# Checklist for Scientific Papers

http://github.com/jpeelle/paperchecklist

| Item | Yes | No | Why not?* | See section |
|------|-----|----|-----------|-------------|
| **Openness** | | | | |
| Made stimuli/materials publicly available? | | | | 1 (page 2) |
| Made data publicly available? | | | | 1 (page 2) |
| Made data analyses publicly available? | | | | 1 (page 2) |
| Submitted a preprint? | | | | 3 (page 3) |
| Published in an open access journal? | | | | 4 (page 3) |
| Published original figures CC-by prior to journal? | | | | 2 (page 2) |
| **Statistics** | | | | |
| Preregistered the experiments? | | | | 5 (page 4) |
| Documented a rule for stopping data collection (e.g., number of participants, BF) before starting? | | | | 6 (page 4) |
| Explained all conducted tests? I.e., included a statement such as "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study." | | | | 7 (page 5) |
| Collected an informative data set? | | | | 8 (page 5) |
| Plotted the data informatively? | | | | 9 (page 6) |
| Avoided p values? | | | | 10 (page 6) |
| Included effect sizes? | | | | 10 (page 6) |
| Included confidence intervals (preferably Bayesian)? | | | | 11 (page 7) |

\* D = don't know how, H = too hard, E = ethical constraints make it impossible, S = scary, B = bad for science, O = other (specify)

*The goal of this checklist is not to lay down a set of immutable rules, but to promote discussion about—and implementation of—best practices in research. For more see the README file on github.*

# 1    Making stimuli/materials, data, and analyses publicly available

There are many reasons to think that making our data and materials publicly available are a good thing for science, some of which I summarize in my blog post on open science:

> My own take is that sharing materials and data promotes more accurate science because (a) it facilitates checking of one's data and analysis by others, and (perhaps equally importantly) (b) our internal checks on data organization and accuracy are typically better if we know the materials will be publicly available. So, even if no one ever looks at them, the mere fact that we are sharing them will improve the accuracy of our research. I know this is certainly true in my own work. It may slow down the process, but I think this simply reflects a speed/accuracy tradeoff shifting towards more accuracy. In the long run this seems like only a good thing for scientific discovery.

Making materials publicly available can also save time down the road—most of us agree to make data available "upon request", but this can take many frustrating hours if the files have been moved or if the person responsible for a research project has moved on. Although organizing and making available materials and data at the outset takes time now, it can save a lot of time later.

Many datasets can easily be shared using websites such as GitHub, figshare, or the Open Science Framework. For neuroimaging data, `openfmri.org` is an excellent choice for raw data. At the very least, unthresholded statistical maps can be uploaded to `neurovault.org` (Gorgolewski et al., 2015), which allows browsing and for other researchers to download maps for meta analyses.

See also:

- Gorgolewski KJ et al. (2015) NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics* 9:8. doi:10.3389/fninf.2015.00008

- Morey R et al. (2016) The Peer Reviewers' openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science* 3:150547. doi:10.1098/rsos.150547

- Rouder JN (In press) The what, why, and how of born-open data. *Behavior Research Methods*. doi:10.3758/s13428-015-0630-z

- Vines, TH et al. (2014) The availability of research data declines rapidly with article age. *Current Biology*, 24, 94–97. doi:10.1016/j.cub.2013.11.014

# 2    Open-sourcing figures

Most traditional journals will copyright the content of an article, including its figures, meaning that permission must be requested from the journal for any future use. In some (many?) cases this will be granted for free for academic purposes, but not always (a recent figure I reproduced from a *JAMA* article cost $250). Even if a paper isn't published in an open access journal, as the author you can license your figures before submitting to the journal. For example, if you use the popular Creative Commons Attribution-By license, the work is already licensed and so the journal does not own the

copyright to the figure. Other authors (including you) are then free to easily re-use your figures, provided they provide attribution, which is generally what we want in science. For an example, see `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2722435`.

## 3 Submitting preprints

Preprints generally refer to author-formatted manuscripts submitted to websites without pre-publication peer-review, which can be accessed by anyone. Preprint servers include arXiv and bioRxiv, and PeerJ.

[As a side note, it's been suggested that the term "preprint" is a misnomer in the digital age—other suggestions (some more serious than others) include dispatch, sciencegram, or (maybe someday) "paper". But we're all talking about the same thing.]

The argument for preprints might be summarized like this. Traditional scientific publishing takes too long, and thus delays the publicizing of research (which slows down science, and unfairly prevents authors from staking a claim to their ideas—important in competitive fields). In addition, research is frequently published in journals that charge for access, restricting who can benefit from the science. Finally, publishing a preprint may allow authors to get faster feedback on their work, resulting in a higher quality final product.

