Checklist for Scientific Papers

http://github.com/jpeelle/paperchecklist

Item	Yes	No	Why not?*	See section
Openness				
Made stimuli and materials publicly available?				2 (page 2)
Made data publicly available?				
Made data analyses publicly available?				
Submitted a preprint?				4 (page 3)
Published in an open access journal?				5 (page 4)
Published original figures CC-by prior to journal?				3 (page 3)
Statistics				
Preregistered the experiments?				6 (page 5)
Documented a rule for stopping data collection (e.g., number of participants, BF) before starting?				7 (page 6)
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."				8 (page 6)
Collected an informative data set?				9 (page 7)
Plotted the data informatively?				10 (page 8)
Avoided p values?				11 (page 8)
Included effect sizes?				11 (page 8)
Included confidence intervals (preferably Bayesian)?				12 (page 9)

* D = don't know how, H = too hard, S = scary, E = ethical constraints make it impossible, 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 Some general thoughts on scientific practices

Science is about discovering truths about the world around us and how we act in it. We make mistakes, of course, and go down wrong paths, but we continually strive to improve our ideas and test our theories so that we can ultimately increase our knowledge. What I've tried to do here is keep note of some suggestions and writings that push me towards performing the process of discovery in the best way possible.

Most of us (I hope) don't set out to intentionally skew our results or act for personal gain rather than scientific truth. It's been important for me to realize that, nevertheless, many of our intuitions and common practices can lead to distorting the pattern or interpretation of results. Science is still a human an endeavor and will always be prone to our own mistakes and biases. I think the important question is, how can we, as scientists, minimize bias and error, and maximize the usefulness and truthiness of our results? Doing so is not always easy, but will hopefully provide us with better—more accurate, reliable, and reproducible—results.

See also:

- Neuroskeptic (2012) The nine circles of scientific hell. *Perspectives on Psychological Science* 7:643-644. doi:10.1177/1745691612459519
- Video of Open Science session at Psychonomic Society 2016: https://www.youtube.com/watch?v=CmOkGEe4a7A

2 Making stimuli/materials, data, and analyses publicly available

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

> [S]haring 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 materials and data takes time now, it can save *a lot* of time later.

When sharing data, it's probably best to share it on a third party hosted site, rather than a personal or institutional website (unless your institution has a server set up specifically for data sharing). In many cases data posted online ceases to be accessible due to websites disappearing, links changing, and so on. Most third party hosting sites are a safer bet for the long-term availability of your materials.

Many datasets can easily be shared using websites such as GitHub, figshare, or the Open Science Framework. For neuroimaging data, `openneuro.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
- Gorgolewski KJ, Poldrack RA (2016) A practical guide for improving transparency and reproducibility in neuroimaging research. *PLOS Biology* 14:e1002506. doi:10.1371/journal.pbio.1002506
- 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 (2016) The what, why, and how of born-open data. *Behavior Research Methods* 48:1062–1069. 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

3 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 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 typically 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 (for example, in review papers), provided they provide attribution, which is generally what we want in science.

A couple of examples:

- Peelle JE, Wingfield A (2016) The neural consequences of age-related hearing loss. *Trends in Neurosciences* 39:486-497. doi:10.1016/j.tins.2016.05.001 (Figure 1A)
- Schönbrodt FD, Wagenmakers E-J (2018) Bayes Factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review* 25:128–142. doi:10.2139/ssrn.2722435

4 Submitting preprints

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

The argument for preprints can be summarized in the following way. 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 preprints may be riskier for early stage scientists and trainees, particularly those not coming out of "superstar" labs. And, if comments are public, then feedback can also be public, which may be intimidating. 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 and the fact that it is scary.

See also:

- Bhalla N (2016) Has the time come for preprints in biology? *Molecular Biology of the Cell* 27:1185–1187. https://doi.org/10.1091/mbc.e16-02-0123
- Niko Kriegeskorte's "The selfish scientist's guide to preprint posting" http://nikokriegeskorte.org/2016/03/13/the-selfish-scientists-guide-to-preprint-posting/

5 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 read it. Open access contrasts with traditional publishing models, 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 traditional journals will make content freely available after an embargo period.)

