Open Science Manual

Relationship Science & Social Psychology Lab
Department of Psychology, Haverford College
Prof. Benjamin Le (@benjaminle)

Table of Contents

About This Document

This manual was assembled and is being updated by Professor Benjamin Le (@benjaminle), who is on the faculty in the Department of Psychology at Haverford

College. The primary goal of this text is to provide guidance to his senior thesis students on how to conduct research in his lab by working within general principles that promote research transparency using the specific open science practices described here. While it is aimed at undergraduate psychology students, hopefully it will be of use to other faculty/researchers/students who are interested in adopting open science practices in their labs.

Thanks & Acknowledgements

The first draft of this manual was assembled during a sabbatical generously provided by Haverford College (2017-18 academic year), and additionally supported by a Project TIER Faculty Fellowship. I'd also like to thank Tim Loving (formerly of University of Texas) for initially putting the open science movement on my radar screen, Lorne Campbell (University of Western Ontario) for the continual inspiration and advice, Carmel Levitan (Occidental College) for helping craft this vision, Kevin McIntyre (Trinity University) for the awesome preregistration template, Richard Ball and Norm Medeiros (both from Haverford College and the brains behind Project TIER) for their tireless work in promoting open science in undergraduate education, and Emma Waldner (Haverford College class of '18) for being a good sport and piloting many of the things discussed in this document in her senior thesis project.

Sharing, Usage, & License

If this document has been valuable to you, or if you have feedback/suggestions, please let me know (ble@haverford.edu). I'm very interested in continuing to improve the open science practices we're using in my lab and collaborating with colleagues who have a similar vision for their research.

1. Background

Jump back to
Table of Contents >>>

Around 2011-12, social and personality (SP) psychology began to experience a "replication crisis" and started to question whether published findings and commonly accepted results in the field would replicate, or if the statistically significant effects reported in a range of articles were due to Type I error that capitalized on publication bias or the decisions researchers were making during the research process (e.g., researcher flexibility and/or hypothesizing after-the-fact).

Read More:

- How Reliable Are Psychology Studies? (Yong, 2015; *The Atlantic*)
- A Reproducibility Crisis? (Weir, 2015; APA *Monitor*)
- Many Scientific Studies Can't be Replicated. That's a Problem (Achenbach, 2015, *The Washington Post*)
- The Replication Crisis in Psychology (Diener & Biswas-Diener, 2018)
- Estimating the Reproducibility of Psychological Science (Open Science Collaboration, 2015; *Science*); more info about this project here.
- Psychology is in Crisis Over Whether it's in Crisis (Palmer, 2016; *Wired*)
- The Reproducibility Crisis is Good for Science (Baker, 2016; *Slate*)

The issues encountered by SP psychologists are not exclusive to this particular subdiscipline (John et al., 2012 [PDF]; Open Science Collaboration, 2015), but SP psychology became the flashpoint for much of the subsequent discussion. The problems uncovered in SP research can certainly be generalized to other areas of psychology and other data-driven disciplines, and the procedures outlined here can be used/modified by researchers in other areas to help assure the integrity of their data and conclusions drawn from analyzing those data. In addition, **the general principles of transparency, reproducibility, and replicability, collectively known as *open science*, apply to all scientific disciplines.**

Video >>>

Dr. Simine Vazire talks about **open science** (~5 minutes)

Based on readings (e.g., Simmons et al., 2011 [PDF]), conversations with colleagues (e.g., Kevin McIntrye, who created the Open Stats Lab; Lorne Campbell, Tim Loving, Carmel Levitan), online discussions (e.g., Open Science Psychology, PsychMAP, Psychological Methods Discussion Group), listening to podcasts (e.g., The Black Goat), blogs (e.g., Andrew Gelman), and my involvement in Project TIER (Teaching Integrity in Empirical Research), **the guidelines and procedures articulated here will be employed in my lab (see statement on open science on my website), and are freely available for other researchers and teachers to use in their labs and classes.** Please note that I work at a small liberal arts college with a strong tradition of encouraging and supporting undergraduate research, and the procedures described in this document have been adopted for this context. They do not represent the entirety of open science practices and are not necessarily optimized for other research settings (e.g., large institutions/graduate programs), but are consistent with the goal of promoting research transparency more broadly.

Before getting into the specifics of what we'll be doing in my lab (e.g., the protocols for preregistration, open lab notebooks, codebooks, syntax, and archiving documentation), it's useful to provide the lay of the land to understand why these protocols are necessary, as well as introduce some terminology.

