

AczelBalazs\_17\_22

# **Metascience from the Psychologist's Perspective**

**Akadémiai Doktori Értekezés**

**Aczél Balázs**

2022

## Table of Contents

<i>Acknowledgments</i> .....	3
<i>Abbreviations</i> .....	4
<i>List of Papers by the Author Directly Used in the Dissertation</i> .....	5
<b>1. INTRODUCTION AND AIMS OF STUDIES</b> .....	<b>6</b>
1.1. <i>Problems with the Publication Practice</i> .....	11
1.2. <i>Lack of Transparency</i> .....	17
1.3. <i>Issues in Statistical Practice</i> .....	21
<b>2. PUBLICATION PRACTICE</b> .....	<b>30</b>
2.1. <i>A Billion-Dollar Donation: Estimating the Cost of Researchers' Time Spent on Peer Review</i> .....	30
2.2. <i>Documenting contributions to scholarly articles using CRediT and tenzing</i> .....	49
2.3. <i>Researchers working from home: Benefits and challenges</i> .....	65
<b>3. TRANSPARENCY</b> .....	<b>85</b>
3.1. <i>A Consensus-Based Transparency Checklist</i> .....	85
3.2. <i>A survey on how preregistration affects the research workflow: Better science but more work</i> .....	93
3.3. <i>Seven Steps Toward More Transparency in Statistical Practice</i> .....	126
3.3.1. <i>Situational factors shape moral judgments in the trolley dilemma in Eastern, Southern, and Western countries in a culturally diverse sample</i> .....	153
<b>4. STATISTICAL PRACTICE</b> .....	<b>189</b>
4.1. <i>The role of human fallibility in psychological research: A survey of mistakes in data management</i>	189
4.2. <i>Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation</i> .....	217
4.3. <i>Discussion points for Bayesian inference</i> .....	235
4.4. <i>One statistical analysis must not rule them all</i> .....	248
4.5. <i>Consensus-based guidance for conducting and reporting multi-analyst studies</i> .....	251
4.6. <i>SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies</i> .....	273
<b>5. SUMMARY AND CONCLUSIONS</b> .....	<b>293</b>

## Acknowledgments

Most of the presented papers are large-scale collaborations that I led or co-led, therefore, it is not possible to list all those who contributed to the success of my projects. Nevertheless, the core idea of these projects came out of long, and often passionate discussions with my closest collaborators. I'm profoundly grateful to the thought-provoking conversations that I had (while waiting in airport check-in queues, during boring conference presentations, or while having some beer) with my closest collaborators: Barnabas Szaszi, Marton Kovacs, Bence Bago, and Bence Palfi (in no particular order). Throughout the years, I had the honour to work with outstanding and innovative minds. I have to mention E-J Wagenmakers, Alexandra Sarafoglou, Don van Ravenzwaaij, Gustav Nilsson, Felix Holzmeister, Alex Holcombe, Rink Hoekstra, and Zoltan Dienes (in reverse alphabetical order) with whom I'm always happy to collaborate.

The members of my research lab generate a lot of positive energy inside and outside the walls of the lab and I must be thankful to their ongoing support. I also wish to express my gratitude to the Affective Psychology Department of ELTE, that they acknowledged and supported my metascientific ambitions and tolerated my absence from some department meetings. I also thank Zsolt Demetrovics for his support and encouragement throughout the years, and for successfully nagging me into writing up this thesis. I conducted most of my work without funding, but the Bolyai János Research Scholarship, the National Excellence Program, the ELTE Promising Researcher Prize, Rector's Excellence Prize at ELTE, and the travel funds of my faculty provided me with financial support to reach my research aims. The Center for Open Science, the Society for the Improvement of Psychological Science, and the Psychological Accelerator are organisations that provide a lot of energy and workforce for ambitious metascientists. I don't think that without their hard work metascience would be such a successful enterprise.

Last, but not least, I'm in debt to my wife, daughters, and the whole family for their understanding and support of my obsessions in science.

## Abbreviations

**AIPE** - Accuracy in Parameter Estimation. A sample size estimation method used for parameter estimation. The approach aims to find the required sample size, such that the confidence interval has a certain expected width.

**APC** – Article Processing Charge

**APP** - A Priori Procedure. The approach aims to plan a sample size based on how close the researcher wishes both sample means to be to their respective population parameter, and how confident the researcher wants to be in this.

**BFDA** - Bayes Factor Design Analysis. This technique provides an expected sample size such that compelling evidence in the form of a Bayes factor can be collected for a given effect size with a certain long-run probability when allowing for sequential testing.

**BF** - Bayes Factor

**CARKing** - Critiquing After the Results are Known

**CRedit** - Contributor Role Taxonomy

**FORRT** - Framework for Open and Reproducible Research Teaching

**OSF** – Open Science Framework

**PARKing** - Preregistering After Results are Known

**ROPE** - Region Of Practical Equivalence. The region of effect sizes considered practically equivalent to zero under the HDI-ROPE method.

