

2017-01-09-course-intro

Rick O. Gilmore

2017-02-08 10:34:59

Prelude

<http://ed.ted.com/lessons/is-there-a-reproducibility-crisis-in-science-matt-anticole>

Today's Topics

- Introduction to the course
- Scientific misconduct
- Is there a crisis of reproducibility?

Introduction to the course

Goals

- What are transparent, open, and reproducible science practices?
- Why are they important?
- How can one implement them?

Topics

- Tour of course website: <http://psu-psychology.github.io/psy-511-reproducible-research-spring-2017/>

Course structure

- 40 min lecture
- 10 min break
- ~ 2 hrs discussion/hands-on activities

Evaluation

Component	Points	% of Grade
Class participation	2 pts/class * 15 weeks = 30	25
Assignments	5 pts * 12 assignments = 60	50
Final projects	30	25
TOTAL	150	100

Tools we'll use

- R via RStudio
- The web
- git via GitHub

- Jupyter
- Slack, <http://psy511.slack.com> for communication, sharing, Q&A

Scientific misconduct

The Hauser and Stapel cases

- Marc Hauser
 - <http://www.sciencemag.org/news/2014/05/harvard-misconduct-investigation-psychologist-released>
- Diederik Stapel
 - Commentary

Marc Hauser

- Evolutionary/Comparative Psychologist, Professor at Harvard
- Resigned 2011 after internal investigation found him responsible for research misconduct

2012 report by NIH Office of Research Integrity (ORI)

- Fabricated data in Figure 2 of (M. D. Hauser, Weiss, and Marcus 2002); paper retracted.
- Falsified coding in unpublished data.
- Falsely described methodology in manuscript submitted to Cognition, Science, and Nature. Corrected manuscript prior to publication as (Saffran et al. 2008)
- Falsely reported results and methodology for one study in (M. D. Hauser, Glynn, and Wood 2007)

2012 report by NIH Office of Research Integrity (ORI)

- Made false statement in Methodology section of (J. N. Wood et al. 2007)
- Engaged in research misconduct by providing inconsistent coding of unpublished data

“Respondent neither admits nor denies committing research misconduct but accepts ORI has found evidence of research misconduct as set forth above and has entered into a Voluntary Settlement Agreement to resolve this matter. The settlement is not an admission of liability on the part of the Respondent.” <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-12-149.html>

Hauser’s response

“...Although I have fundamental differences with some of the findings in the ORI report, I acknowledge that I made mistakes. I tried to do too much, teaching courses, running a large lab of students, sitting on several editorial boards, directing the Mind, Brain & Behavior Program at Harvard, conducting multiple research collaborations, and writing for the general public. I let important details get away from my control, and as head of the lab, I take responsibility for all errors made within the lab, whether or not I was directly involved...”

“I am saddened that this investigation has caused some to question all of my work, rather than the few papers and unpublished studies in question. Before, during and after the investigation, many of my lab’s research findings were replicated by independent researchers...”

Diederik Stapel

- Dean of the School of Social and Behavioral Sciences at Tilburg University
- teacher of Scientific Ethics course
- Fraud investigation launched when 3 grad students noticed anomalies – duplicate entries in data tables
- Stapel confessed, lost position, gave up Ph.D., wrote a book

Flawed science: The fraudulent research practices of social psychologist Diederik Stapel

- *“...found to have committed a serious infringement of scientific integrity by using fictitious data in his publications, while presenting the data as the output of empirical research.”*
- Previous suspicions/concerns about data “too good to be true” were ignored.
- Fraud established in 55 publications.

How did it happen?

- *“admitted to having committed fraud in much of his research. He has also conceded that he has been sloppy, kept inadequate records and presented data in the best light possible”*
- *“...No explicit attention was paid to the combating or prevention of scientific misconduct within the scientific research environment in the Netherlands, and there was certainly no formal Code for research integrity...”*
- *“...Whatever measures existed were more implicit, and any action was at the discretion of individual researchers....”*

Flawed science: The fraudulent research practices of social psychologist Diederik Stapel

How did it happen?

