How to do cumulative science as a masters student

Version 2.2, 2019-06-28 – comments welcome: alecristia@gmail.com (Alex Cristia)

Recent times have seen remarkable examples of poor science (e.g., data fabrication) which, albeit unfortunate in themselves, have spurred interest in how to change scientific practices to optimize the discovery of Truth. A number of specific proposals have been put forward, and their likely positive and negative impacts are today the subject of much discussion. Some of these ideas are very likely to stick around, since there are few costs and considerable gains to their application. Indeed, salient journals such as *Psychological Science* have created "badges" to set out work that applies such recommended practices.

As a research masters student, you are now practicing science. What will your practice be like? The present document seeks to provide you and your advisor with entry points to this debate, to help you make informed decisions. While some of the concepts we will use have been around for decades, others are relatively new. Therefore, it is likely that you and your advisor will need to discuss which of these items you can *realistically* apply during your masters project.

Throughout your project's life cycle

A) Build traceability. Think about how you are going to keep track of what you did, when and why, from as early as possible in the project life. Will you keep a paper notebook, an online notebook, a wiki, a blog, a file on dropbox? Will you use git and/or version control? Where will you store your files, and how? How will you back them up? When you close a project, you should be able to provide a relatively complete pipeline, at least from data collection to conclusions, and ideally also from inception to implementation.

Entry points: tips for "old" style; tips for "new" style; example; general tips for data analysis; a more "advanced" version

B) Facilitate the integration with previous and future literature. How does this study build on previous ones? How will future studies profit from what I find? This goal can shape your decisions at many points. For example, at the implementation stage this consideration may lead you to use very similar methods to maximize comparability, or conversely very different ones to be able to "triangulate" a prediction (meaning provide a logical proof from a completely different angle); at the reporting stage, this consideration should lead you to report effect sizes that can be integrated with others in a meta-analysis (instead of or in addition to more "elegant" albeit unique analyses).

Entry points: Cumming (2013)

When looking for an idea

C) Consider carrying out a replication. One of the most astounding features of the new trends is the value that replications have acquired. Many journals (including e.g. *Psychological Science*) are opening their pages to straight replications through pre-registered reports. If you are considering doing a replication, read more to find out how to decide which work to attempt to replicate, and how to do it properly.

Entry points: Nosek & Lakens (2014); particularly p. 622 of Nosek et al. (2012); if you are a modeler, check out ReScience

Once you have identified a potential empirical project

D) Understand what came before. All empirical research builds on previous work, and your study will be a new brick laid on top of the others. So you should ask yourself, if you try to put a new brick on this "wall", will it crumble? How large are the effects, and how reproducible are they? Might (some of) these results be, in fact, the consequence of a confound, a bias in the experiment or the field, ...? Should you still venture into it? How will *you* avoid bias?

Entry points: Ioannidis (2005), Lakens (2013)

When you are designing your empirical project

E) Know what you are looking for, so as to avoid false positives. What is the question you want to answer? What is the most precise way to answer it? How can you tell if you have answered the question? False positives are cases where a result is thought to be meaningful when in fact it was due to chance. One common situation that inflates the risk of false positives is that of multiple testing, when one looks at many,

many aspects of a dataset – eventually, one is bound to find something that "looks good," even if it is by chance. This false-positive risk inflation can, fortunately, be prevented by keeping one idea in mind: Be clear about what your research goal is. This touches many aspects of the design, from the hypothesis (avoid wording that is open to post hoc re-interpretations - "the distinctiveness between 2 faces" → did you mean "number of pixels differentiating 2 pictures" so the exact same face but with black or white pigmentation gets a high score, or did you mean "the number of facial features", or...?) to the actual study implementation (make sure each one of your experimental conditions and each one of your potential covariates is justified given your research questions). Define which specific statistical method(s) you will use to address your research question(s). You should be able to draw a picture representing your expected results.

Entry points: Simmons et al. (2011)

F) Establish a rule to stop data collection, so as to avoid false negatives. How many data points do I need to be sure of what my result is? A second risk that you are facing is that of false negatives, a result that is not statistically significant when, in actual fact, it should be. The risk of false negatives is inflated by under-powered research. Therefore, you need to make an informed decision as to how much power you will require, and thus how much data you need to collect. You may want to rely on *a priori* power analyses to calculate the ideal sample size, if enough people have used implementations very similar to yours. A simpler one is picking an exact (if arbitrary) number of participants based on previous research, and hope for the best. Many other stopping rules exist, including some said to optimize effort.