**RR** – (a) Registered Report; (b) Robustness Region

**SESOI** - Smallest Effect Size Of Interest. The region of effect sizes considered practically equivalent to zero under the TOST method.

**TOST** - Two One-Sided Tests. A frequentist statistical testing approach aimed at establishing equivalence between two groups.

**TPR** - True Positive Rate. The long-run probability of finding evidence for an effect, given that it exists. In our paper, statistical power is subsumed under TPR.

## List of Papers by the Author Directly Used in the Dissertation

- Aczel, B.**, Hoekstra, R., Gelman, A., Wagenmakers, E. J., Klugkist, I. G.,... & van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, 4(6), 561-563.
- Aczel, B.**, Kovacs, M., Van Der Lippe, T., & Szaszi, B. (2021). Researchers working from home: Benefits and challenges. *PLOS One*, 16(3), e0249127.
- Aczel, B.**, Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ... & Wagenmakers, E. J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1(3), 357-366.
- Aczel, B.**, Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1), 1–8.
- Aczel, B.**, Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *ELife*, 10, e72185.
- Aczel, B.**, Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4(1), 4–6.
- Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., ... & **Aczel, B.** (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, 6, 880-895.
- Holcombe, A. O., Kovacs, M., Aust, F., & **Aczel, B.** (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLOS One*, 15(12), e0244611.
- Kovacs, M., Hoekstra, R., & **Aczel, B.** (2021). The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management. *Advances in Methods and Practices in Psychological Science*, 4(4), 25152459211045930.
- Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & **Aczel, B.** (2022). SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies. *Advances in Methods and Practices in Psychological Science*, 5(1), 25152459211054060.
- Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., **Aczel, B.** (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7), 211997
- Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & **Aczel, B.** (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, 5(11), 1473–1480.
- Wagenmakers, E.-J., Sarafoglou, A., **Aczel, B.** (2022). One statistical analysis must not rule them all. *Nature*, 605(7910), 423-425.

## 1. INTRODUCTION AND AIMS OF STUDIES

“Certain features of the mental life of the scientist [...] affect the trustworthiness of his product, and in particular make the findings of science subject to weakness and passion like other human constitution.”  
Watson (1938). *Scientists are human*. Watts. London (p.21)

In his utopian novel, *New Atlantis*, Sir Francis Bacon depicted a fantasy island where the inhabitants fully dedicate their lives to the pursuit of science. Their fictional institute, *Salamon's House*, reflects Bacon's ideal organisation of a future scientific community (Bacon, 1627). He finds that a detailed specialisation of labour, roles, duties, and the supply of required equipment are key to the future of human knowledge and discovery. Bacon's vision, that inspired the foundation of the first scientific academies, underscores that beyond its principles and methods, science is a human enterprise. Four hundred years later, researchers work in a *Salamon's House*, more complex and intricate than Bacon imagined. How psychologists' research can understand this complexity and decrease its intricacy is the main focus of this thesis.

All those who carry the main roles in science – researchers, publishers, institutes, funders – have questions about science itself. *Can I trust what I read? Could the publication system work more efficiently? Are journal impact factors and citation good indicators of research excellence? How can funders best inspire innovation?* Metascience, also known as the science of science, aims to use scientific methodology to study science and answer such questions. In this thesis, I illustrate how psychologists can contribute to metascience through their specific perspective and methodology.

Today, psychologists play a very active role in the development of metascience. Perhaps psychologists' history and research scope are what predispose their interest in the workings of science. History of psychology shows that scientific self-reflection was always part of the quest to find its place among other disciplines. For example, the first major debate within the earliest period of psychology was about whether psychology as a science should first concentrate on the elements of consciousness or the adaptive function of mental activities. Edward Titchener argued that if psychology wants to follow the biological sciences, it must start its investigation from morphology, understanding the structure before concentrating on the function. In contrast, John

Dewey, William James, and the functionalists propagated that psychology should become a Darwinian science and take a pragmatic, function-centred perspective (Pléh, 2010).

The greatest turning point in the development of psychology, the behaviourist revolution, was also about the quest to become a true scientific discipline. John B. Watson's manifesto (1913) proclaimed that in order to eliminate the barriers between psychology and other natural sciences, psychology must abandon all its subjective and speculative aspects (such as introspection and the study of consciousness) and should focus on solely the objective investigation of behaviour.

The latest major direction-change in psychology, the cognitive turn, was also about what psychologists may investigate within the demarcation lines of science. While the neobehaviourists and Skinner insisted that behaviour must be studied without any recourse to inner mental states, Chomsky convincingly argued that human language cannot be understood without studying the mind and that it is within the principles of science that the hypothesis contains not-directly observable components, such as cognitive functions (Leahey, 2004).

