Detecting Bad Science: Reviewing and Improving Social Science Research

René Bekkers, 09 October 2024

This intensive course will make you a *Bad Science Detective* with a good purpose: to improve social science research. In case you missed it: science is in a credibility crisis, and this diagnosis also holds for the social and behavioral sciences. At least half of researchers in the social and behavioral sciences in the Netherlands admit to have engaged in questionable research practices (Gopalakrishna et al., 2022a). Yes – for every pair of scholars you randomly choose, one is engaging in bad science. For every sixteen researchers you choose, one has even committed fraud or fabricated data. Though the number of retractions by academic journals for fraud, fabrication, plagiarism and other integrity violations is rising, most bad science is still undetected. The quality control system that peer review is commonly believed to be is a very lax one, and easy to fool (Smith, 2006). As a result, it should be no surprise that half of all studies do not replicate, and published effects are only half the original size upon replication (Open Science Collaboration, 2015). In sum: you cannot trust research to be valid and reliable, even when it is peer reviewed and published in the most prestigious journals.

How then can you tell the difference between good and bad science? What signals tell you something about the quality of research? As a bad science detective, you'll be able to call bullshit on the texts that your professors require you to read – including their own work. At the same time, we will collectively improve the chances that bad science is identified. With a higher discovery rate of bad science, researchers will be more careful and the quality of research will improve (Gopalakrishna et al., 2022b). In addition, by identifying the weaknesses in the work of others, you learn in which aspects you can improve your own research.

Audience

This course is developed for students in research master programs and PhD candidates in the social and behavioral sciences broadly conceived, including psychology, neuroscience, data science, computational social science, sociology, political science, public administration, organization science, social geography, epidemiology, human health and life sciences, economics, marketing, management and business administration. You will find this course useful if you seek to uncover regularities or test hypotheses using empirical data on human cognition and behavior.

Entry requirements

You can enroll in this course if you are a PhD candidate or a student in a research master program or equivalent (e.g., advanced postgraduate research program, master program at a research-oriented institution of 120 ECTS).

Support

Please address practical questions about the Winter School to graduatewinterschool@vu.nl.

Learning objectives

At the end of this course, you will be able to use analytical tools and software to identify the weaknesses of research and evaluate the quality of research in the social and behavioral sciences.

The primary analytical tool is iQUESST – identifying Questionable Social Science through Transparency. The iQUESST acronym refers to the evaluation of research quality with respect to the:

i information on
QU the Question that the research answers: how informative would potential answers to the research question be for practice and theories?
using an
E Estimation method: does it provide an answer to the question, and is it the best choice?
S Sample: how useful is it to make inferences about the target population?
S Stringent design: are the data and methods the best possible stress test of the research claims?
T through Transparency of the research: what does the research report tell you about the data and methods used to produce the results?

The secondary analytical tool is the four validities framework (Vazire et al., 2022), to which iQUESST roughly corresponds as follows:

1. Construct validity ≈ *QU*estion: poorly defined and badly operationalized constructs, ill-documented measures, and hypothesizing after results are known;

2. Internal validity ≈ Estimation: selective attrition, non-causal mediation, lack of random assignment, reverse causality, incorrect chronology, omitted variable bias;

3. External validity ≈ Sample: constraints on generality due to survivorship bias, selection bias, biased samples, or selective response;

4. Statistical conclusion validity ≈ Stringency: problems with outliers, missing values, model misspecification, false assumptions, p-hacking, researcher degrees of freedom, the garden of forking paths, or low power.

The website https://detectingbadscience.wordpress.com/ describes 36 potential flaws in research reports. For each flaw you can find a description, a strategy to identify the flaw, and insights from meta science research, and a solution. For each form of bad science there's also a description of its good science counterpart.

Learning activities and skills
In this course, you will read research reports individually, give one plenary presentation, engage in group discussion in four workshops, and individually write a critical review. Through these activities, you learn to develop an eagle eye for weaknesses in research reports, use tools to identify statistical anomalies such as Statcheck (Epskamp & Nuijten, 2016) and Papercheck (DeBruine & Lakens, 2024), improve your knowledge about research design, data analysis and methodology, discover effective ways to repair and prevent weaknesses, and effectively present your insights.

