Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Wissenschaftstheorie & Einführung in das wissenschaftliche Arbeiten

Open Science: Science in practice

Dr. Blazej Baczkowski (Błażej Bączkowski)

Intro: motivation for teaching “Open Science” — my take

- This lecture does NOT aim to undermine science / research / academic psychology. On contrary, it aims to show the struggles / difficulties the community faces due to various factors (some are in our control and some are not), and to show how the community aims to overcome them to produce better science / research / academic psychology!
- **YOU** are also part of the process:
  - regardless whether a student, a future practitioner, a researcher, or a tax-payer, you are a recipient of science (and it may affect you in one or the other way)
  - **you can have an impact on the “credibility revolution” that is happening NOW!**
  - some general things from my perspective might be worth mentioning:
    - I would recommend not taking things at face value, but also not being critical just for the sake of it — provide a logical argumentation (re-think lecture 4 and 5)
      - there is always something one can learn (finding good solutions is important but delineating problems that really matter perhaps even more; re-think lecture 2 and 3)
    - I would encourage you to see science (and your studies) as **ongoing epistemic activity** and not as a fixed body of knowledge that you have to consume / acquire
      - uncertainty seems inevitable but it is not mean to be treated as ignorance (re-think lecture 5 on scientific method and 6 on bayesianism)

Terminology

- Reproducibility — testing the reliability of a prior finding using the same data and the same analysis strategy
  - *outcome reproducibility* failure suggests that the original result may be wrong
  - *process reproducibility* failure indicates that the original result cannot be verified
  - requires the availability of (raw/ source) data and analysis strategy (e.g., a computer code and necessary software); therefore data sharing and code reduces *process reproducibility failures*, which can reveal more *outcome reproducibility failures*
- Robustness — testing the reliability of a prior finding using the same data and a different analysis strategy
  - if evidence is fragile (across reasonable variation in analysis), then it suggests the finding is contingent on specific decisions of the researcher
  - fragility does not meant that finding is wrong, but it is a risk factor for generalizability (see, for example, over-fitting)
- Replicability — testing the reliability of a prior finding with different data
  - central in scientific method
  - seems simple: “do the same study again and see if the same outcome recurs”, but it is not clear what counts as *the same study and outcome*
  - generic variations:
    - close replication — an experimental procedure is repeated as closely as possible (exact replication does not exist)
    - systematic replication — an experimental procedure is largely repeated (some intentional changes can be introduced)
    - conceptual replications — a different experimental procedure is used to test the same hypothesis / phenomena (allows to test the generalisability)

Nosek et al., (2022) 10.1146/annurev-psych-020821-114157

The replicability / reproducibility crisis — some generic aspects

- methodological crisis in which the findings of many scientific studies are not replicated (or reproduced)
- replication is an essential part of scientific method, and the crisis called into question scientific knowledge
  - “media” have probably played a role (yet, such exposure is necessary and useful!)
- the crisis is not specific to psychology but present also in other disciplines ranging from medicine, cancer biology, economics, and chemistry
  - in fact, considerable efforts have been undertake in psychology to counteract the situation (“credibility revolution”)!!
- near-100% replicability would indicate that we adopt extremely conservative research agenda, study phenomena we already understand well, and test hypotheses that have extreme high prior odds
  - this would lead to “zero” research progress
  - non-replicability to some extent is healthy and indicates seeking novelty / innovation (at the end of the day, we aim to change the boundary of our knowledge -> justified true beliefs :)
- “Low replicability is partly a symptom of tolerance for risky predictions and partly *a symptom of poor research practices.* ***Persistent*** **low replicability is a symptom of poor research practices.**” (Nosek et al., 2022)

Mis-conduct?
— this is not the topic for today

Published: 16 March 2006

A specific amyloid-β protein assembly in the brain impairs memory

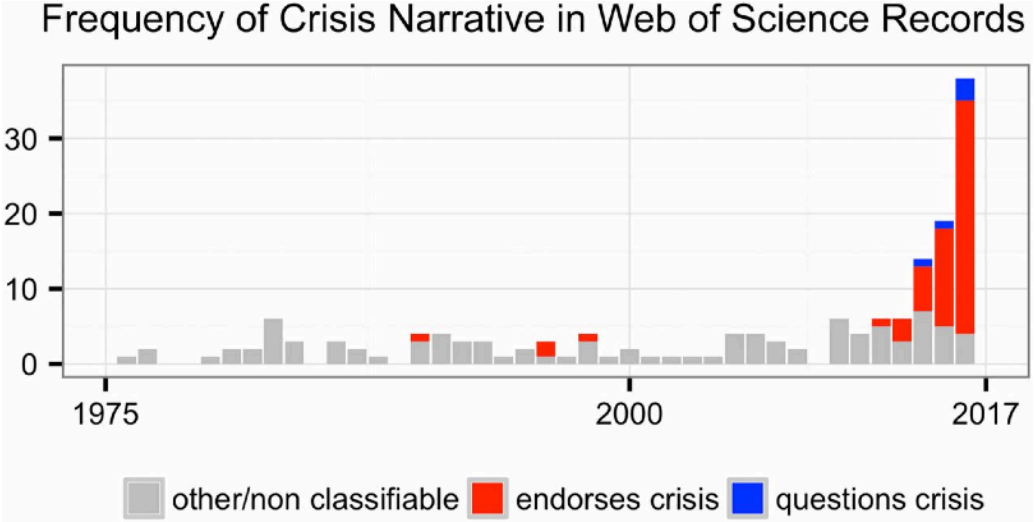
Sylvain Lesné, Ming Teng Koh, Linda Kotilinek, Rakez Kayed, Charles G. Glabe, Austin Yang, Michela Gallagher & Karen H. Ashe

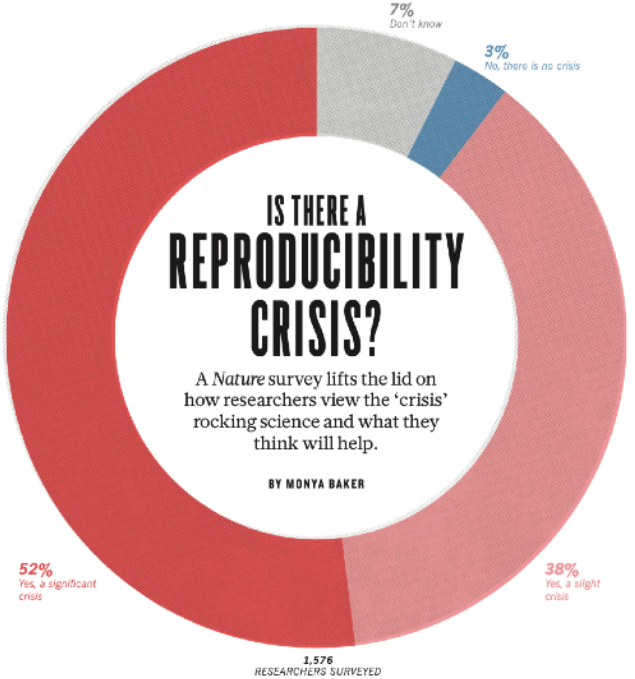
Nature **440**, 352–357 (2006) | Cite this article

14 July 2022 Editor's Note: The editors of Nature have been alerted to concerns regarding some of the figures in this paper. Nature is investigating these concerns, and a further editorial response will follow as soon as possible. In the meantime, readers are advised to use caution when using results reported therein.