Another historical aspect that triggers psychologists' scientific self-reflection is the constant presence of alternative approaches that strive to provide their own explanations to or treatments for psychological issues. Occultism, astrology, and spiritualism, for example, have an overlap of interest with psychology in that they try to provide frameworks for people's psychological life, except their methods fall outside of the methodology of science (Leahey, 2004). Parapsychology, however, has a more borderline status as it aims to use scientific methodology to study psychic and other paranormal claims. These claims make psychologists uncomfortable when drawing the boundaries of its discipline. Daryl Bem's 2011 study, for example, caused a major uproar in wide circles of psychology. The study which was published in the *Journal of Personality and Social Psychology* (Bem, 2011), claimed to provide evidence that future events can influence people's present behaviour. For example, he claimed that his participants' memory was improved for words that they rehearsed in the future. This precognition ability, of course, is regarded as pseudoscientific as it is against our basic understanding of causality and leads to apparent logical contradictions (Lobo & Crawford, 2003). Not surprisingly, the publication of this claim in a major journal of the discipline caused extensive waves, partially contributing to the so-called crisis of confidence (Pashler & Wagenmakers, 2012), leading to major reflections (Gelman, 2016; Kahneman, 2012; Nelson et al., 2018) and new manifestos (e.g., Munafò et al., 2017).

Beside their historical predispositions, psychologists' interest in the science, in a broad sense, can be linked to the relevance of their expertise to metascience. If science is a what scientists

do, then the study of their behaviour is, in a great part, a psychological question. It has been long argued that science cannot be understood without taking into account human nature. In 1938, the philosopher and psychologist John Dewey wrote the following to the foreword of D. L. Watson's book *Scientists are Human*: "... the pursuit of science and the products of science are relative to the mental world of the scientist, to the organization of his personality in all its phases, and that this in turn is relative to the social organization that subsists. [...] We have to know also what the conditions of present social life are doing to the scientist, and what, in consequence, the scientist does to and with science." (D. L. Watson, 1938, p. ix). Watson dedicated his whole book to the tenet that "... when we say that *scientists are human* we are directing attention to the fact that science – far from being the work of an abstract automaton or unemotional mechanism – is inextricably intertwined with the paradoxes and tragic imperfections of human nature." (1938, p. 8).

In his influential writings, Thomas Kuhn (1962) expressed that the full understanding of the dynamics of science demands "the competence of the psychologist even more than that of the historian" (p.86). Abraham Maslow's less known book, *The Psychology of Science: A Reconnaissance*, (1966) continued Kuhn's idea by analysing the dichotomy of 'normal' and 'revolutionary' science from a purely psychological perspective (Kožnjak, 2017). When exploring the question of how scientists deal with simultaneous discoveries, the sociologist Robert Merton also emphasised the psychological nature of these questions. He called for "advancing the sociology and psychology of science" (Merton, 1973, p. 372). He argued that topics such as scientists' resistance to scientific discovery (Barber, 1961) should be considered in psychological investigations.

By the early 2000s, a considerable literature developed adjacent to the disciplines of philosophy, history, and sociology of science: the *psychology of science*. According to its definition, this new field "applies the empirical methods and theoretical perspectives of psychology to scientifically study scientific thought and behavior (hence, it is a "metascience"). At its core, psychology of science is the empirical study of the biological, developmental, cognitive, personality, and social influences of scientific thought and behavior." (Feist, 2008, pp. 3–4). This psychology of science was later claimed to be a subdiscipline of psychology (Feist, 2011).

Although the term "psychology of science" didn't really take off, psychologists, with an increasing pace, discover that their concepts and methodology can be easily used in the study of



science and they put science in the focus of their investigation. Dorothy Bishop, for example, argued that "... we need to understand how cognitive constraints lead to faulty reasoning if we are to get science back on course and persuade those who set the incentives to reform. Fortunately, as psychologists, we are uniquely well positioned to tackle this issue." (2020, p. 3) She lists four cognitive constraints that influence how researchers process, understand, or remember information: (1) *Confirmation bias*, the tendency to seek out and remember evidence that supports a preferred viewpoint; (2) *Misunderstanding of probability*, the failure to understand how estimation scales with sample size; (3) *Asymmetric moral reasoning*, that the errors of omission are judged less seriously than errors of commission; (4) *Reliance on schemata*, meaning that perceiving and/or remembering tend to be in line with pre-existing knowledge, leading to omission or distortion of irrelevant information. She gives a detailed account how these constraints can bias experimental designs, data analyses, and scientific reporting.

I use another example to demonstrate how various aspects of the scientific workflow can be explored by psychologists. The order of authors in a manuscript carries importance since it has influence on the scientific credit that the scientists receive and the visibility of the authors. Whereas first and last authors often receive the main credit (Tschardt et al., 2007) first authors have higher visibility (Baum et al., 2022) and have higher chance for promotion (Einav & Yariv, 2006). Who receives these positions and what it matters in the scientific community is a social psychological question. The group dynamics and power distribution among the authors has a lot of influence on these decisions (Bartlett & Mercer, 2000). Early career researchers often lack the power and experience to negotiate a fair representation of their contribution. It's possible that new formatting requirements of the first page, in-text citation styles, and the rules of the reference list arrangement counteract these biases (Baum et al., 2022). Without targeted studies, however, it is difficult to decide how authorship rules and institutional guidelines could foster a transparency and equality between authors.