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

1. Many researchers are at institutions who may *not* have institutional access, due to the fees involved. For example, Erin McKiernan tweeted: "For those arguing there's no access problem, please come work here in Mexico for a month. Then tell me what you think about #openaccess."
2. As a scientist at a research university in the United States, I frequently find journals to which my university does not subscribe. I ask friends at other universities, or on twitter, for PDFs. (I'm not the only one—see the twitter tag #icanhazpdf for examples.)
3. As a scientist at a research university in the United States, 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). Thus, I'm frequently without access to journals to which my institution has a subscription.

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? As an author, I'd prefer to have my papers get cited more, and thus I try to make my own papers as easy to access as possible.

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 sometimes higher than for a traditional journal (which makes money through subscription fees). A second challenge is that some open access

journals are sometimes perceived to be of lower quality than their traditional counterparts. So, the theoretical and practical advantages of an open access publication must sometimes be tempered by these concerns.

See also:

- Martin Eve's "Open access in a time of illness"
  https://www.martineve.com/2016/04/07/open-access-in-a-time-of-illness/
- Tennant JP, Waldner F, Jacques DC et al. (2016) The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research* 5:632. doi:10.12688/f1000research.8460.1

6 Preregistration

In many studies, 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 for this outcome, 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", but also 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, preregistration 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 §7), 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).

Nothing fancy is needed to preregister a study: simply a place to make available the analysis plan, and document when the plan was generated. Preregistration is more credible if it is hosted by a third party to make the dates and edits transparent. The website aspredicted.org makes the process particularly painless.

"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. Publishing results regardless of the significance of the outcome particularly helps with meta-analyses, which try to judge the overall effect across multiple studies (and whose accuracy depends on the availability of both positive and negative outcomes).

(If one isn't ready to officially preregister a study, a potential baby step in that direction is to create an internal document with an analysis plan to see how that goes—although this isn't really preregistration, it is still a useful step in constraining and explaining our analyses.)

See also:

- Chambers CD (2013) Registered Reports: A new publishing initiative at *Cortex*. *Cortex* 49:609–610. doi:10.1016/j.cortex.2012.12.016
- Nosek BA, Lakens D (2014) Registered reports: A method to increase the credibility of published results. *Social Psychology* 45:137-141. doi:10.1027/1864-9335/a000192
- 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

7 Deciding on when to stop data collection ahead of time

One aspect about data that can seem surprising at first is that statistical results jump around as more data are collected, particularly with small sample sizes (see Simmons et al., 2011, Figure 2). 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 practices on the current result, bias can occur.

Consider a not uncommon situation in which a researcher collects behavioral data from 20 participants and would like to know whether they need to run more participants. If the effect of interest is $p < .05$, then 20 participants seems like a nice round number and they may decide to stop. If the effect comes out at $p = 0.9$, this seems like a clear answer, too, and the researcher may also stop (quite possibly putting the experiment in the proverbial file drawer). But if the effect is $p = .08$, they 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). (If 30 participants don't work, how about 31? 32? 37?)

There are ways to conduct sequential analyses that are preplanned (Lakens, 2014), and Bayesian statistics may be less affected (Rouder, 2014, Schönbrodt & Wagenmakers). These situations aside, it may be 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. doi:10.3758/s13423-014-0595-4
- Schönbrodt FD, Wagenmakers E-J (2018) Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review* 25:128–142. doi:10.3758/s13423-017-1230-y

8 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. For example, with a typical cutoff

of $p < .05$, if I conduct 20 tests, I am likely to find one "significant" difference by chance (when no true difference exists). 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.)