Read more here:
https://www.science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease

A narrative of “crisis”



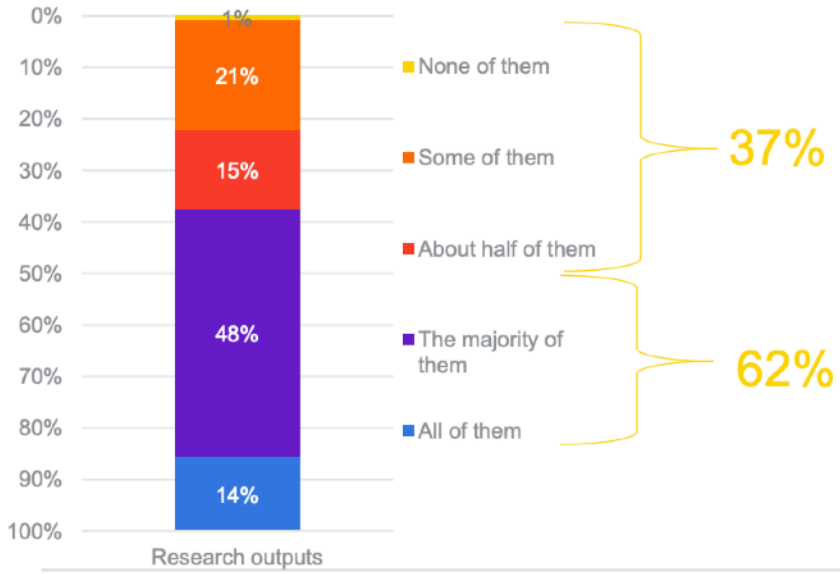


Fanelli (2018) https://doi.org/10.1073/pnas.1708272114

Baker (2016) https://doi.org/10.1038/533452a

Trust in research by researchers

Thinking about the various research outputs that you interacted with (or encountered) last week what proportion of the outputs would you consider trustworthy?



Base: All respondents (n=3133)

Which of the following mechanisms do you employ to compensate for any lack of confidence you have in the content you are considering reading/accessing?

Base: All respondents that do not think all research outputs are trustworthy (n=2715)

Source: https://www.elsevier.com/connect/trust-in-research

Summary of recent evidence for replicability status in psychology

associations between personality traits and outcomes (self-reports). 90% achieved sig. in the same directions with effect sizes 91% as large as found in the original studies

studies selected from Science and Nature; 62% achieved sig. in the same direction effect size 50% as large as in the original studies

3 psych journals 36% achieved sig. effect size of 49% as large as in the original studies

56% achieved sig. in the same direction effect size 53% as large as in the original studies