1.1. Reproducibility vs. Replicability

For the purposes of this document, *reproducibility* refers the act of reproducing results generated from a particular data set or study (sometimes known as "computational reproducibility"). For example, if my lab collects data and reports analyses from that dataset (e.g., in a publication), another lab should be able to generate identical results from that same data. Analysis that have been verified by independent labs should be more trusted than work that is not, or cannot be, verified. This, of course, assumes that my lab makes our data and analytic plans available to other labs; open science requires accessibility, transparency, and cooperation with regards to sharing data and experimental materials. **In short, *reproducibility* is desirable, and is facilitated by open science.**

Video >>>

Dr. Katherine Corker talks about **replication** (~6 minutes)

While reproducibility is an important goal to strive towards, SP psychologists have been especially concerned with *replicability*: other labs should be able to replicate results generated in my lab by collecting new data using similar procedures. Research findings that have been obtained from by multiple researchers from different studies/datasets should be more trusted than results coming from a single data set or isolated lab. Access to the materials used in my lab would be useful for their replication efforts, as would some insight into the the decision making processes we made along the way while collecting and analyzing our data. Again, open science provides a rationale and mechanism for making these things available. **In short, *replication* is desirable, and is facilitated by open science.**

1.2. Barriers to Replication & Reproducibility

There are multiple possible causes for results that are not reproducible or replicable. Some, such as outright fraud (point 1.2.1 below), are relatively rare, with broad consensus among scientists that such behavior is unethical and unacceptable. There is also agreement that other causes, like unintentional errors in the research process (1.2.2 below), are undesirable and that we should employ procedures that minimize these errors. There is less consensus, however, in the extent to which research materials and data should be made widely available so that reviewers, editors, colleagues, and other readerships have full access to verify the accuracy of the results. **Open science requires that these materials are accessible.**

Type I errors, or *false positives*, occur when the data shows a “statistically significant” effect (e.g., differences between groups or associations between variables), when in fact, there is no significant effect, due to random error or chance. For example, imagine if researchers ran a study to see if people with odd-numbered birthdays (e.g., May 1, 3, 5, etc.) were more extroverted than those with even-numbered birthdays (May 2, 4, 6, etc.). This seems like a crazy idea; there’s no good reason think that one’s birthday is related to differences in personality. If the data do show a significant difference, at commonly accepted levels for assessing statistical significance, between those with odd- and even-birthdays, you’d suspect that we made a Type I error in interpreting these results if we said that “people with odd-numbered birthdays are significantly more extroverted than those born on even-numbered days!"

Conclusions stemming from Type I errors in a particular dataset are unlikely to be replicated in subsequent studies; if an effect was due to random error, it's unlikely that other studies on that topic will show that same pattern of results. However, there are many reasons Type I errors can occur and make their way into the published literature, and below they are divided between those that are a function of how articles are selected for publication (1.2.3 below) and the decisions that researchers make throughout the research process (1.2.4 below).

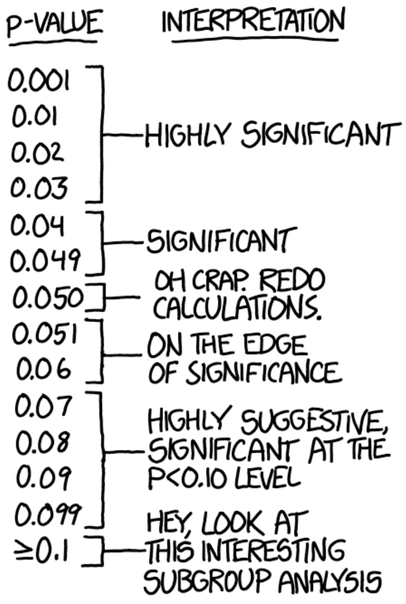
1.2.1. Data Fabrication & Research Fraud

Although cases of fabricated data have garnered a lot of press (see here and here for examples), these instances are relatively rare and not the root cause underlying most failed attempts to replicate or reproduce results. As such, the procedures described in this document are not primarily designed to address fraud, however, detailed and transparent documentation that describes data collection and analyses can certainly increase the credibility of that research, especially if made accessible.

1.2.2. Errors & Unintended Inaccuracies

Errors in data preparation, analyses, or reporting can produce or communicate results that are incorrect (e.g., a common reason *erratum,* or more precisely, *corrigendum*, are reported in journals upon discovery of such mistakes). These errors are typically due to failures in common research practices and end up introducing mistakes to the data files, analyses, or published results. The procedures described here, which are influenced by the Project TIER protocol, aim to provide a complete record of the processing (e.g., "cleaning") and statistical analysis of quantitative data, such that results should be reproducible with the data and documentation that accompanies each study. Ideally, these materials are made available to other researchers to reproduce, or potentially identify errors in, the data analysis. In addition, adherence to these protocols should increase attention to detail during the research process, decreasing the likelihood of errors occuring in the first place.

1.2.3. Type I Errors: Publication Bias & the File Drawer Effect


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

From: xkcd.com/1478/

Publication bias refers to the preference for research findings that are statistically significant within peer-reviewed journals (i.e., by editors, reviewers, readers); "significant" results are more likely to be published than non-significant results. There are undoubtedly many studies that don't show significant effects that have gone unpublished and are locked away in researchers' labs or buried on their hard drives (a.k.a., the "file drawer effect"). There is also a bias towards publishing novel findings, with replications less likely to be published. Essentially, readers see one article (or maybe a small handful of papers) showing a particular significant effect, but they don't have any knowledge of, or access to, the nonsignificant papers on that topic or the (un)successful replications. Given this bias, if reading only the published literature, how does one know the ratio of significant to nonsignificant results on a particular topic? It could be that the published paper reports results that are a function of at Type I error, and that there is an ocean of unseen evidence that does not support the published effect.