The aim of this thesis is to demonstrate that psychologists can play an important role in the development of metascience and their perspective and methodology are indispensable for the understanding and improvement of science. The thesis focuses on three lines of studies conducted in the following metascientific topics: (1) *Problems in the publication practice*; (2) *Lack of transparency*; and (3) *Issues in statistical practice*. These topics are among the main challenges that science currently need to face (Hardwicke et al., 2020) and psychologists can supply their perspective to each. Table 1 provides an overview of the topics and aims of the presented studies.

**Table 1** Summary Table of the Studies Presented in this Thesis

<b>Chapter</b>	<b>Topic</b>	<b>Aim</b>
<b>2.</b>	<b>Problems with the Publication Practice</b>	
2.1.	Peer review system	To estimate reviewers' time and salary-based contribution to publishers
2.2.	Article contributorship	To develop a web-application to assist authors in collecting and reporting required contributorship information
2.3.	Efficiency of researchers' work	To assess the benefits and challenges of researchers working from home
<b>3.</b>	<b>Lack of Transparency</b>	
3.1.	Transparency in social sciences	To develop a consensus-based checklist to report transparency-related aspects of social science studies
3.2.	Preregistration	To assess how preregistration can help the workflow of empirical studies
3.3.	Transparency practices in statistics	To provide concrete recommendations to promote transparency in statistical practice
3.3.1.	Good transparency and statistical practices in psychology	To present an empirical study that follows the recommended transparency and statistical practices
<b>4.</b>	<b>Issues with Statistical Practice</b>	
4.1.	Research data management mistakes	To identify the most frequent and most serious data management mistakes
4.2.	The strength of evidence in psychology	To quantify the evidence in non-significant results in psychological studies
4.3.	Conducting and reporting Bayesian analyses	To develop a consensus-based thinking guideline and reporting template for Bayesian analyses
4.4.	Alternative statistical analyses	To discuss the importance to explore alternative analyses
4.5.	Multi-analyst studies	To develop a consensus-based guidance on how to prepare and run multi-analyst studies
4.6.	Sample size estimation and justification	To develop a tool for calculating and justifying required sample sizes

## 1.1. Problems with the Publication Practice

In most of the disciplines, the fundamental communication platform of science is the journal article. The history of academic journals goes back to the 17<sup>th</sup> century and since then it proved to be a useful asset for scientists (Spier, 2002). The format came with some advantages beyond the periodicity of the spread of scientific information. For example, an interest of early scientists was to claim priority in research discoveries and they found that publishing in academic journals is a good method for that (Merton, 1963). In academia, journal articles serve not just the function of communication but, one way or another, they are the basis of performance evaluation systems as well (Rijcke et al., 2016). Since the print of the first volumes, tens of millions of articles has been published (Jinha, 2010) and the number of scholarly journals is beyond 100,000<sup>1</sup>. The growth rate of scientific publications show a steady trend for many decades now (Larsen & Von Ins, 2010) with doubling its volume every 15-17 years (Bornmann et al., 2021; Fortunato et al., 2018). Whereas it's hard to imagine science without writing journal articles, academic publication system is a greatly contentious topic. Researchers and research institutions alike strongly relate to debated issues such as predatory journals, article processing charges, Open Access mandates, "publish or perish" academic climate, authors' copyright questions, self-archiving, retractions, the peer review system, conflict of interest, impact factor, citation index, or publication bias. This thesis has no ambition to cover all issues but to highlight some current aspects in the ongoing debates.

One neuralgic subject of academic publishing is its business side. Publishing used to be solely on paper, requiring a printing house that edits, prints, and distributes the printed volumes. The world of production radically changed with the advent of digitalisation, the Internet, and the dominance of online communication. From this view point, it's easy to assume that online publication could be virtually free, but there are many parts of a journal article production that require resources (Grossmann & Brembs, 2021). Content acquisition, as a start, requires a staff to search and assign reviewers, communicate with reviewers and authors, execute plagiarism checks, obtain and maintain an online submission system, collect APCs etc. In the actual content preparation, copyediting, typesetting, language editing, graphs and other formatting, and technical checking of the manuscripts are among the tasks to complete. Further costs are associated with web hosting and uploading the materials to indexing platforms, such as Scopus, that all require

---

<sup>1</sup> Ulrich's Periodicals Database: <https://www.ulrichsweb.com/ulrichsweb/faqs.asp>

subscriptions. Grossmann and Brembs (2021) found that the associated costs of article production range from US\$200 per article to US\$1,000, but a representative scholarly paper would cost around US\$400 without profit. Although these expenses are not negligible, the authors state that the publication costs are only 15% of the subscription price, making academic publishing with its 40 percent profit margin allegedly the most profitable business in the world (“Time to Break Academic Publishing’s Stranglehold on Research,” 2018).