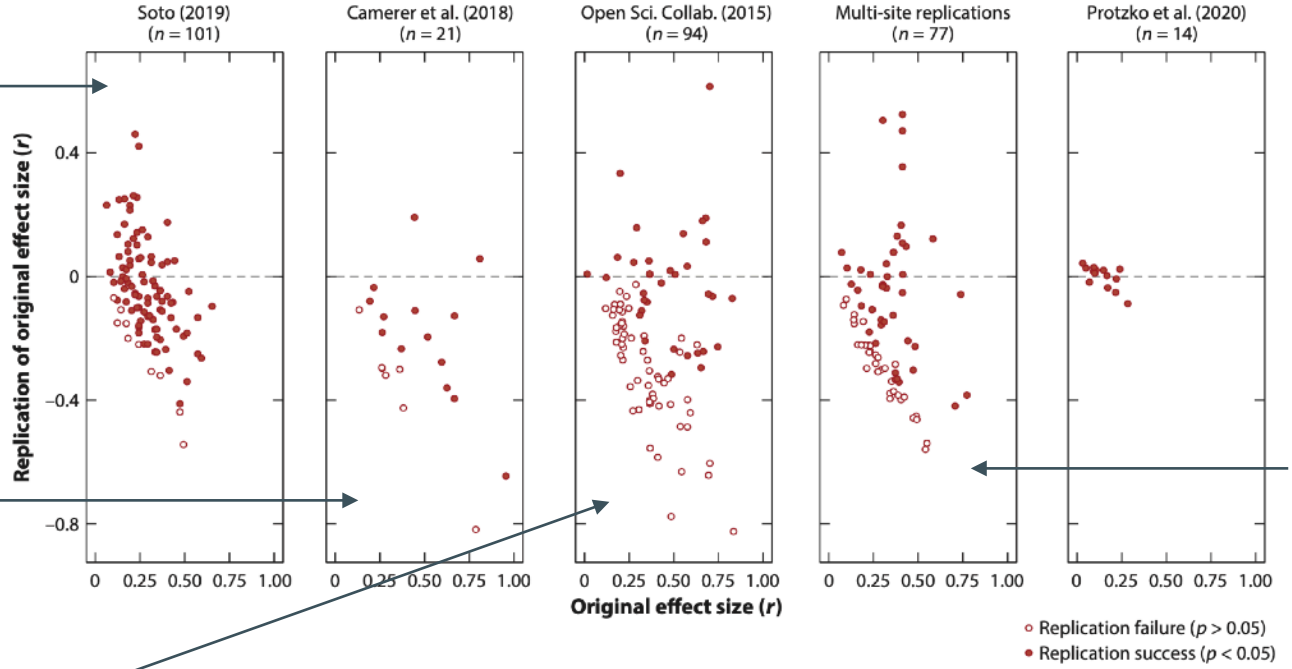


Figure 1

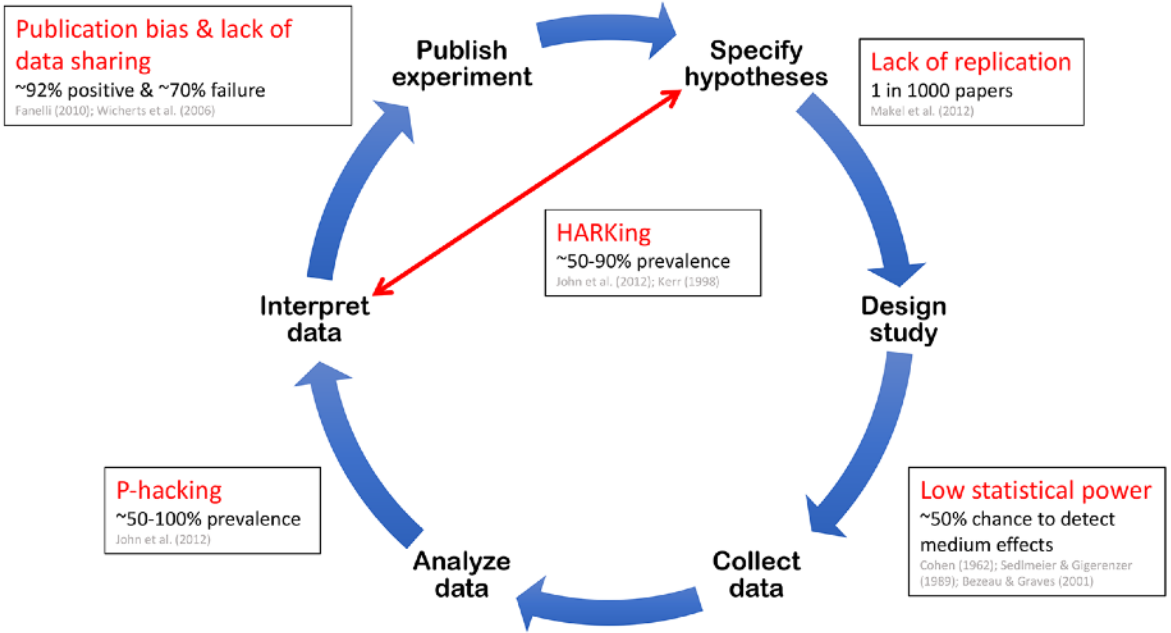
Replication outcomes for three systematic replication studies (Camerer et al. 2018, Open Sci. Collab. 2015, Soto 2019), multi-site replication studies, and a prospective best-practice replication study (Protzko et al. 2020). Values above zero indicate that the replication effect size was larger than the original effect size. Solid circles indicate that replications were statistically significant in the same direction as the original study. Studies with effects that could not be converted to *r* or original studies with null results are excluded.

Nosek et al., (2022) 10.1146/annurev-psych-020821-114157

Part of the problem — a mixture of factors

- dichotomous thinking (Did the study replicate? Yes / No) despite the recognition that empirical evidence rarely provides clear-cut answers
  - usually based on *p-value* of less than .05 (null hypothesis significance testing)
  - but other statistical inference methods are not free from limitations; at the end of the day we do probabilistic inference (re-think the problem of induction in lecture 5 & 6)
  - but "yes / no" answers are useful heuristics
- theoretical im-maturity?
  - operationalisation (precise definitions needed to formulate precise predictions)
  - boundary conditions
  - auxiliary hypotheses (implicit?)
- low prior probability of the hypothesis (surprising findings that caught attention)
  - "extraordinary claims need extraordinary evidence"

Questionable research practices in the context of the hypothetico-deductive method



CENTER FOR OPEN SCIENCE

Lack of replication impedes the elimination of false discoveries and weakens the evidence base underpinning theory. **Low statistical power** increases the chances of missing true discoveries and reduces the likelihood that obtained positive effects are real. Exploiting researcher degrees of freedom (**p-hacking**) manifests in two general forms: collecting data until analyses return statistically significant effects, and selectively reporting analyses that reveal desirable outcomes. **HARKing**, or hypothesizing after results are known, involves generating a hypothesis from the data and then presenting it as a priori. **Publication bias** occurs when journals reject manuscripts on the basis that they report negative or undesirable findings. Finally, **lack of data sharing** prevents detailed meta-analysis and hinders the detection of data fabrication. Source: www.cos.io/initiatives/registered-reports

Part of the problem: the incentive system? — pressure to publish and have ‘nice’ results

“What part of a research study — hypotheses, methods, results, or discussion — should remain beyond a scientist’s control? The answer, of course, is **the results**: the part that matters most for publishing in prestigious journals and advancing careers. This **paradox** means that the careful scepticism required to avoid massaging data or skewing analysis is pitted against the drive to identify eye-catching outcomes. Unbiased, negative and complicated findings lose out to cherry-picked highlights that can bring prominent articles, grant funding, promotion and esteem.”

Chambers (2019). Nature, doi.org/10.1038/d41586-019-02674-6

File Drawer Problem

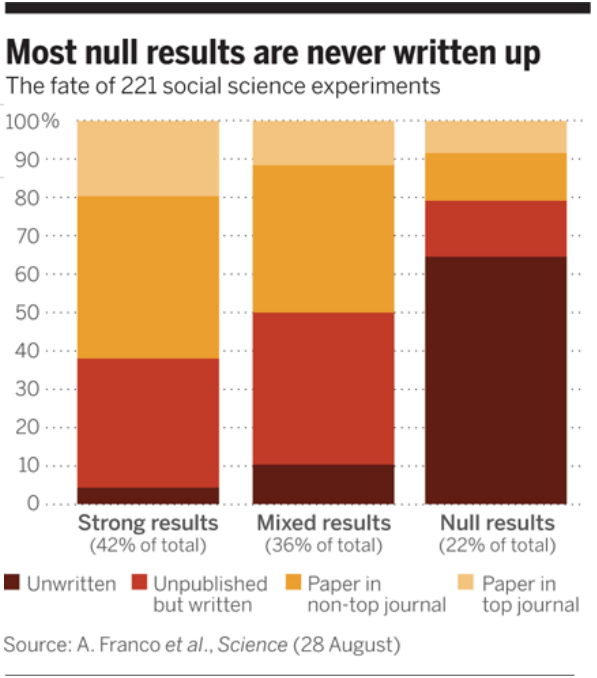
Publication bias in the social sciences: Unlocking the file drawer

ANNIE FRANCO, NEIL MALHOTRA , AND GABOR SIMONOVITS Authors Info & Affiliations

SCIENCE · 28 Aug 2014 · Vol 345, Issue 6203 · pp. 1502-1505 · DOI: 10.1126/science.1255484

The file drawer is full. Should we worry?

Experiments that produce null results face a higher barrier to publication than those that yield statistically significant differences. Whether this is a problem depends on how many null but otherwise valid results might be trapped in the file drawer. Franco *et al.* use a Time-sharing Experiments in the Social Sciences archive of nearly 250 peer-reviewed proposals of social science experiments conducted on nationally representative samples. They find that only 10 out of 48 null results were published, whereas 56 out of 91 studies with strongly significant results made it into a journal.