- *“...no culture of transparency, information sharing, peer review and joint responsibility in this Faculty. It was easy for researchers to go their own way...”*
- *“...[researchers] did not share their data with other members of the various research groups, and there was no peer review...”*
- *“People trusted each other’s scientific integrity: no feedback, no warnings, no correction.”*

Flawed science: The fraudulent research practices of social psychologist Diederik Stapel

Recommendations

- Avoid verification biases (pp. 48-50)
- Provide complete information about methodology, sufficient to permit replication
- Avoid statistical errors
- If it’s “too good to be true”, it probably isn’t.
- Journals should avoid bias toward *“elegant, concise, attractive findings”*
- *“The data on which an experimental psychology doctoral dissertation is based must as a rule be collected and analysed by the PhD students themselves.”*

Is there a crisis of reproducibility?

The Reproducibility Project

- Reproducibility Project: Psychology
- (Collaboration 2015)
- Center for Open Science (COS), produces Open Science Framework (OSF)

(Collaboration 2015)

- 39/98 (39.7%) replication attempts were successful
- 97% of original studies reported statistically significant results vs. 36% of replications

So, did the studies replicate?

- (Gilbert et al. 2016)
 - Sampling error differences predicts < 100% reproducibility
 - Samples !=
 - Only 69% of original PIs “endorsed” replication protocol. Replication rate 4x higher (59.7% vs. 15.4%) in studies with endorsed protocol.
 - CI of *expected* effect sizes given sample/methodological variability? Many Labs project
-
- (Collaboration 2015) “...*seriously underestimated reproducibility of psychological science.*”

Issues

- Kudos to (Collaboration 2015) and (Gilbert et al. 2016) for addressing these issues openly
 - Data from (Collaboration 2015)
 - Data from (Gilbert et al. 2016)

Issues

- Reproducibility of “psychological science” vs. a specific finding
- What is the *true* effect size of a particular manipulation?
- Domain-specific differences in/challenges to reproducibility
- Possible confusion about types of reproducibility

Examples of differences that affect the approach to reproducibility in distinct scientific domains

- (Goodman, Fanelli, and Ioannidis 2016)
- Degree of determinism
- Signal to measurement-error ratio
- Complexity of designs and measurement tools
- Closeness of fit between hypothesis and experimental design or data
- Statistical or analytic methods to test hypotheses

Examples of differences that affect the approach to reproducibility in distinct scientific domains

- Typical heterogeneity of experimental results
- Culture of replication, transparency, and cumulating knowledge
- Statistical criteria for truth claims
- Purposes to which findings will be put and consequences of false conclusions
- (Goodman, Fanelli, and Ioannidis 2016)

What does research reproducibility mean?

- Methods
- Results
- Inferential

(Goodman, Fanelli, and Ioannidis 2016)

What does research reproducibility mean?

- Methods reproducibility
 - “...*the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results.*”

(Goodman, Fanelli, and Ioannidis 2016)

What does research reproducibility mean?

- Results reproducibility
 - “(*previously described as replicability*) refers to obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible.”

(Goodman, Fanelli, and Ioannidis 2016)

What does research reproducibility mean?

- Inferential reproducibility
 - “...*refers to the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study*”

(Goodman, Fanelli, and Ioannidis 2016)

Steps toward improving transparency and openness

- Transparency and openness promotion (TOP) guidelines in publishing
 - (Nosek et al. 2015)

TOP guidelines

- Citation
- Data transparency
- Analytic methods (code) transparency
- Design and analysis transparency

- Preregistration of studies
- Preregistration of analysis plans
- Replication

Who's signed on and who hasn't?

- List of signatories
- Declines
 - (Lash 2015)
 - Implementation would run counter to efforts to “...maintain an editorial policy that encourages creativity and novelty, resists regimentation of research practices to the extent practicable, and invites challenges to current scientific habits and conventions through innovation in epidemiologic theory and practice.”