Some have argued against preprints. Like many things that go against traditional models they may be riskier for early stage scientists and trainees, particularly those not coming out of "superstar" labs. Preprints may not be appropriate for every manuscript, but should at least be considered. (In our lab, there is frequently no compelling reason to avoid submitting a preprint, apart from the extra work involved.)

## 4 Publishing in open access journals

There are different flavors of open access, but the basic idea is that once a paper is published, anyone can access and read it. Open access contrasts with traditional access, in which once a paper is published only readers who have paid for access to the journal or have institutional access can read it.

Some people argue that in practice open access is not important, because the people reading research papers are almost certain to have institutional access. There are several reasons I disagree with this sentiment:

1. As a scientist at a research university, I frequently find journals to which my university does not subscribe. I ask friends at other universities, or on twitter, for PDFs.

2. As a scientist at a research university, I frequently find myself working on manuscripts while not connected to my campus network (I do not pay extra fees for off-campus access that my university requires).

Of course, for specific papers I am very interested in, I will find a way to track down the paper. But many times, I'm exploring a new topic, or looking for a reference for a particular point. If there are two papers that seem appropriate, and I can access one but not the other, guess which one I'll end up reading and citing? So, as an author, I'd prefer to have my papers get cited more, and thus be more available.

People apart from researchers with institutional affiliations do, in fact, read research articles, and in many cases have helped to fund the study through tax dollars. Having paid for the research, as it were, it is only fair they are able to access it. (For some of the work in my lab, this may not happen often, but for clinical and medical research this point takes on whole new importance.)

One challenge to open access is the publication fee, which is frequently higher than for a traditional journal. A second challenge is that some open access journals are perceived to be of lower quality than

their traditional counterparts. So, the theoretical and practical advantages of an open access publication sometimes must be tempered by these concerns.

# 5   Preregistration

In many research studies and clinical trials, the researchers have hypotheses ahead of time regarding aspects such as the likely (or interesting) outcome, how many participants are needed to achieve a meaningful result, and so on. Unfortunately, because manuscripts are often written after the data collection has completed, it can be tempting to alter these predictions post-hoc. "Had we been thinking clearly, we certainly *could* have predicted such-and-such an outcome"—which can make for a better "story", or even change the statistical significance (predicted tests are sometimes not as rigorously controlled for). Preregistration involves documenting *a priori* aspects of the design and predictions in order to make transparent things that were truly planned or predicted before data collection (or manuscript writing), and those that were not.

(Of course, as researchers we are entitled to perform analyses that were not originally planned, it just forces us to acknowledge that these analyses were not originally planned.)

Preregistration is intended to help avoid a number of specific problems that can creep in to research studies, including:

1. Failure to report null results

2. Arbitrary stopping of data collection (see §6), which can make it easier to achieve a significant p value.

3. Changing of outcome measures (which may result in a significant result, when one was not found using the originally specified outcome measure)

"Registered reports" is a term used by some journals (including *Cortex*, *eLife*, and *Royal Society Open Science*) to describe research studies whose design has been submitted to the journal ahead of time. In some cases, the manuscript can be accepted in principle based on the motivation and design, regardless of the outcome. Doing so is intended to lower the bar to publishing null (or confusing) results: if the design is solid and adequately powered, the findings should be published, and are equally informative regardless of whether they support the original hypothesis.

See also:

- Chambers CD (2013) Registered Reports: A new publishing initiative at *Cortex*. *Cortex* 49:609–610.

- Scott S. Pre-registration would put science in chains. `http://www.timeshighereducation.co.uk/comment/opinion/pre-registration-would-put-science-in-chains/2005954.article`

# 6   Deciding on when to stop data collection ahead of time

One aspect about data that can seem surprising at first is that statistical results can jump around as more data are collected, particularly with small sample sizes. That is, the effect size—or the p value of a particular effect—can go from small to large, or significant to not significant, as more data enter in to the analysis. So, one one hand, having more data "always" leads to a more accurate estimate of an effect. But, if a researcher bases their data collection on the current result, bias can occur.

Consider a not uncommon situation in which a student collects behavioral data from 20 participants and would like to know whether he needs to run more participants. If the effect of interest is p < .05, then

20 participants seems like a nice round number and he may decide to stop. If the effect comes out at p = 0.9, this seems like a clear answer, too. But if the effect is p = .08, he may decide to run 10 more people to "see if the effect is real". The problem is that, if one continues to collect data until a significance value is achieved, it is more likely to happen (and thus the p value is invalid).