Publication bias (Publikationsbias) — a failure to publish (disseminate) results of studies on the basis of the direction or strength of the results

- Two studies can be execute with similar quality but they may produce qualitatively opposite results (small effect size, sample variability), and one is published while the other not.
- This may motivate researchers to (a) “find good” results (data dredging, data snooping, p-hacking), or (b) decide not to communicate the study (assuming people will not be interested), or (c) yet something else
  - for example, running many tests and reporting only those that came out as “statistically significant” (selective outcome and analyses reporting)
- published studies are no longer a representative sample of the available evidence
  - the bias is propagating: distorted systematic reviews and meta-analyses — some naive meta-analyses done in the past might be wrong (or needs to be re-done with some statistical corrections)

Publications in commercial academic journals

- academic journals serve as permanent and transparent forum for the presentation (and discussion) of research
- articles mostly present original research (but also reviews or opinions) that are *peer-reviewed*
- journals differ in topics, type of papers they publish, and their prestige (*impact factor*)
- they also vastly differ in their *access* by the public (paywall vs open access) and processing costs (inequality across countries and institutions regardless of the potential scientific impact)

Practical advice
- literature search

Scientific search engines:

- ✓PubMed (https://www.ncbi.nlm.nih.gov/pubmed/advanced)
- ✓ScienceDirect (http://www.sciencedirect.com/science/search)
- ✓Web of Science (https://webofknowledge.com)
- ✓Google Scholar (https://scholar.google.de/)

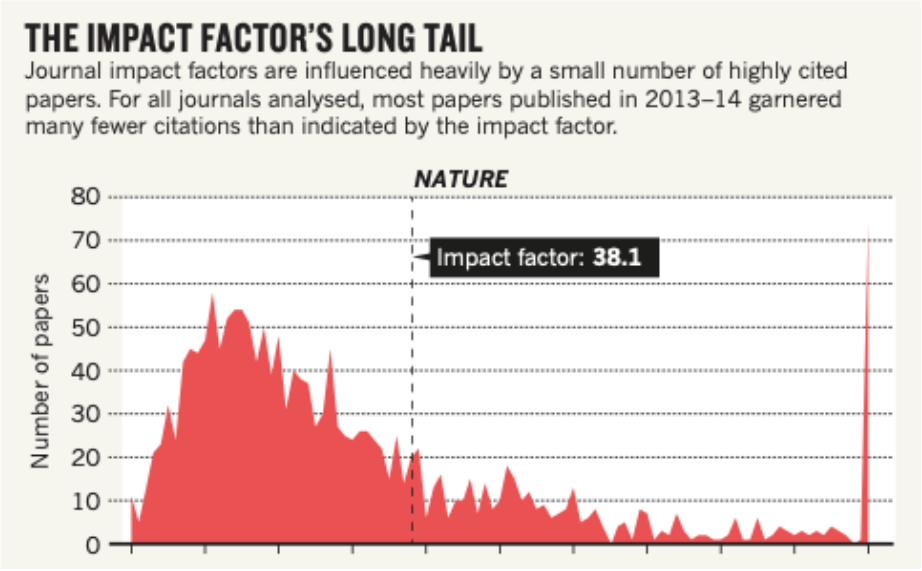
Databases:

- ✓PsycINFO (www.apa.org/psycinfo)
- ✓PSYNDEX (https://www.zpid.de/PSYNDEX)
- ✓PsycARTICLES (www.apa.org/pubs/databases/psycarticles)

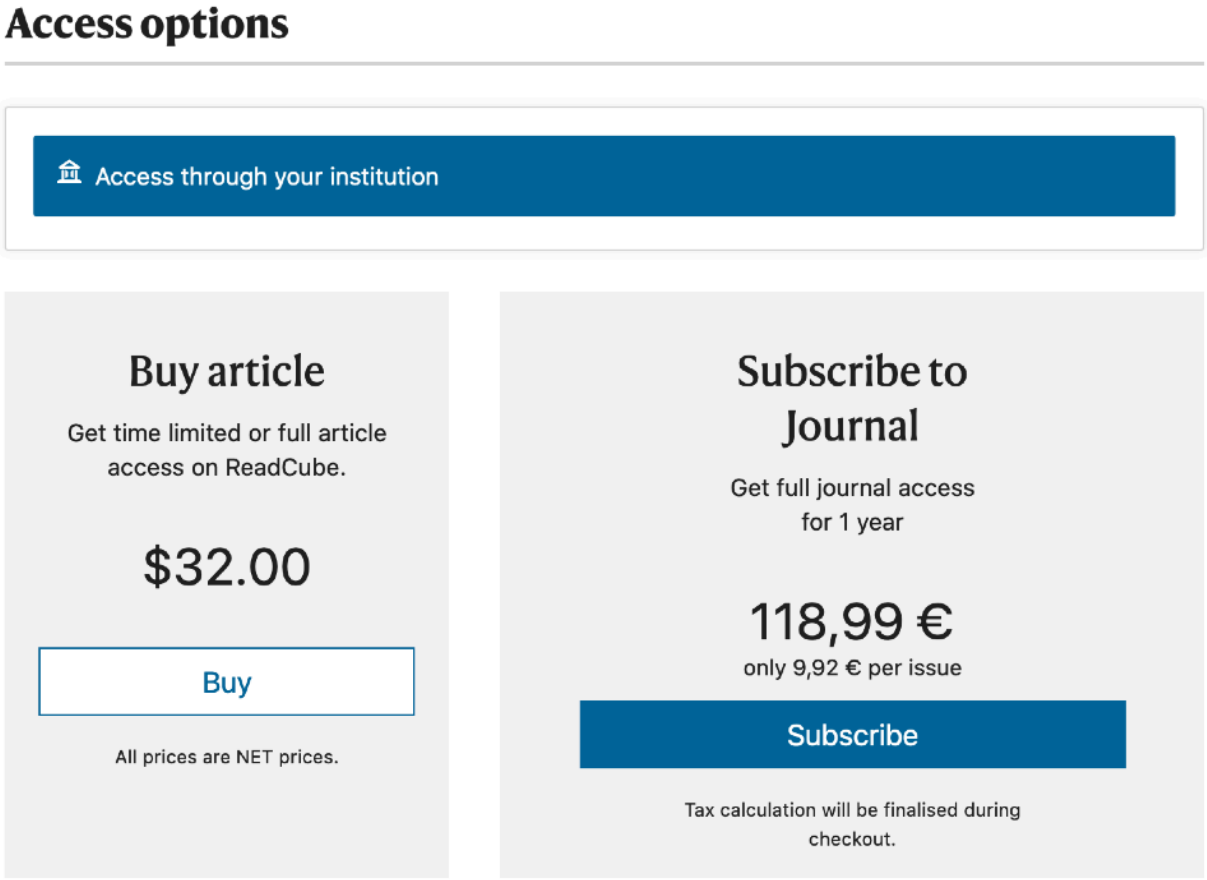
Peer Review Process

Commercial academic journals and impact factor

- Impact factor (IF) is scientometric index calculated by Clarivate (analytics company) that reflects the yearly average number of citations of articles published in the last two years in a given journal (conceived in the 1970s as a useful tool for research libraries to judge the relative merits of journals when allocating their subscription budgets)
  - “Science" IF in 2021 is 63.7
  - “Psychological Science” IF in 2018 is 4.9
  - “Journal of Experimental Psychology: General” IF in 2020 is 4.91



Access to *Nature Human Behaviour* without subscription



Costs: *Nature Human Behaviour* publishing options

Authors submitting primary research articles to *Nature Human Behaviour* have the option of publishing their research using either:

1 – Traditional subscription publishing model – an article is submitted and is assessed by our editors. If suitable it will be put through Peer Review, and if successful (subject to amendments), will be eligible for publication. Published articles are made available to institutions and individuals who subscribe to *Nature Human Behaviour* or who pay to read specific articles.