A direct consequence of the business model of academic publishing is that access to scientific products is controlled by the arrangements between the publisher and the authors. At the start of the Open Access advocacy (Laakso et al., 2011), a steep adoption of open journals was anticipated:

“... within ten years, open journals are likely to dominate scholarly communication.” (Getz, 2005, p. 15)

“This analysis suggests that Gold OA could account for 50 percent of the scholarly journal articles sometime between 2017 and 2021, and 90 percent of articles as soon as 2020 and more conservatively by 2025.” (Lewis, 2012, p. 493)

These predictions proved to be overly optimistic. In 2008, 8.5% of scholarly journal articles were found to be freely available on the publisher’s site and through other platforms an overall 20.4% of the manuscripts were Open Access (Björk et al., 2010). This proportion visibly increased in the coming years. A study of a random sample of 2011 papers found 50% of them to be freely available (Van Noorden, 2013). Nevertheless, sampling from all publication years, a 2018 study found no more than 28% of the scholarly literature open access (Piwowar et al., 2018). The small increase was partly due to the slow but steady growth of popularity of hybrid Open Access arrangements. Hybrid type of Open Access is an arrangement in which the authors are offered to buy the Open Access status of their article (see Table 2 for Open Access classifications). The general prevalence of this format is difficult to estimate, but in 2019 the Open Access share of Elsevier hybrid journals reached only 3.7% (Jahn et al., 2022).

Table 2. Main classification Terms in Types of Open Access

Open Access Classification	Description
Diamond	readers or authors are not charged for immediate Open Access
Gold	published in an Open Access journal, but authors need to pay for it
Green	paywalled on the publisher site but available in an Open Access repository
Bronze	freely readable on the publisher site but without associated license
Hybrid	Open Access can be bought, otherwise paywalled
Delayed	freely available only after an embargo period
Black	openly shared only on pirate sites

Open Access fees vary. While in 2018, no fee was more than US\$913 (Crawford, 2019), today payments can be significantly higher. For example, the *Proceedings of the National Academy of Sciences* charges up to US\$ 4,215 per article for processing with a surcharge of US\$4,975 for immediate Open Access<sup>2</sup>. *Nature* authors are also offered to make their work freely accessible for a sum of US\$11,390 (Else, 2020). For many journals, additional fees apply for submission<sup>3</sup> or for requests such as colour figures and extra pages<sup>4</sup>.

This state of academic publishing received heavy criticism from both the research community and the public. As a start, any limitation to free access to scientific knowledge is against the ethos of science in which scientific findings should be public property (Merton, 1973). Open Access to scientific knowledge is also a moral obligation towards global human equality. Article 27 of the *United Nations Declaration of Human Rights* states that " Everyone has the right to freely participate in the cultural life of the community, to enjoy the arts and to share in scientific

<sup>2</sup> <https://www.pnas.org/author-center/publication-charges>

<sup>3</sup> <https://web.archive.org/web/20150516032726/http://www.econ.ucsb.edu/~tedb/jfees.html>

<sup>4</sup> <https://web.archive.org/web/20190804021107/https://sites.agu.org/publications/files/2014/08/pubfeetablefinalAug2014.pdf>

advancement and its benefits." (United Nations, 1948). Most scientific articles are products created from public funding by researchers paid by public institutes. In the traditional publication system, the right of this public value lands in the hands of publishers who can freely tag them with market price. A negative consequence of this arrangement is that less affluent institutions, scholars, and members of the public have no free and legal access to the vast majority of scientific knowledge. The increase of article processing charges similarly create unequal access for researchers to publish their scientific work (Jain et al., 2020).

Articles with Open Access, however, come with a number of advantages. As a start, they are readable by a much wider audience than by paywalled papers. As a consequence, they can make more impact, facilitate innovations, and they are more often cited. The rate of open access citation advantage is difficult estimate, but the calculations put the figure between 5% and 83% (J. A. Evans & Reimer, 2009; Hua et al., 2017; Langham-Putrow et al., 2021). It's important to add that increased access to scientific knowledge can also foster scientific education and literacy (Zuccala, 2010) as well as public policy (European Commission, 2012).

