*Instructor Syllabus*

Estimating the credibility

of past research

Most readers of published scientific articles make reasonable assumptions: that the descriptions of the studies are largely accurate, that the studies they report actually occurred, that the studies were conducted as described, that the results are reported accurately, and their conclusions are likely to be valid. Unfortunately, there is good reason to be skeptical, and there is growing recognition that the credibility and trustworthiness must be proved rather than assumed.

In the 1980s, after a series of academic fraud scandals, an editorial in Science reassured readers that “99.9999 percent of reports are accurate and truthful” (Koshland, 1987). Twenty years later, some were no longer as confident, as Ioannidis (2005) presented mathematical arguments to support his belief that “most published research findings are false.” Much meta-science research in since then has supported this provocative idea that an undesirably large proportion of scientific claims have lower credibility and trustworthiness than previously believed.

Consumers of scientific publications - including other scientists, students, practitioners, policy makers, and members of the public - are faced with the difficult questions of what to believe and how to determine this. This seminar course will equip you with the practical and technical skills to estimate the credibility and trustworthiness of published research findings.

# BASIC INFORMATION

## Type

Seminar

## Level

Master

## ECTS

5

## Mode

Weekly meetings

## Learning Goals

* Acquire practical skills to assess the credibility and trustworthiness of published research
* Recognize the importance of transparency and scrutiny in science
* Identify the incentive structures that determine scientists behavior.

## Grading

Weekly at-home assignments

# SCHEDULE

|  |  |
| --- | --- |
| **Week** | **Topic** |
| **1** | Introduction and organization |
| **2** |  |
| **3** |  |
| **4** |  |
| **5** |  |
| **6** |  |
| **7** |  |
| **8** |  |
| **9** |  |
| **10** |  |
| **11** |  |
| **12** |  |
| **13** |  |
| **14** |  |

# WEEK 1: Introduction and organization

## Readings

Bernal, J. D. (1939). *The social function of science*.

Alexander, R. (2013) Reinhart, Rogoff... and Herndon: The student who caught out the profs. BBC news.

Anon (2023) There is a worrying amount of fraud in medical research. The Economist, Simply Science newsletter.

Smith, R. (2021) Time to assume that health research is fraudulent until proven otherwise? thebmj.com

Van Noorden, R. (2023) How many clinical trials can’t be trusted? Nature Features.

## Assignment

To be completed **after** this class.

*Answer the following question in 2000 to 5000 characters:*

Why do scientists do science? We are often told that we do science to discover truth. If this is the only reason errors and fraud would be rare, but there is growing concern they are not (Van

Noreen, 2023; Simply Science, 2023; Smith, 2021). Bernal (1939) describes other reasons why scientists do science. Which of these reasons have been emphasized in your own studies so far, and which have not? Do you agree with Bernal?

# WEEK 2:

## Preparation Assignment

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*(11), 1359-1366. <https://dx.doi.org/10.1177/0956797611417632>

**1.** Read Simmons et al. (2011).

**2.** Simmons et al. discuss the problem of peeking & optional stopping, i.e. the practice of starting with a smaller sample and deciding whether or not to collect more data depending on a significance test. They claim that this dramatically increases the false positive rate. But why, actually? Imagine two identical studies (between-subjects experiment with 2 conditions and one DV). In study A, data is collected from 50 subjects and then a significance test is conducted for differences between conditions. In Study B, data are also collected from 50 people, but the significance test is performed after every 5 people (so 10 tests in total). Why is the false positive probability in study B higher than in study A, even though the same amount of data is collected?

# WEEK 3: Science’s Fraud Problem

## Preparation Assignment

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8*(4), 363–369. <https://doi.org/10.1177/1948550616673876>

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203-210. <https://doi.org/10.1037/h0041593>

\* Guéguen, N., Meineri, S., & Fischer-Lokou, J. (2014). Men’s music ability and attractiveness to women in a real-life courtship context. *Psychology of Music, 42*(4), 545–549. <https://doi.org/10.1177/0305735613482025> (\* *retracted*)

Servick, K. (Sep 21, 2018). Cornell nutrition scientist resigns after retractions and research misconduct finding. *ScienceInsider.* <https://www.science.org/content/article/cornell-nutrition-scientist-resigns-after-retractions-and-research-misconduct-finding>

**1.** Read the papers Brown & Heathers, Festinger & Carlsmith, Gueguen et al., and Servick.

**2.** Run the GRIM test described in Brown & Heathers on all suitable means reported in Festinger & Carlsmith using this tool: <http://nickbrown.fr/GRIM/GRIM.html>

**3.** Report, what you did and the results of this test. If you found inconsistencies, try explaining what might have caused them.

# WEEK 4: Scientific Utopia: Promoting Truth over Publishability

## Preparation Assignment

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615-631. <https://doi.org/10.1177/1745691612459058>

\* Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109*(38), 15197-15200. <https://doi.org/10.1073/pnas.1209746109> *(\* retracted)*

**1.** Read Nosek et al. and Shu et al.

**2.** Nosek et al. look to promote “truth over publishability”. Explain what this has to do with honesty, as characterised by Shu et al.

**3.** How could the empirically tested techniques described across three experiments in Shu et al. be implemented in science to achieve (some of) the goals laid out in Nosek et al.?