2 – Gold Open Access – same publishing process as above. The difference is that when an article is accepted for publication, the author/s or funder/s pay an Article Processing Charge (APC). The final version of the published article is then free to read for everyone. The APC to publish Gold Open Access in *Nature Human Behaviour* is €9,500.

This is begging the question: what is included in that price?

Psychology today is different than a decade ago — evidence of change (“credibility revolution”)

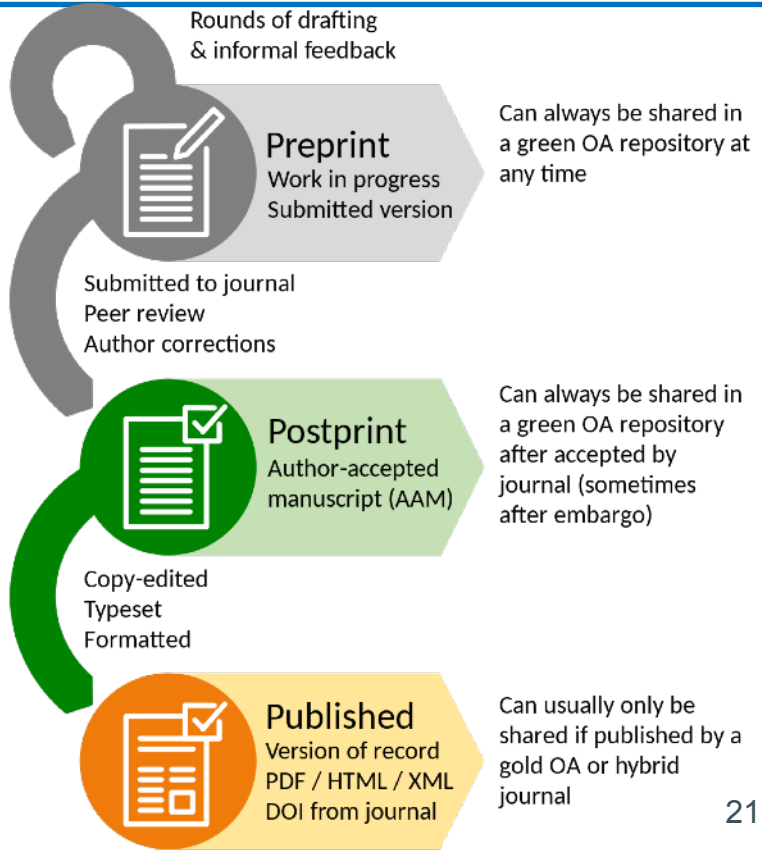
- social context starts promoting rigour, transparency, and replicability
  - open data, open materials (e.g., computer code), pre-registrations (see next slides)
- initiatives started to emerge and grow
  - Centre for open science (https://www.cos.io)
  - German reproducibility network (https://reproducibilitynetwork.de)
- publishing system undergoes transformation
  - pre-prints and peer community journals next to commercial journals (see next slides)
- funding agencies and hiring committees start changing their policies
  - open access publications and “open science” practices become required
- challenges still exist!!!
  - for example, “openness” becomes a norm but “quality control” not yet

Pre-prints
— sharing manuscripts without commercial journals

- advantages: prompt dissemination, increase number of citations, chances for collaboration, increased transparency, **open access**, may communicate negative findings (or exploratory work), and others…
- criticism/disadvantages: **lack of peer-review**, questionable quality (?), concerns about pre-mature data, risk of pre-mature media coverage, information overload, and others…
- preprint repositories related to psychology: **PsyArXiv, bioRxiv**

PsyArXiv

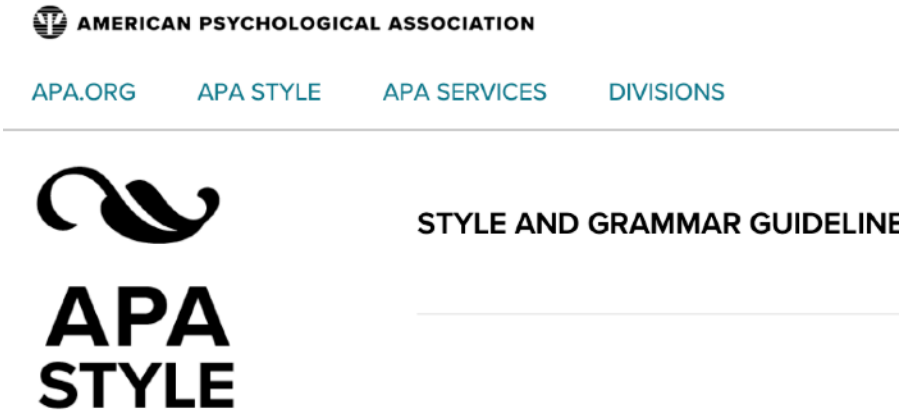
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY



Peer review without gatekeeping (eLife, Oct ’22) — eliminating accept/reject decisions after peer review

“From next year, we will no longer make accept/reject decisions at the end of the peer-review process; rather, **all papers** that have been peer-reviewed will be published on the eLife website as Reviewed Preprints, accompanied by an eLife assessment and public reviews. The authors will also be able to include a response to the assessment and reviews.”

Practical advice
- style of a scientific manuscript or student assignment



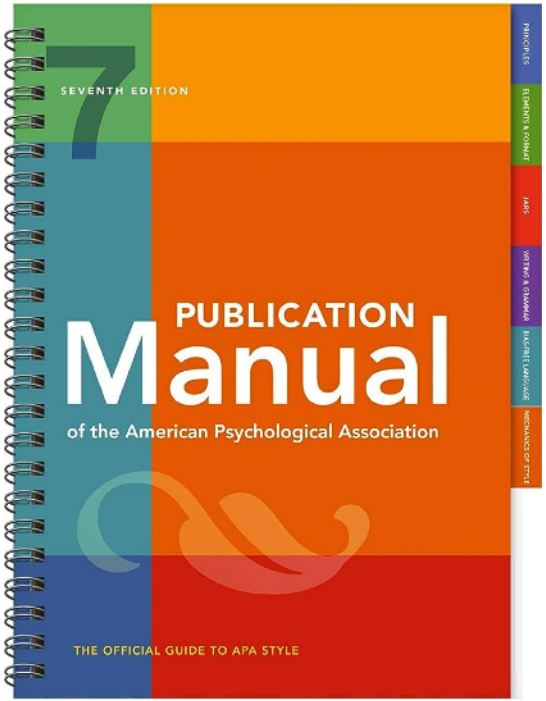
https://apastyle.apa.org/

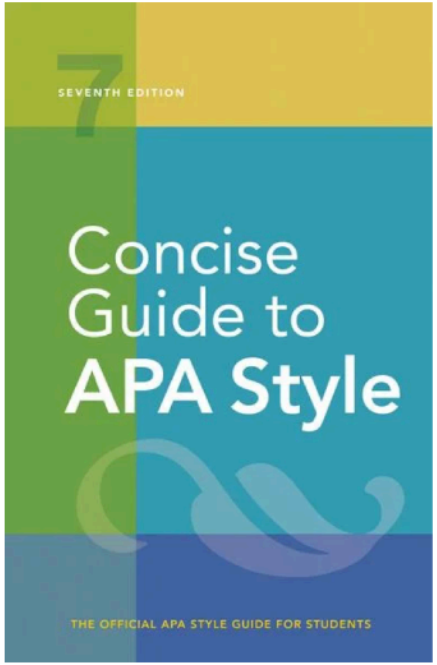
https://www.scribbr.com/knowledge-base/

https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/index.html

See the APA materials uploaded to STINE and templates at
https://apastyle.apa.org/style-grammar-guidelines/paper-format/sample-papers