Another expected benefit of online publication formats (Harter & Kim, 1997) was that the delay from submission to publication could radically decrease. As printing and posting the manuscripts are avoidable and that editing the text can be easily done in word processor templates, the workflow is expected to be simpler and faster. In 1980, the average publication delay, from submission to publication, between 25 journals was found to be 18.9 months (Yohe, 1980). In 2013, Björk and Solomon sampled 135 journals (Björk & Solomon, 2013) and found that for Business/economics journals this time was 18 months, for chemistry papers 9 months. At least among medical journals, the average turnaround time for journal articles did not change since 2013 (Christie et al., 2021; Horbach, 2020). The extreme delays, however, became less frequent since the early 2000s (Himmelstein & Powell, 2021). Publication delay has a number of negative consequences. It is a certain source of frustration for the authors (Ross-Hellauer et al., 2017) affecting their career advancement and funding opportunities. Furthermore, publication delay negatively affects the accuracy of the impact factor of the journal, as the delay can invalidate citations (Guo et al., 2021; Shi et al., 2017).

Another frustration point for the authors of journal articles is the formatting and administration of their submissions. As formatting and submission requirements differ among journals, authors often need to spend a lot of time with manuscript submission. A survey among researchers found that resubmission requirements is a major hinder to progress. Reformatting can

delay publication by at least two weeks but not rarely over three months (Jiang et al., 2019). The extra effort has a clear economic price with an estimated annual US\$ 1.1 million accounting for a research team's time – which is estimated to be more than 1.5 million hours every year (Khan et al., 2018).

Peer review is an inherent part of academic publishing. Its main function is to uphold quality standards so that nothing should be published that doesn't satisfy the expectations of the field and journal. These expectations can be theoretical, methodological, can relate to robustness, relevance, or novelty, according to the policy of the given journal. "Peer" refers to the practice that the reviews are typically performed by an invited professional with relevant competencies. Who a peer is to an author is hard to define. Should the peer be from the same discipline? Should they be an active researcher of the same topic with the same methodology? Can a reviewer be expert of only one aspect of the paper? These questions are famously unanswered (Smith, 2006). Similarly, it's hard to define what counts as a review and what difference it achieves in the submitted manuscript. From an empirical viewpoint, one should be able to tell from a manuscript whether it has been peer reviewed or not. Would anyone notice if the editor swapped the 'publish' and 'reject' collections? – was the provocative question of Robin Fox, editor of *Lancet*. The comments of other esteemed editors only support our doubt. Richard Smith, the editor of *British Medical Journal*, jokingly said: "When I was editor of the BMJ I was challenged by two of the cleverest researchers in Britain to publish an issue of the journal comprised only of papers that had failed peer review and see if anybody noticed. I wrote back 'How do you know I haven't already done it?'" (Smith, 2006, p. 178). Systematic reviews are not more informative. Jefferson and colleagues (Jefferson et al., 2002) reviewed the pre-2000 studies of peer review system and found that out of the little they tell about the system the effects remain uncertain. Peer review, therefore, is mostly based on faith in the system rather than facts (Peh, 2022; Smith, 2015).

### ***Solutions to the Problems in the Publication System***

Despite the continuous effort to reform the academic publication system, the achieved changes are nowhere near radical. Only around one in four papers is freely available and the publication process is mostly unchanged for the majority of the journals. Still, there has been a number of innovative approaches with visible success. For example, journals, such as *F1000Research*<sup>5</sup> provide a rapid dissemination to authors with open peer review. In open peer

---

<sup>5</sup> <https://f1000.com/>

review, the articles are published first, and the invited reviewers' reviews are posted on the site along the original manuscript or its updated version. This journal published not just traditional manuscripts, but posters and presentation slides as well. Open Peer Review, however, is not a general practice. *Nature's* known experiment with this version of peer review was found to be discouraging in 2006 (Nature, 2006). They offered authors the Open Peer Review option for their non-desk-rejected manuscript, but only 5% of the authors were interested and they and the editor thought that the reviews brought little value to the assessment. A more recent, cross-disciplinary survey found that the majority of scholars support transparency and Open Science but they are against opening reviewer identities to authors as it would have negative effects to the process and junior researchers (Ross-Hellauer et al., 2017).

*PeerJ*<sup>6</sup> show another alternative to the traditional business model. In this open access journal, researchers can become lifetime members by a one-time membership fee that provides them the right to a reduced article processing fee every year for their whole life. Other “megajournals” such as *PLOS One* or *Scientific Report* use a special type of review model where the review assesses only the scientific soundness of the manuscripts and does not make judgment over the potential contribution of the given study.

Archiving scholarly manuscripts in public repositories became common for most disciplines. Preprints mostly receive digital object identifiers (DOI), go through plagiarism checks, increase the chance of early feedback, and help authors gain early credit for their work (Callaway & Powell, 2016). Preprints are allowed or even encouraged by most journals<sup>7</sup> and funders started to support preprints as well (Callaway, 2017). Tools and procedures have been introduced to improve and speed up the publication process. For example, ASWG<sup>8</sup> provides a software that automatically checks for common problems in manuscripts related to transparency and reproducibility. Some repositories, such as PsyArxiv, facilitate the submission process by providing a direct submission option from the preprint server to APA journals<sup>9</sup>. At least for COVID-related studies, medical journals managed to significantly decrease the publication delay (Brierley et al., 2022; Horbach, 2020). An increasing list of journals<sup>10</sup> accept the initial submission without formatting requirements with the motto of “Put science first and formatting later” (Khan