You can make good use of the eagle eye for weaknesses not only to improve your own research, but also to provide suggestions for the work of others as a peer reviewer. In science, a good peer review is not merely critical: it is also constructive (Bekkers, 2020). Therefore, you will also be able to suggest improvements in the validity of research in the process of peer review. You will know the principles of reproducible science (Munafo et al., 2017).

Forms of tuition
The course consists of four meetings, on two days per week: Mondays and Thursdays. Course meetings take 4 hours, and are scheduled in the afternoon, from 13.00 to 17.00. Each hour is 50 minutes of class time, followed by a 10 minute break.

Meeting 1, Monday 6 January. We start by getting to know each other, reviewing the course design and activities, and discussing the credibility crisis. We get started with nominations for target papers, and create a schedule for the presentations in the second and third meeting.

Meeting 2, Thursday 9 January. The second day is a workshop in which we discuss the first half of the target papers.

Meeting 3, Monday 13 January. On the third day we discuss the second half of the target papers.

Meeting 4, Thursday 16 January. We discuss improvements of the research designs and analyses of the target papers.

Course credits and work load

Completing the course work successfully is worth 2 credits in the European Credit Transfer System (ECTS). This corresponds to a work load of 56 hours, distributed as follows:

Reading required articles, selecting and reading target papers	6
Data analysis	20
Writing	12
Preparation for presentation	2
Course meetings	16
Total	56

Preparation for course meetings

For the first meeting, you study the syllabus, the readings below, and complete assignment 1:

- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162-168. https://doi.org/10.1177/09637214211067779

For the second and third meeting, you complete assignment 2, identifying weaknesses in a paper, and you present them in class. In the meeting, you comment on the presentations of other participants and receive their comments in a group discussion. After the discussion in class, you revise the assessment based on the feedback and discussion of group participants.

For the fourth meeting, you complete assignment 3, studying the weaknesses of the target paper of another participant, and suggesting improvements for the group discussion.

Assignments

1. Describe how your bachelor or master program training taught you to think about the quality of research. Which aspects of research have you learned indicate that the research is of high quality, and which aspects indicate low quality? Reflect on these criteria referring to the readings for meeting 1. Which aspects of research that you learned about in your training do not necessarily indicate research quality, and by which criteria should they be replaced?

2. Select a paper from the list of target papers compiled in the first meeting, and read it thoroughly on your own.

a. Identify all possible weaknesses of the paper, starting from iQUESST and the examples of weaknesses given in the first meeting. Which study limitations did the authors note themselves? Which weaknesses does the paper have that the authors themselves did not describe?
b. Search for replications or commentaries on the target paper in papers that cite the target paper according to Google Scholar. What was the result of the replication? Which of the study weaknesses may have contributed to the replication showing different results than the original result? Which weaknesses do studies citing the target paper identify?
c. Summarize the study weaknesses you have identified in a five minute presentation. List all weaknesses you have identified, and explain one of them in detail so that participants who have not read the target paper can understand.

3. Study the target paper for another course participant, and suggest improvements to the weaknesses you find. Did you find additional weaknesses? Explain to what extent and how the improvements can effectively repair the weaknesses you identified.

4. Write a short report on the weaknesses of your target paper, and suggest improvements. Explain how the improvements effectively repair the weaknesses. If, in your view, repairing the problems is not possible and the paper is a total loss, explain why.

Assessment
You successfully complete the course if you've participated in the course meetings, demonstrated the ability to identify weaknesses in research reports discussed during the meetings and in the assignments, and submitted a sufficiently detailed and constructively critical review of a target paper. The review is sufficiently detailed if you can describe the quality of the research using iQUESST. You receive extra praise if you detect statistical anomalies, incorrect interpretations, plagiarism, data fabrication, or undisclosed deviations from preregistrations. We use the four eyes principle: you read the review composed by another participant and check if you understand the report (Bekkers, 2021).