Practical advice
- style of a scientific manuscript or student assignment

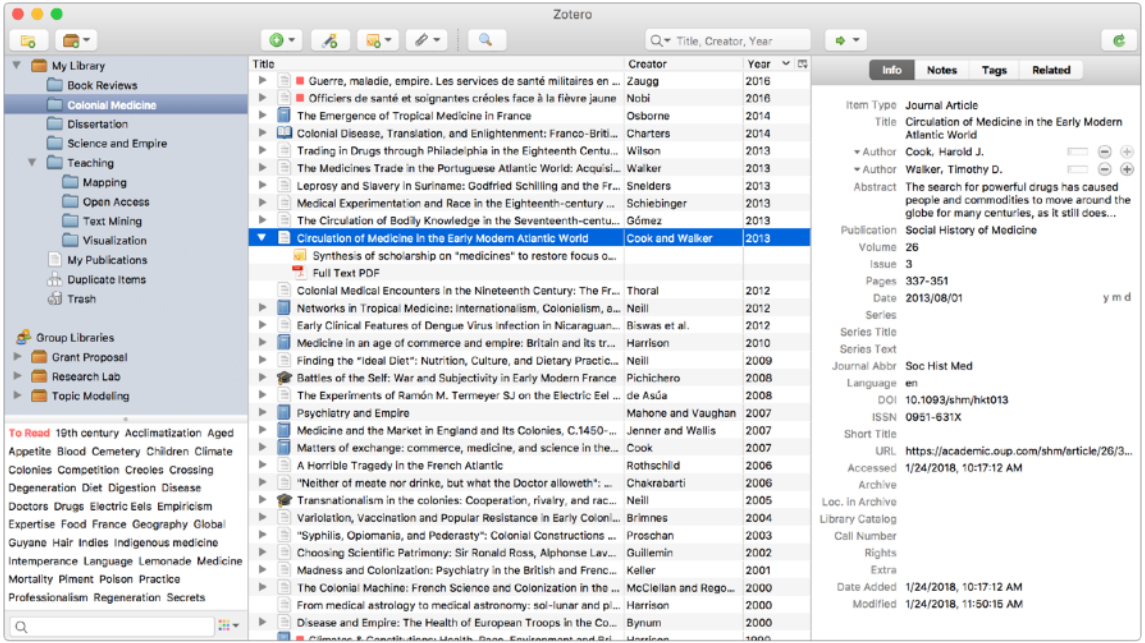




Practical advice
- reference manager (example)

zotero

https://www.zotero.org/



Pre-registration
— the distinction between prediction and postdiction

- When you preregister your research, you're simply specifying your research plan in advance of your study and submitting it to a registry.
- Preregistration separates *hypothesis-generating* (exploratory) from *hypothesis-testing* (confirmatory) research. Both are important. But the same data cannot be used to generate *and* test a hypothesis, which can happen unintentionally and reduce the credibility of your results. Addressing this problem through planning improves the quality and transparency of your research. This helps you clearly report your study and helps others who may wish to build on it.
- Source: https://www.cos.io/initiatives/prereg

Increasing transparency

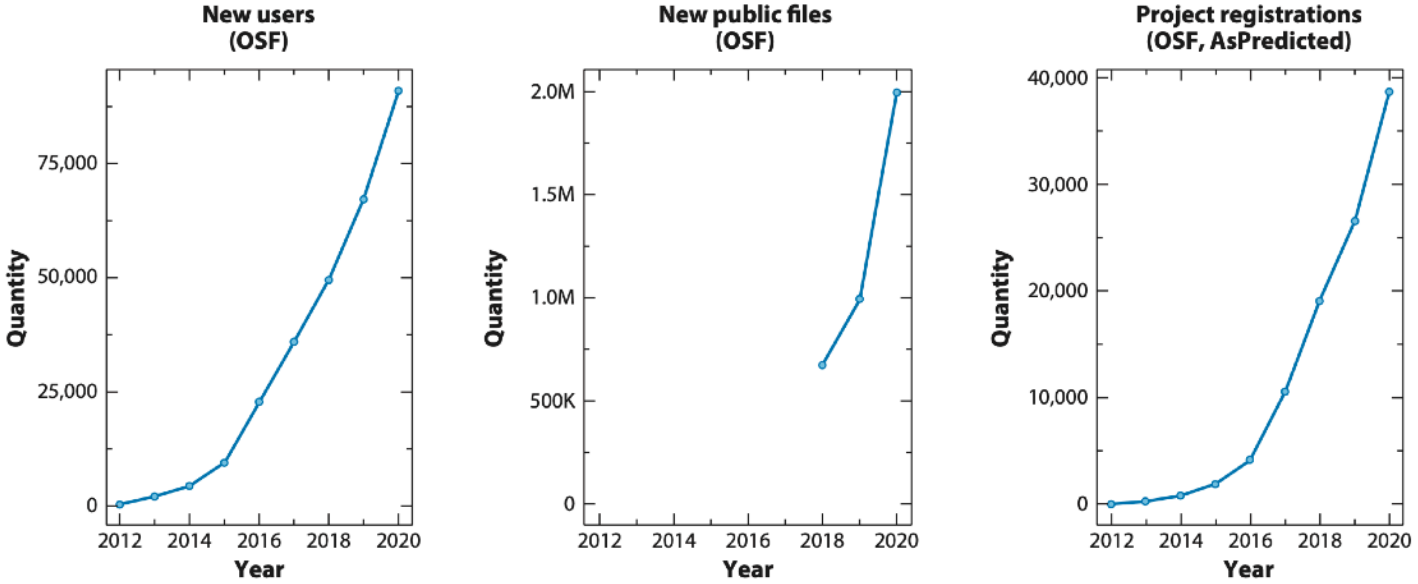


Figure 4

Yearly counts of users, sharing of files (research data, materials, code), and registration of studies on OSF and AsPredicted, two popular services for psychologists and allied disciplines. Data for new public files (sharing) prior to 2018 are not available. Abbreviation: OSF, Open Science Framework.