---

<sup>6</sup> <https://peerj.com/>

<sup>7</sup> [https://en.wikipedia.org/wiki/List\\_of\\_academic\\_publishers\\_by\\_preprint\\_policy](https://en.wikipedia.org/wiki/List_of_academic_publishers_by_preprint_policy)

<sup>8</sup> <https://scicrunch.org/ASWG>

<sup>9</sup> <https://help.osf.io/article/188-submit-to-journal>

<sup>10</sup> <https://asntech.github.io/format-free-journals/>



et al., 2018). Some journals share the reviews of a manuscript in case it got resubmitted to another journals from the same publisher (called “cascading peer review”) (Maunsell, 2008). Peer Community In<sup>11</sup> solicits reviews of preprints and journals can consider these reviews for the publication of their manuscript.

## 1.2. Lack of Transparency

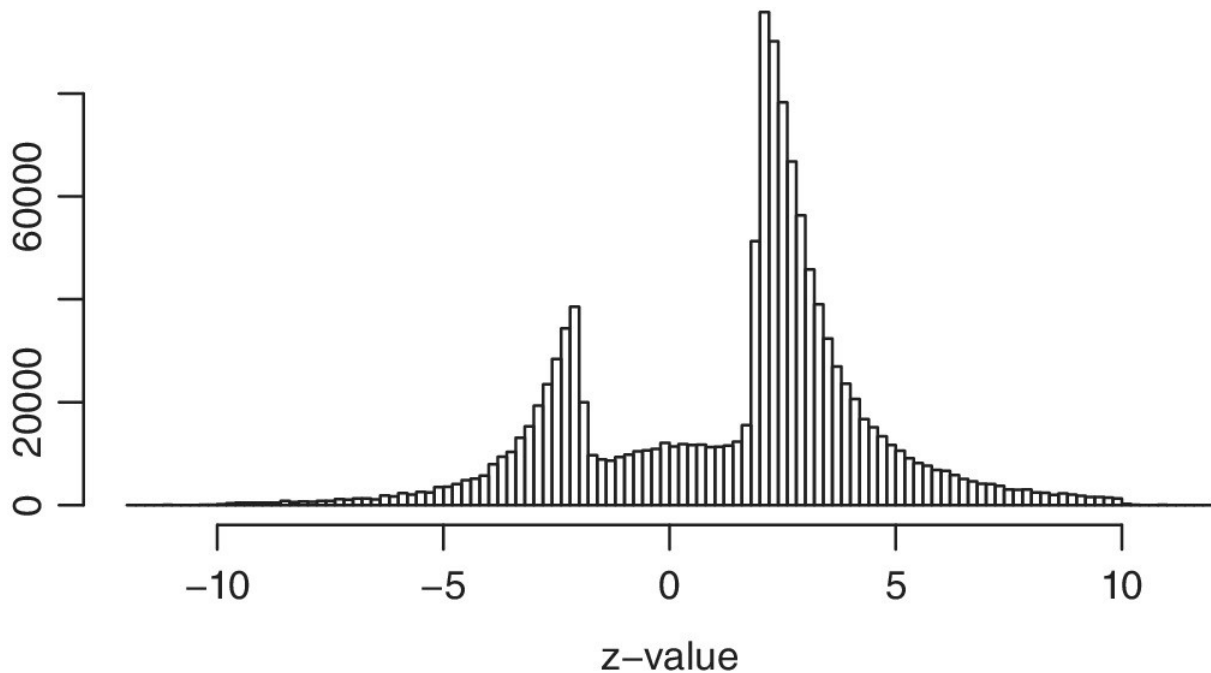
The Royal Society’s motto ‘*Nullius in verba*’ means ‘take nobody’s word for it’. It reflects the epistemological paradigm change that the Scientific Revolution and empiricism brought to Western thinking. Within empiricism, referring to authority lost from its convincing value, instead claims were required to build on observation and experimentation (Leahey, 2004). The success of early modern science was probably due, in large part, to its transparency. For example, Galilei, when describing his discovery of the moons of Jupiter in 1610, meticulously reported his day-by-day logs of current weather, telescope properties, timing, methods, analysis, and conclusions in his publication (Galilei, 2016). Robert Boyle, the father of modern chemistry, was not only keen to painstakingly report all details of his successful and failed experiments but he also emphasised the importance of replicating the observations in front of his peers (Bishop & Gill, 2020). He also insisted on the proper referencing of claims in scientific writing to link knowledge to its source (Boyle, 1661/1911). Today, transparency is a central obligation of scientists. Not just research institutes and journal policies but also ethical codes, such as the Declaration of Helsinki (World Medical Association, 2001), require researchers to publish complete and accurate reports. Nevertheless, the lack of transparency in all aspects of science is a generally identified problem as important information remains hidden in all levels of knowledge creation (Hardwicke et al., 2020).