On the other hand...

- New editorial from European Journal of Personality
 - <https://twitter.com/i/web/status/818460818102697985>

Barriers to greater openness and transparency

- Technical
- Cultural
 - Old habits
 - Fear of embarrassment, retaliation, effects on career
- Economic
 - Intellectual property

Other concerns to consider

- File drawer effect
- Pro-novelty/anti-replication bias in publication
- Pro-effects bias in publication
- Pro-counterintuitive effects bias in publication in high profile/impact journals

Is psychological/neural science sufficiently powered?

- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLOS Medicine, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Szucs, D., & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. bioRxiv, 071530. <https://doi.org/10.1101/071530>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>

(Szucs and Ioannidis 2016)

“We have empirically assessed the distribution of published effect sizes and estimated power by extracting more than 100,000 statistical records from about 10,000 cognitive neuroscience and psychology papers published during the past 5 years. The reported median effect size was $d=0.93$ (inter-quartile range: 0.64-1.46) for nominally statistically significant results and $d=0.24$ (0.11-0.42) for non-significant results. Median power to detect small, medium and large effects was 0.12, 0.44 and 0.73, reflecting no improvement through the past half-century. Power was lowest for cognitive neuroscience journals. 14% of papers reported some statistically significant results, although the respective F statistic and degrees of freedom proved that these were non-significant; p value errors positively correlated with journal impact factors. False report probability is likely to exceed 50% for the whole literature. In light of our findings the recently reported low replication success in psychology is realistic and worse performance may be expected for cognitive neuroscience.”

References

- Collaboration, Open Science. 2015. “Estimating the Reproducibility of Psychological.” *Science* 349 (6251): aac4716. doi:10.1126/science.aac4716.
- Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson. 2016. “Comment on ‘Estimating the Reproducibility of Psychological Science’.” *Science* 351 (6277): 1037–7. doi:10.1126/science.aad7243.
- Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. “What Does Research Reproducibility Mean?” *Science Translational Medicine* 8 (341): 341ps12–341ps12. doi:10.1126/scitranslmed.aaf5027.
- Hauser, Marc D., David Glynn, and Justin Wood. 2007. “Rhesus Monkeys Correctly Read the Goal-Relevant Gestures of a Human Agent.” *Proceedings of the Royal Society of London B: Biological Sciences* 274 (1620): 1913–8. doi:10.1098/rspb.2007.0586.
- Hauser, Marc D., Daniel Weiss, and Gary Marcus. 2002. “RETRACTED: Rule Learning by Cotton-Top Tamarins.” *Cognition* 86 (1): B15–B22. doi:10.1016/S0010-0277(02)00139-7.
- Lash, Timothy L. 2015. “Declining the Transparency and Openness Promotion Guidelines.” *Epidemiology* 26 (6). LWW: 779–80. http://journals.lww.com/epidem/Fulltext/2015/11000/Declining_the_Transparency_and_Openness_Promotion.1.aspx.
- Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. “Promoting an Open Research Culture.” *Science* 348 (6242): 1422–5. doi:10.1126/science.aab2374.
- Saffran, Jenny, Marc Hauser, Rebecca Seibel, Joshua Kapfhamer, Fritz Tsao, and Fiery Cushman. 2008. “Grammatical Pattern Learning by Human Infants and Cotton-Top Tamarin Monkeys.” *Cognition* 107 (2): 479–500. doi:10.1016/j.cognition.2007.10.010.
- Szucs, Denes, and John PA Ioannidis. 2016. “Empirical Assessment of Published Effect Sizes and Power in the Recent Cognitive Neuroscience and Psychology Literature.” *BioRxiv*, August, 071530. doi:10.1101/071530.
- Wood, Justin N., David D. Glynn, Brenda C. Phillips, and Marc D. Hauser. 2007. “The Perception of Rational, Goal-Directed Action in Nonhuman Primates.” *Science* 317 (5843): 1402–5. doi:10.1126/science.1144663.