Simmons et al. (2011) demonstrated this with example experiments and simulations with a large number of "experimenter degrees of freedom" to demonstrate that with enough tests it is possible to find "significant" results, and even to write these up in a convincing way. Explanations can be especially convincing when all of the tests aren't described, but only the ones leading a significant result. To avoid these situations, Simmons et al. (2012) suggest a "21 word solution" that makes it clear to readers whether all tests were (or were not) explained in the text:

> We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

See also:

- Video from Neuroskeptic on p-hacking: https://www.youtube.com/watch?v=A0vEGuOMTyA
- Dorothy Bishop's take, featuring "The amazing Significo" http://deevybee.blogspot.co.uk/2016/01/the-amazing-significo-why-researchers.html
- Carp J (2012) On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience* 6:1-13. doi:10.3389/fnins.2012.00149
- 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

9 Collecting 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; we'd like to know *how large* the effect is. 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 or 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, of course, varies, but estimates are typically in the range of hundreds of participants, rather than tens of participants. So—based on my experience and reading—generating an informative, accurate, and hopefully reproducible dataset typically requires many more participants than is typical for the field.

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. *Nature Reviews Neuroscience* 14:365-376. doi:10.1038/nrn3475
- Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* 2:e124. doi:10.1371/journal.pmed.0020124

- Kelley K, Maxwell SE (2003) Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods* 8:305-321. doi:10.1037/1082-989X.8.3.305
- Lakens D, Evers ERK (2014) Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science* 9:278-292. doi:10.1177/1745691614528520
- Maxwell SE (2000) Sample size and multiple regression analysis. *Psychological Methods* 5:434-458. doi:10.1037/1082-989X.5.4.434
- Schönbrodt F, Perugini M (2013) At what sample size do correlations stabilize? *Journal of Research in Personality* 47:609-612. doi:10.1016/j.jrp.2013.05.009

10 Plotting data informatively

Bar graphs have been the standard in many years for showing means and 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 raincloud 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

11 Potentially 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.

More than 20 years later, the situation is largely unchanged. Difficulties with NHST include:

- The true null hypothesis is always false (i.e., means between two conditions will not be identical to an infinite precision)
- A dichotomous estimate ("statistical significance") rather than an interval
- Misinterpretation of what p values reflect

See also:

- Geoff Cumming's "Dance of the p values" https://www.youtube.com/watch?v=5OL1RqHrZQ8
- Cohen J (1994) The earth is round (p < .05). *American Psychologist* 49:997–1003. doi:10.1037/0003-066X.49.12.997
- Loftus GR (1996) Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science* 5:161–171. http://www.jstor.org/stable/20182423
- Taleb NN (2016) The meta-distribution of standard p-values. http://arxiv.org/abs/1603.07532
- Wagenmakers EJ, Morey RD, Lee MD (2016) Bayesian benefits for the pragmatic researcher. https://webfiles.uci.edu/mdlee/BayesianBenefitsSubm.pdf
- Wagenmakers EJ, Verhagen AJ, Ly A (2016) How to quantify the evidence for the absence of a correlation. *Behavior Research Methods* 48:413–426. doi:10.3758/s13428-015-0593-0
- Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. 70:129–133. doi:10.1080/00031305.2016.1154108

12 Using confidence intervals

The first point here is that any sort of interval or distribution suggests (quite correctly) that any number our statistical models produce is an *estimate* rather than the "truth". You would react very differently if I said "I will give you $100" compared to "I will give you an amount between $0 and $210, or maybe I will even ask you to *give* me $10."

Similarly, in a research study I may find that hearing loss is related to participants' memory for speech, with a standardized beta of 0.3. How much confidence do I have in this number of 0.3? It matters if the likely range is 0.28–0.32, or if it might be 0.0–0.6 (particularly because 0.0 would not be very interesting!).

So, thinking about distributions rather than point estimates is probably closer to what we actually care about.

Given that we care about distributions, there are many ways to report uncertainty around an estimate. The standard error (commonly used for error bars) gives an indication of the standard deviation of the population mean (i.e., a distribution of sample means). But a point estimate, plus and minus 1 standard error, is only 68% likely to contain the true mean.

Confidence intervals, on the other hand, should contain the population mean with a given accuracy (that is, the population mean should be within the 95% confidence interval 95% of the time). Thus, confidence intervals (CIs) are better suited for inference (that is, for drawing conclusions based on data). Reporting CIs instead of point estimates, and using CIs instead of standard errors, is a step in the right direction.

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. At the very least, informed by their simulations, it is worth checking assumptions about whether a confidence interval functions as we expect it to.

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