1.2.4. Type I Errors: HARKing, P-Hacking, & QRPs

HARKing refers to "hypothesizing after results are known"; essentially, that researchers collect data, do a wide range of exploratory analyses, and then "cherry pick" from those results and (mis)represent those as being derived from *a priori* hypotheses. This is also known as "data dredging" and "fishing." While it might surprise some students to learn that this practice occurs, it may be relatively common and has even been recommended as the preferred way to frame a research project (e.g., Bem, 2002 [PDF]). Engaging in numerous analyses and only reporting the (few) significant results inflates Type I error.

This isn't to say that exploratory analyses should be avoided; instead, one should simply clearly state (in publications/presentations etc.) which predictions were made *a priori* and which results were generated by exploratory analyses.

P-hacking is a general term that refers to the many different decisions that researchers make during data collection, preparation, and analyses that maximize the chance of obtaining significant effects. Is has been described as "researcher degrees of freedom" or "researcher flexibility," and this process of "trying things different ways" to see what leads to significant ($p < .05$) results is not necessarily done to be deceptive or manipulative, and in retrospect these decisions may be easy to justify or seem to make good sense (a.k.a., "motivated reasoning"). However, they have the cumulative effect of greatly inflating the likelihood of Type I errors (Simmons et al., 2011 [PDF]), and we generally focus on those (spurious?) significant effects and lose track of all of the ways of analyzing the data that do not yield significant effects. In short, these practices should be avoided:

a. Collecting data and doing interim analyses during data collection. If results are significant, stop collecting data; if not significant, keep collecting more data until it becomes significant.
b. Collecting data on several dependent variables (DVs), but only reporting results from the DVs that show significant effects. Similarly, creating DVs in different ways (e.g., a subset of items) to maximize significant effects.
c. Running multiple tests and only reporting on those that yielded significant results.
d. If initial results do not show significant effects, re-run analyses with control variables or interaction terms, or doing subgroups analyses.
e. Eliminating data (e.g., "outliers" or particular demographic groups) in an attempt to find stronger results.
f. Designing an experiment with multiple conditions, but ignoring or collapsing conditions if analyses based on all conditions is not significant.
g. Run multiple experiments and only report those that "worked."

Activity >>>

P-hack your way to significance @ fivethirtyeight.com! (full article here)

Video >>>

John Oliver talks about **p-hacking** (first 5 minutes of this video; warning: NSFW language

While these **questionable research practices (QRPs)** may seem obviously problematic given what we know now, prior to 2011, they were commonplace in many labs across the various subdisciplines in psychology (John et al., 2012 [PDF]). In addition, some of these practices sound perfectly reasonable, both in terms of theory (e.g., add control variables) and cleaning the data (e.g., remove outliers), and if stated a priori, they maybe be quite acceptable. But based on what we have learned about the impact of researcher degrees of freedom and motivated reasoning on false positives, we should avoid doing things things after the fact without expliciting stating they were done *post hoc*. It is no longer acceptable to misrepresent analyses as being planned when in fact results were generated via trial-and-error through various the methods of researcher degrees of freedom described above.

1.3. What’s the Problem?

At the risk of stating the obvious, scientists should seek unbiased answers to questions and aim to uncover the truth with their work. So anything that compromises these goals should be avoided and actively discouraged.

There are also some important practical implications to keep in mind. First, keeping scientific findings, especially nonsignificant effects, hidden can lead to wasting resources on investigating research questions that have already been shown not to be fruitful. How would you like to spend years chasing an idea, not knowing that others have already gone down that path unsuccessfully? In addition, at a broader level, interventions and policy maybe be developed based on scientific evidence. What if it turns out that the respective evidence underlying these interventions/policies was not as strong as what is found in the published literature?

1.4. What are Possible Solutions?

So, what solutions do we have to address these problems with reproducibility and replicability?

- Commit to your hypotheses and data analytic strategies prior to commencing with data collection/analysis.
- Create detailed documentation to accompany your research and data analysis, such that other researchers can replicate your study design and/or reproduce your results from your data.
- Make your study hypotheses, design/materials, and data accessible to other researchers, even if you do not end up publishing that work.
- Cooperate with other researchers who attempt to replicate your work.
- Change the academic publishing incentive system, such that replications and rigorous work reporting non-significant results are valued.

Essential Reading:

- Enhancing Transparency of the Research Process to Increase Accuracy of Findings: A Guide for Relationship Researchers (Campbell et al., 2014; in *Personal Relationships*) [PDF]

1.5. Purpose & Organization of this Document

The rest of this document provides suggestions and tools that we are using in my lab at Haverford College to work towards these solutions, in the form of:

Section 2: Preregistration of your hypotheses and data analytic plans
Section 3: Using an "Open Lab Notebook" with your research group
Section 4: Creating codebooks to document your dataset
Section 5: Using annotated code/syntax for all statistical analyses
Section 6: Strategies for managing, archiving, sharing hypotheses, research materials, and data using OSF
Section 7: A comprehensive checklist for a senior thesis research project in my lab, including the open science practices described throughout this document