Transparency of the Use of Artificial Intelligence
In your assignments, include an AI Tools section, describing which generative artificial intelligence tools you have used in producing the contents such as ChatGPT, Bing, Claude, Copilot, Elicit, Gemini, HuggingChat, NotebookLM, Perplexity, or ResearchRabbit. You are allowed to use such tools, as long as you identify that you have used them, and how you have used them. Do so in sufficient detail for others to be able to reproduce your findings. This means that you specify the software version, settings, date of usage, the prompts and commands, and output with a URL or a screendump. You do not have to disclose the use of spellcheckers, translation services or writing style assistants such as Grammarly.

Whenever you use AI-generated content, independently verify the claims made and insert references to sources supporting the claims including DOIs (for scholarly publications) or URLs (to non-scholarly sources such as Wikipedia). If you use chatbots, make sure that you do not feed them private, confidential or sensitive information. If you plan to use ChatGPT, document how you use it through the Shared Links service, https://help.openai.com/en/articles/7925741-chatgpt-shared-links-faq

Bias Badges, Bad Science Bingo, Worst Science Awards, and Meme Competition
To insert some fresh air in the serious atmosphere in which we work to find fatal flaws and other bad stuff, we'll also have a bad science bingo, a series of worst paper awards, and a meme competition. Throughout the course we'll award bias badges to papers with flaws on the bad science bingo card. Also we will give out a series of worst paper awards for bad science. Which paper will win the worst science award for the highest number of problems? Which paper will win the worst citation award for being cited most frequently for the wrong reasons? Which journal will win the award for the highest impact factor having published a paper with a fatal flaw? In the meme competition, the best winner generates the loudest laugh with a meme about bad science, but obviously everyone who laughs is a winner.

Readings

- One target article from a preselected list of candidate articles.
- One replication or a commentary on the target article.
- Bekkers, R. (2020). How to Review a Paper. https://osf.io/7ug4w
- Bekkers, R. (2021). How to organize your data and code. April 2, 2021. https://renebekkers.wordpress.com/2021/04/02/how-to-organize-your-data-and-code/
- DeBruine, L. & Lakens, D. (2024). papercheck: Check Scientific Papers for Best Practices. R package version 0.0.0.9002, https://scienceverse.github.io/papercheck/, https://github.com/debruine/papercheck
- Epskamp, S. & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute p values. Retrieved from http://CRAN.R-project.org/package=statcheck.
- Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022a). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PloS one*, 17(2), e0263023. https://doi.org/10.1371/journal.pone.0263023
- Gopalakrishna, G., Wicherts, J. M., Vink, G., Stoop, I., van den Akker, O. R., ter Riet, G., & Bouter, L. M. (2022b). Prevalence of responsible research practices among academics in The Netherlands. *F1000Research*, 11(471), 471. https://doi.org/10.12688/f1000research.110664.2
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M.B., Rohrer, J.M., Romero, F., Scheel, A.M., Scherer, L.D., Schönbrodt, F. & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716
- 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://doi.org/10.1177/0956797611417632
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, 99: 178–182. https://dx.doi.org/10.1177/014107680609900414
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162-168. https://doi.org/10.1177/09637214211067779

Further resources

Blogs

Retraction Watch: a website tracking retractions “as a window into the scientific process”, https://retractionwatch.com/

Data Colada: Uri Simonsohn, Leif Nelson and Joe Simmons “thinking about evidence, and vice versa”. The blog identified multiple cases of data fabrication. https://datacolada.org/

Nick Brown's blog: The adventures of a self-appointed data police cadet, https://steamtraen.blogspot.com/

Andrew Gelman’s blog on statistical modeling, causal inference, and social science: https://statmodeling.stat.columbia.edu/

Podcasts

The Studies Show: a weekly podcast series about the latest scientific controversies, with Tom Chivers and Stuart Ritchie. Episode 49 is about scientific publishing: https://www.thestudiesshowpod.com/p/episode-49-scientific-publishing

Nullius in Verba: a podcast in which Daniel Lakens and Smriti Mehta discuss “what science is and what it could be”. Episode 42 is about the quality of research, https://nulliusinverba.podbean.com/e/reading-papers/

Everything Hertz: Dan Quintana and James Heathers discussing “methodology, scientific life, and bad language”, https://everythinghertz.com/