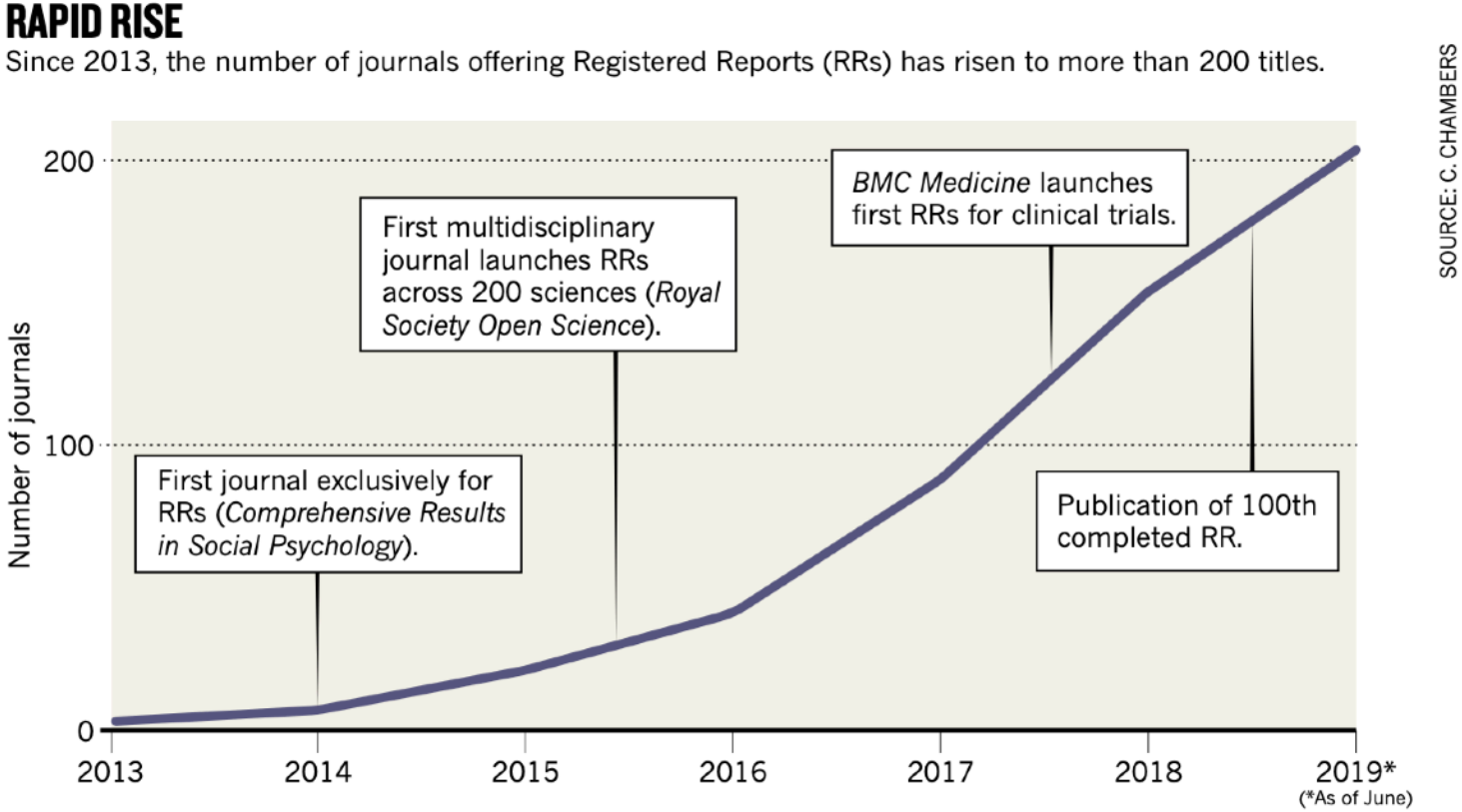
Nosek et al., (2022) 10.1146/annurev-psych-020821-114157

Registered reports (RR):
Peer review before results are known to align scientific values and practices

- Registered Reports is a publishing format that emphasizes the importance of the research question and the quality of methodology by conducting peer review prior to data collection. High quality protocols are then provisionally accepted for publication if the authors follow through with the registered methodology.
- This format is designed to reward best practices in adhering to the hypothetico-deductive model of the scientific method. It eliminates a variety of questionable research practices, including low statistical power, selective reporting of results, and publication bias, while allowing complete flexibility to report serendipitous findings.

https://www.cos.io/initiatives/registered-reports

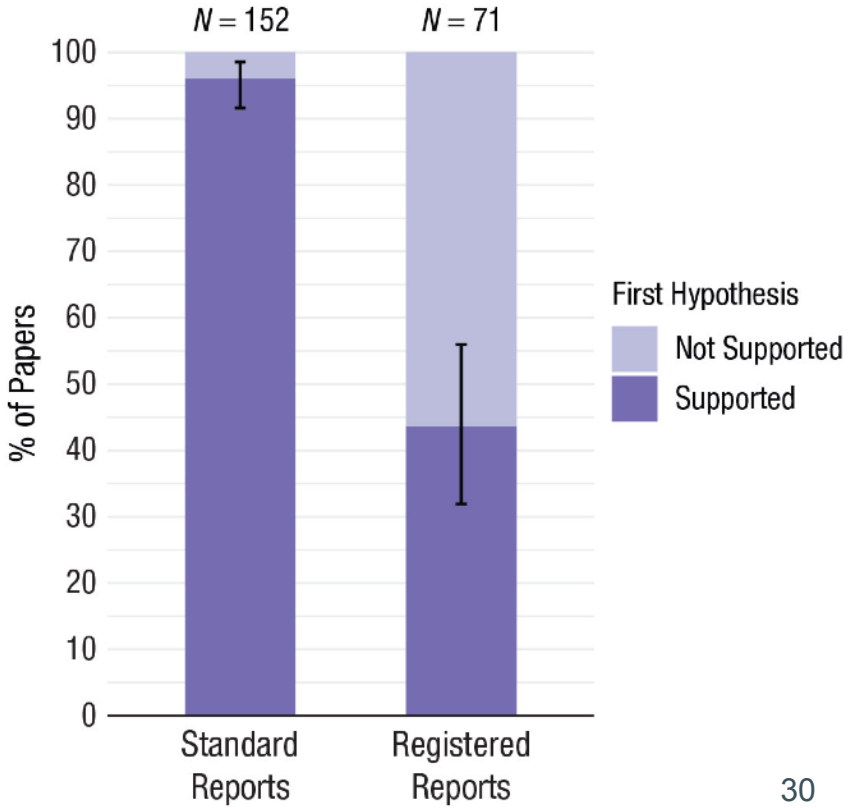
Increase in the number of journals offering RRs