# WEEK 6: Publication Bias and the File-drawer problem

## In-Class Materials

Assign students to groups. Give each group three bowls (white, black, and grey), and two types of marbles (red ones, and blue ones. Of the blue ones, a few are bicolored, with extra green). Instructions:

1. Each marble represents one study testing the same hypothesis.
2. Red marbles are nulls, blue marbles are positives.
3. The black bowl represents all studies conducted in a field, the white bowl the published literature.
4. Each group sets the size of their field (k), i.e. the total number of studies conducted, between 10 and 30.
5. Each group sets a true discovery rate for their field (P), i.e. the proportion of positive studies in k, between 5% and 95%, and writes it down in secret. This represents the “true effect size”.
6. They put k marbles in the black bowl of which P% is blue. For example, in a field of k = 20 studies and discovery rate of P = 80%, 16 marbles are blue und 4 marbles are red.
7. The group sets a publication bias rate (the chance of null studies to get published) on a scale from 5% to 95% in steps of 5%.
8. They play publish or perish: One student randomly grabs a marble from the black bowl. If it’s blue then it gets put in the white bowl, if it’s red another student rolls a d20. If the roll x 5 is equal or lower than the publication bias rate, the red marble gets added to the white bowl. If it’s higher, it gets added to the grey bowl. For example, if the publication bias rate is 30%, rolls from 1 to 6 would mean a red marble gets added to the white bowl, whereas rolls of 7+ mean it gets added to the grey bowl.
9. Repeat until the black bowl is empty.
10. The groups now do meta-analyses on each other’s fields and try to estimate the true effect size. They exchange white bowls and have to guess the proportion of red to blue marbles that were in the black bowl.
11. Reveal that the blue with sprinkles of green represent p-hacked studies and should be treated as red marbles.

# WEEK 7: Measurement Schmeasurement, negligence, and error

## Preparation Assignment

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. <https://doi.org/10.1177/2515245920952393>

Xu, I., Passell, E., Strong, R. W., Grinspoon, L., Wu, C. Y., Jung, L., … Germine, L. (2022). No evidence for consistent reliability across 36 variations of the emotional dot probe task in 9,000 participants [Preprint]. *PsyArXiv.* <https://doi.org/10.31234/osf.io/58z4n>

**1.** Read the papers by Flake and Fried and by Xu et al.

**2.** How is it possible that the results from Xu et al. suggest that the reliability of the Dot Probe Task is close to zero, but at the same time there are dozens of studies in which the task was used and is described as a useful measure of attentional bias? Provide one optimistic and one pessimistic explanation.

# WEEK 8: Science Communication and Publicity

## Preparation Assignment

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., ... & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ, 349*, g7015. <https://doi.org/10.1136/bmj.g7015>

Young, C., Majolo, B., Heistermann, M., Schülke, O., & Ostner, J. (2014). Responses to social and environmental stress are attenuated by strong male bonds in wild macaques. *Proceedings of the National Academy of Sciences, 111*(51), 18195-18200. <https://doi.org/10.1073/pnas.1411450111>

1. Read Sumner et al. and Young et al.
2. Imagine working in University of Göttingen‘s press office. You learn of the publication of Young et al. and think that this could be a nice study for a press release. Write a press release of 400-500 words in which you commit as many of the mistakes and problematic practices identified by Sumner et al.
3. Explain what in your own press release might be problematic. Be specific (don’t just say “it’s exaggerated”).

# WEEK 9: Peer Review and Quality Management in Science

## Preparation Assignment

Rennie, D. (2003). Editorial peer review: its development and rationale. In F. Godlee & T. Jefferson (Eds.), *Peer Review in Health Sciences* (2nd Ed.)*,* 1-13. BMJ Books.

Students are randomly assigned to read one of three additional papers:

**A** | Elson, M., Huff, M., & Utz, S. (2020). Metascience on peer review: Testing the effects of a study’s originality and statistical significance in a field experiment. *Advances in Methods and Practices in Psychological Science, 3*(1), 53-65. <https://doi.org/10.1177/2515245919895419>

**1.** Read Rennie (2003) and Elson et al. (2020).

**2.** What are arguments in favour of doing field experiments on peer review?

**3.** What are ethical reasons for not conducting this type of research, and do they outweigh the benefits?

**B** | Bornmann, L., Mutz, R., & Daniel, H. D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS One, 5*(12), e14331. <https://doi.org/10.1371/journal.pone.0014331>

**1.** Read the paper by Rennie (2003) and Bornmann et al. (2010).

**2.** What implications does the low interrater reliability of peer reviewers have for quality management in science?

**3.** Are there good reasons why the interrater reliability is low or why it might even be desirable?

**C** | Bedeian, A. G. (2003). The manuscript review process: The proper roles of authors, referees, and editors. *Journal of Management Inquiry, 12*(4), 331-338. <https://doi.org/10.1177/1056492603258974>

**1.** Read the papers by Rennie (2003) and Bedeian (2003).

**2.** On the basis of the results by Bedeian, what is the best way to characterise the relationship between authors and reviewers?

**3.** Name 2 important methodological weaknesses of the study by Bedeian, and how they affect the interpretability of the results.