

57. Taubman, Sarah L., Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. 2014. “Medicaid Increases Emergency-Department Use: Evidence from Oregon’s Health Insurance Experiment.” *Science* 343 (6168): 263–68. doi:10.1126/science.1246183.

58. Turner, Erick H., Annette M. Matthews, Eftihia Linardatos, Robert A. Tell, and Robert Rosenthal. 2008. “Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy.” *New England Journal of Medicine* 358 (3): 252–60. doi:10.1056/NEJMsa065779.

59. Wydick, Bruce, Elizabeth Katz, and Brendan Janet. 2014. “Do in-Kind Transfers Damage Local Markets? The Case of TOMS Shoe Donations in El Salvador.” *Journal of Development Effectiveness* 6 (3): 249–67. doi:10.1080/19439342.2014.919012.