Lecture 05: The Open Science movement in Psychology

Doing better

Dr. Gordon Wright

October 30, 2023

## Key topics today

The week ahead (week 5)

* Personal Tutor Meeting about essay writing - bring your questions
* Design & Analysis Quiz due this week (week 5)
* Open Science
* Labs - Critical Proposal and Power Calculations

## Personal Tutor Meeting Week 5

This week (week 5) your PT session is all about essay writing

|  |  |
| --- | --- |
|  | Some of you have expressed doubts about this. Please see this as an opportunity to get answers to any questions.  Make sure to use your feedback! |

# Any Questions?

## So…

## But what does that mean?

## Open Science Collaboration (2015)



[Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–943. http://www.jstor.org/stable/24749235](https://www-jstor-org.gold.idm.oclc.org/stable/24749235?sid=primo#metadata_info_tab_contents)

## The replication crisis

* The Open Science Collaboration (2015) (c.f. Brian Nosek) conducted 100 replications of psychology studies published in three psychology journals
* While 97 of previous studies reported significant results, only 36 were significant in the replication attempt. And effects were smaller than originally reported…

## Violin plots

|  |
| --- |
| Violin Plots of Replication Results |

## Raincloud plots

|  |
| --- |
| Raincloud Plots of Replication Results |

## Why aren’t we replicating?

* Some point the finger at scientific fraud (i.e. bad scientists making up their data)
* However, others point to more systematic problems
* Low statistical power
* Questionable research practices (QRPs)
* Publication bias

## Statistical power

* Since 1960s, sample sizes in standard psychology studies have remained too small – giving them low power
* Low power is normally a problem because it means that you don’t find significant effects
* An underappreciated downside of low power is that if you do find effect, it is probably spuriously exaggerated
* This will mean that when you try to replicate it, it will be smaller (not significant)

## We are training you in best practice

If you have had trouble finding an effect size in your Personality Essay or Critical Proposal…

This is either because the new best practice hasn’t been adopted, or the research team dropped the ball.

## Power plot



Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. <https://doi.org/10.1098/rsos.160384>

# Power and Power Calculations in Psychology

## **What is power?**

* Power is the probability of rejecting the null hypothesis when it is false.
* Power depends on the significance level, sample size, and effect size of a test.
* Power is important for planning and evaluating studies.

## **How to calculate power?**

* Use online tools or statistical software like G\*Power.
* Specify the type of test, the alpha level, the effect size, and the desired power or sample size.
* For complex research designs, you may need to calculate a number of potential effect sizes

## **Why is power low in psychology?**

* Small sample sizes are common in psychological research.
* Effect sizes are often unknown or overestimated.
* Researchers may not use power analysis or understand its meaning.

## **How to improve power in psychology?**

* Increase sample size or use more sensitive measures.
* Use meta-analysis or replication to estimate effect sizes.
* Educate researchers and reviewers about power and its implications.

## Questionable Research Practices (QRPs)

**Selective reporting of participants**

E.g., excluding data from some participants

. . .

**Selective reporting of manipulations or variables**

E.g., measuring many different variables in a study, but only writing up the variables that ‘worked’ (were significant)

. . .

**Optional stopping rules**

E.g., continuing to add participants to a sample until it is just significant (p<.05)

## QRPs Continued

**Flexible data analysis**

E.g., Adding covariates (without good reason) to ‘improve’ statistical results

. . .

**HARKing (Hypothesising After Results are Known)**

Running a study, and then generating a hypothesis that fits the results (even if they were not what you originally predicted)

. . .

What these practices all have in common is they involve capitalising on chance to create a significant result (which may not be reliable)

## Novelty and glamour

* Scientists want to communicate their science, but they also want successful careers
* An important metric for success in science is publishing in ‘top journals’ (e.g., Nature, Science)
* Getting published in these journals gets your science out to a wide audience (because lots of people read them) but also carries prestige – you get jobs, grants, funding and prizes from publishing regularly in these journals
* But top journals want to publish novel or surprising results.
* Why do you think that could be a problem?

## Lust for Impact Factors!



[Paulus, F. M., Rademacher, L., Schäfer, T. A., Müller-Pinzler, L., & Krach, S. (2015). Journal Impact Factor Shapes Scientists’ Reward Signal in the Prospect of Publication. *PloS one*, *10*(11), e0142537. https://doi.org/10.1371/journal.pone.0142537](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4640843/)

## Biases in journals: File drawer problem

* Even beyond ‘prestige’ journals, journals are biased to publish positive (i.e. significant) findings
* Because it is much easier to publish positive results, rather than nonsignificant results or failed replications, science has a ‘file drawer problem’
* Scientists don’t try to publish their null results, and/or journals make it hard to publish them
* This means the published literature is biased to contain significant results (that come from a distribution where there is no true effect)

## Let’s work the probabilities

With an alpha level of p=.05, if we have 40 scientists testing any hypothesis we would expect one to find a significant result in one direction, and another to find a significant result in another direction just by random chance



## The credibility revolution?

Recent years have seen several changes to how psychological science is conducted to overcome concerns about reliability – dubbed the ‘credibility revolution’



## Recommendations and changes

Low statistical power? Report power analyses and justify sample sizes



(Taken from guidance to authors at journal Psychological Science)

## Familiar?



## The goal



## The ‘normal’ process



## A better solution?



## Do scientists already ‘know’ which results to trust?

* The unnerving thing about the ‘replication crisis’ seems to be that psychological theories are built on foundations of sand. But is this true?
* Camerer and colleagues attempted to replicate 21 social science studies (including psychology) and found around 13 replicated.
* However, the study also ran a prediction market where scientists (PhD or PhD student) had to bet on which studies would replicate and which wouldn’t
* We should want our journal to publish things that are robust – but if scientists have a good sense of what is reliable, is this really a ‘crisis’?

## Camerer et al. (2018)



## Findings



## Dubious efforts to replicate

Researchers who do replication studies also have flexibility in their design and analysis choices.

There may be a bias to not replicate certain findings (e.g., because you are sceptical of the result in the first place)



## No reason to worry

Some have suggested that low replication rates are not necessarily a sign of bad research

Alexander Bird (philosopher of science) suggests worries about replication reflect base rate fallacy

Most hypotheses are wrong so we wouldn’t expect them to replicate in future studies

What do you think?

## Alexander Bird (2018)



## Are we worry about the wrong thing?

* Other psychologists have argued that focus on replicability, statistical robustness etc. is misguided
* The real problem psychology has is the absence of strong theories
* This “theory crisis” cannot be solved with more and more attention to statistics
* Theory is the thing we should be caring about? Not specific effects in specific studies
* No statistics can help us to test a theory that is poorly thought out

## Summary

You should now know:

* Why scientists are concerned about the reliability of psychological studies
* Steps the scientific community are taking to overcome these worries
* Not everyone is convinced that the ‘crisis’ is as serious as it seems, or whether these changes will help solve psychology’s problems

# Questions?

|  |
| --- |
| **Lab activities**  Power Calculations in preparation for your Ethics Applications  Pay close attention to the lab slides - Step by Step guidance for EVERY ANOVA flavour  Access to detailed Ethics VLE page and resources (Please Review)  Lab Priority is the Critical Proposal Target Paper approval/confirmation |

## References

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. <https://doi.org/10.1126/science.aac4716>