Entry points: Lakens & Evers (2014); see also this presentation I gave at a lab meeting

Right before data collection starts

G) Pre-register. Have you really thought of everything? Pre-registering is a terrific exercise to make sure that you are planning ahead, it lends credibility to your results, and it can further gain you "badges" for traceability in some journals. It consists of a frozen, time-stamped version of your research project, that will help you decide which parts of your results are without false positive/negative risk inflation. At this point, you might want to "pilot" (run 2-3 friends, or create data to look like the real one) and check: Is your code working as you expected? Have you set up your analyses correctly? Is there something about your hypotheses and predictions that does not square with your analysis scheme?

Entry points: Nosek & Lakens (2014)

While you are collecting data

H) Be aware that you are not a passive observer. A great deal has been said about observer bias. The current push for cumulative science has nothing to say on this point, and the only current discussion for this phase of your project bears on whether there are ethical ways to run "preliminary analyses".

Entry points: Sagarin et al. (2014); for observer bias, see e.g. Rosenthal (1994)

After data collection is finished

I) Close your projects. Science costs money and time, and for your project that means your time and that of everyone else who participated. If you don't write up your results, whatever they are, someone else may also spend that time and money – so you've been at the source of a double (or triple, or ...) "waste". So write up a report, whatever the results are. If the study "didn't work", produce a short "file-drawer" report (1-2 pages) and put it somewhere visible (see last point).

Entry points: Ioannidis (2005); Nosek, Spies & Motyl (2012)

J) Be honest. When you are writing the closing report, go back to your pre-registration. Have you followed your established course of action? When you have deviated, why was it? You should believe more your results from the pre-defined analyses than from any other analyses that were unplanned, and therefore exploratory (and subject to false positive inflation). Remember, you are now laying a new brick, and others may use it as *their* foundation.

Entry points: Simmons et al. (2011); Cumming (2013)

K) Make your work visible to the community. How will those coming after you be able to profit from your findings? Ideally, you should plan on posting all files necessary to run the experiment, the data (in as original a format as possible and/or an easier to parse format), and analyses scripts. You might also want to add documentation of earlier stages, additional analyses, etc. There are a number of open science repositories, which vary in terms of the kinds of materials accepted, the structure provided within the project, and the communities using them.

- Your field might already have a vibrant one. Ask your advisor and lab colleagues.
- Otherwise, you can try out dataverse.org (pros: strong in the economics community, online analyses tools) and osf.io (pros: very flexible structure, easy to use).
- Finally, you might be able to capitalize on your literature review by turning it into a "community-augmented meta-analysis;" instructions for this (and good readings on meta-analyses) can be found on cama.acristia.org.
- At the very least, you could post your materials on a website; if your advisor cannot host it, google sites are easy to use and free.

Annotated bibliography

Cumming, J. (2013). The new statistics: Why and how. *Psychological Science, 25*, 7-29.

Proposes 25 rules for new statistical practices, which go from relatively general guidelines to promote knowledge accumulation, to specific instructions to displace "null hypothesis significance testing".

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Provides a logical demonstration that "most published research findings are false", and clear arguments as to why we need to worry about bias at all levels of research, from conception to publication (the latter, related to the file-drawer problem).

Lakens, D., & Evers, E. R. (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*(3), 278-292.

Discusses a number of strategies researchers can use to decide on sample sizes.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 683.

Remarkably clear instructions for calculating effect sizes in the kinds of designs psychologists use (including from published research).

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology, 45*(3):137–141.

Explains the advantages and limitations of peer-reviewed pre-registered reports.

Nosek, B. A., Spies, J. A., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science, 7*, 615.

Bias, file-drawer studies, and other unfortunate practices are entirely expected given the current structure of reward (good researchers are those that publish a lot of really novel things); why it would behoove the field to change this structure for one that is more aligned with the overarching goal of scientific endeavor (finding out the truth); and how this might be accomplished.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9*(3), 293-304.

Proposes a number of strategies researchers can follow to carry out preliminary analyses without inflating the false positive rate.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Introduces the concept of "experimenter degrees of freedom" and clearly illustrate the dangers of multiple testing and preliminary analyses (followed by changes in key methodological decisions, such as participant sample size).