Chambers (2019). Nature, doi.org/10.1038/d41586-019-02674-6

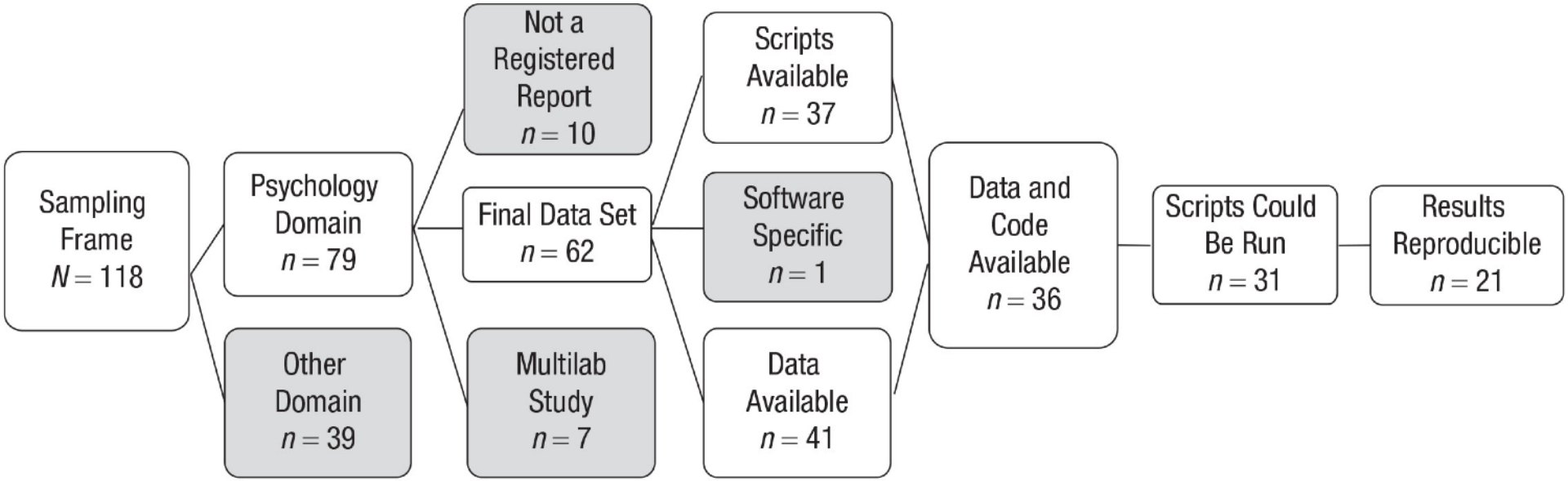
RRs and failures in hypothesis support

"Selectively publishing results that support the tested hypotheses ("positive" results) distorts the available evidence for scientific claims. For the past decade, psychological scientists have been increasingly concerned about the degree of such distortion in their literature. A new publication format has been developed to prevent selective reporting: In Registered Reports (RRs), peer review and the decision to publish take place before results are known. We compared the results in published RRs (*N* = 71 as of November 2018) with a random sample of hypothesis-testing studies from the standard literature (*N* = 152) in psychology. Analyzing the first hypothesis of each article, we found 96% positive results in standard reports but only 44% positive results in RRs. We discuss possible explanations for this large difference and suggest that a plausible factor is the reduction of publication bias and/or Type I error inflation in the RR literature."



Sheel (2021). doi.org/10.1177/25152459211007467

Reproducibility in psychology in RRs (Obels et al., 2020)

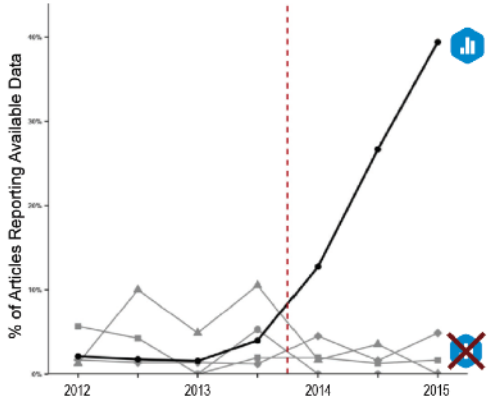


Literature between 2014-2018; computational reproducibility in registered reports at 58%…

I would expect this to be improved in 2022 given increased fluency in computational methods.

Open methods and open data — ‘good climate’ but some challenges remain…

- funding agencies are moving from “open access” to “open data”
  - German Research Foundation (DFG): publicly funded research belongs to the public
  - ethical considerations remain (e.g., neuroimaging data)
- Open materials indicate that the analysis strategy is in principle checkable (verifiable)
  - but who actually checks that and who is responsible?
  - for example, retractions in a journal are costly and imply repetitional damage; who is rewarded for an in-depth review?
- Research institutes differ in providing stable repositories for sharing materials
  - some of us use OSF (permanent storage for free for the next 50+ years) or GitHub (commercial)



https://www.cos.io/initiatives/badges

Schlüsselwörter

- Reproduzierbarkeit, Robustheit, Replizierbarkeit / Reproducibility, Robustness, Replicability
- “Replikationskrise” (Krise der Reproduzierbarkeit) / “replication crisis” (core idea)
- der Publikationsbias / publication bias (“file drawer problem”)
- HARKing
- wissenschaftliches Publizieren: / academic publishing:
    - peer review
    - pre-registration
    - registered reports
    - pre-print
    - open access

Ergänzende Literatur / Quellenmaterial

- 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. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. In Annual Review of Psychology (Vol. 73, Issue 1, pp. 719–748). Annual Reviews. https://doi.org/10.1146/annurev-psych-020821-114157
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. In Proceedings of the National Academy of Sciences (Vol. 115, Issue 11, pp. 2600–2606). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1708274114
- Scheel AM, Schijen MRMJ, Lakens D. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*. 2021;4(2). doi:10.1177/25152459211007467
- Obels P, Lakens D, Coles NA, Gottfried J, Green SA. Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. Advances in Methods and Practices in Psychological Science. 2020;3(2):229-237. doi:10.1177/2515245920918872
- Chambers, C. (2019). What's next for Registered Reports? In Nature (Vol. 573, Issue 7773, pp. 187–189). Springer Science and Business Media LLC. https://doi.org/10.1038/d41586-019-02674-6
- Farid Anvari & Daniël Lakens (2018) The replicability crisis and public trust in psychological science, Comprehensive Results in Social Psychology, 3:3, 266-286, DOI: 10.1080/23743603.2019.1684822
- Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? In Proceedings of the National Academy of Sciences (Vol. 115, Issue 11, pp. 2628–2631). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1708272114

Ergänzende Literatur / Quellenmaterial

- https://osf.io/qvfp8/ (open science crash course by Felix Schönbrodt, Angelika M. Stefan, Lena Schiestel)
- https://osf.io/8njbp (From crisis to renaissance by Felix Schönbrodt)
- Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. In The Leadership Quarterly (Vol. 28, Issue 1, pp. 5–21). Elsevier BV. https://doi.org/10.1016/j.leaqua.2017.01.006
- Michael B Eisen Anna Akhmanova Timothy E Behrens Jörn Diedrichsen Diane M Harper Mihaela D Iordanova Detlef Weigel Mone Zaidi (2022) **Scientific Publishing: Peer review without gatekeeping** *eLife* **11**:e83889.
- https://www.elsevier.com/connect/trust-in-research