There are ways to include sequential analyses that are preplanned (Lakens, 2014), and Bayesian statistics may be less affected (Rouder, 2014, Schönbrodt & Wagenmakers). These situations aside, it's safer to specify a stopping rule ahead of time.

See also:

- Lakens D (2014) Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology* 44:701–710. doi:10.1002/ejsp.2023

- Rouder JN (2014) Optional stopping: No problem for Bayesians. *Psychonomic Bulletin and Review* 21:301–308.

- Schönbrodt FD, Wagenmakers E-J. Bayes factor design analysis: Planning for compelling evidence. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2722435

# 7  Explaining all conditions, tests, and data exclusions

In a null hypothesis significance testing (NHST) framework, the more tests conducted, the better the chance of finding a significant effect even when one doesn't exist. Thus, it's important to be clear about how many tests were conducted, whether any data were excluded, and so on. (As researchers this should also mean that we are clear in our own minds when we are indeed testing a hypothesis vs. conducting exploratory research.)

See also:

- Dorothy Bishop's take, featuring "The amazing Significo" http://deevybee.blogspot.co.uk/2016/01/the-amazing-significo-why-researchers.html

- Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22:1359–1366. doi:10.1177/0956797611417632

- Simmons JP, Nelson LD, Simonsohn U (2012) A 21 Word Solution. doi:10.2139/ssrn.2160588

# 8  Collected an informative data set

As discussed in the section on effect sizes, in many cases we don't just want to know whether a non-zero effect exists, but we'd like to know the magnitude of the effect. For example, if I am studying the effect of hearing loss on speech intelligibility, I'd like to know how much of a change in behavior I might see for a 5 dB change in hearing: it matters whether there is a 25% drop in accuracy for a 1% drop in accuracy.

Unfortunately, many power analyses (and more prevalent rules of thumb) are structured around finding a significant result, and not accurately estimating parameters. The number of participants needed to accurately estimate parameters in a behavioral study, of course, varies, but estimates are typically in the range of hundreds of participants, rather than tens of participants.

See also:

- Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munaf MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365-376.

- Kelley K, Maxwell SE (2003) Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. Psychol Methods 8:305-321.

- Maxwell SE (2000) Sample size and multiple regression analysis. Psychol Methods 5:434-458.

- Schönbrodt F, Perugini M (2013) At what sample size do correlations stabilize? J Res Pers 47:609-612.

# 9   Plotting data informatively

Bar graphs have been the standard in many years for showing means and/or effect sizes, but do a poor job of showing the distribution of the data: data with a myriad of underlying distributions can give rise to identical means (and standard deviations). With modern graphing software (e.g., R or Matlab) it's trivial to produce more informative plots, showing either every data point, or at least a fuller picture of the distribution (for example, in a violin plot). Showing all the data gives readers a fuller set of the truth of the dataset (including the validity of any underlying assumptions that play in to the statistical analyses).

See also;

- Drummond GB, Vowler SL (2011) Show the data, don't conceal them. *Journal of Physiology* 589.8:1861-1863. doi:10.1113/jphysiol.2011.205062

# 10   Avoiding NHST and p values

Writing in 1994, Cohen said:

> Like many men my age, I mostly grouse. My harangue today is on testing for statistical significance, about which Bill Rozeboom (1960) wrote 33 years ago, "The statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

> And today, they continue to continue. And we, as teachers, consultants, authors, and otherwise perpetrators of quantitative methods, are responsible for the ritualization of null hypothesis significance testing (NHST; I resisted the temptation to call it statistical hypothesis inference testing) to the point of meaninglessness and beyond. I argue herein that NHST has not only failed to support the advance of psychology as a science but also has seriously impeded it.

Difficulties with NHST include:

- A dichotomous estimate ("statistical significance") rather than an interval

- Misinterpretation of what p values reflect

- Focus on statistical significance rather than effect size

See also:

- Cohen J (1994) The earth is round (p < .05). *American Psychologist* 49:997–1003. doi:10.1037/0003-066X.49.12.997

## 11 Confidence intervals

So, confidence intervals, as generally understood, seem more in line with what we as researchers want to know.

That being said, Morey et al. (2016) argue that most of us have an incorrect understanding of confidence intervals, which can have nonintuitive and/or unexpected properties. They suggest using Bayesian versions (credible intervals) instead.

See also:

- Cumming G (2014) The new statistics: Why and how. *Psychological Science* 25:7–29. doi:10.1177/0956797613504966

- Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J (2016) The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review* 23:103-123. doi:10.3758/s13423-015-0947-8