Publication bias is a major source of lack of transparency in science. It refers to selective reporting in the publication system: certain characteristics of research findings increase the chances to get submitted and published than others (Ioannidis, 1998). For example, statistically significant findings have 2.2-4.7 higher odds of being fully reported compared to non-significant findings (Dwan et al., 2008, also see Figure 1). Similarly, strong results (where all hypotheses are supported by statistical tests) were found to be 40% more likely to be published (Franco et al., 2014). These patterns can be traced back not just to the selective taste of journals but also to researchers’ reluctance to write-up or submit results that are not newsworthy, thus creating a file drawer effect (Rosenthal, 1979). A less-known mechanism of publication bias is in the observation

---

<sup>11</sup> <http://peercommunityin.org/>

that non-English speakers make their decision whether to publish in an English-language journal or in their native language based on their findings (Egger et al., 1997; Jüni et al., 2002). It's easy to see how these mechanisms of publication bias provide the reader an inaccurate picture and can disadvantageously impact systematic reviews (Jüni et al., 2002) or meta-analyses (Page et al., 2021).



*Figure 1.* The distribution of more than one million z-values from Medline (1976–2019). The figure demonstrates the relatively low frequency of non-significant z-values compared to the significant ones. Adopted from van Zwet & Cator (2021). License: CC BY-NC-ND 4.0

An important way to follow Galilei and Boyle is to comprehensively and accurately report all relevant details of an empirical project. In social and behavioural sciences, researchers are trained to report, among others, sample sizes, participant exclusions, demographic information, effect sizes but in practice not all details are listed in the published articles. Meta-researchers identified a number of repeating reporting issues. Bakker and Wicherts (2011), for example, found that 18% of statistical results were incorrectly reported in psychological journals with around 15% containing incorrect statistical conclusions. In another study, Nuijten and colleagues (2016) screened 250,000  $p$ -values reported in eight major psychological journals and found some inconsistencies in half of them with one in eight affecting the statistical conclusion. Selective reporting of results or experiments with statistically significant results has been repeatedly detected (Chan et al., 2004; John et al., 2012). In general, many studies found reporting research

outcomes to be poor (Avey et al., 2016; Carp, 2012; Goldacre et al., 2019) with often insufficient details about the applied statistical analyses (Counsell & Harlow, 2017).

Lack of transparency in science can also emerge when the research data, analysis code, materials, and protocols are not shared along the publication of findings. Merton's ethos of science clearly states that research findings are not private but public property:

“The substantive findings of science are a product of social collaboration and are assigned to the community. [...] Property rights in science are whittled down to a bare minimum by the rationale of the scientific ethic. [...] The institutional conception of science as part of the public domain is linked with the imperative for communication of findings. Secrecy is the antithesis of this norm; full and open communication its enactment.” (Merton, 1973, pp. 273–274)

Open access, open code, and open materials are propagated not just for their intrinsic values but also for practical reasons. As a start, sharing data makes it possible for other scientists to independently verify the published findings or to conduct robustness analyses. Also, sharing data makes secondary data analysis possible: using the original data for answering different research questions or conducting meta-analyses. Sharing data in public repositories can also prevent data loss. Sharing analysis scripts and code makes reviewers' and readers job easier if they wish to fully understand the details of the published analysis or just check it for errors. Just as sharing analysis code, sharing materials can make science more efficient if not all researchers have to develop these when preparing a similar study.

When legal or ethical constrains allow, data sharing should not be a difficult task. Public repositories, such as Open Science Framework<sup>12</sup> or GitHub<sup>13</sup>, offer free storage space and assistant services. Data librarians and data stewards help researchers in many institutions on licensing and managing their data and analysis scripts. In practice, however, researchers' reluctance of sharing data is a major obstacle of transparent science. In psychology, the issue is repeatedly investigated. In 2006, Wicherts and colleagues sent out email requests to obtain datasets of 141 published empirical articles (Wicherts et al., 2006). After six month and 400 emails, they received only 38 positive reactions and datasets for 64 studies. Another study assessed the public availability of research data of articles published in high-impact journals from 2009 and found that a mere 9%

---

<sup>12</sup> <https://osf.io/>

<sup>13</sup> <https://github.com/>

deposited full primary research data online (Alsheikh-Ali et al., 2011). Gabelica (2022) and colleagues asked authors of 1,792 biomedical papers to share their study data – as they indicated in their paper to share them upon request. After all, only 6.7% of them fulfilled what they promised in their Data Availability Statement and shared usable data (C. Watson, 2022). Data sharing is becoming more difficult with time. It was found that data availability declines with article age mostly to do with obsolete storage devices or non-working emails addresses (Vines et al., 2014). One would think that at least the data of the most important findings should be available. Data Ark is an attempt to preserve the datasets the most-cited recent articles of psychological science and psychiatry (Hardwicke & Ioannidis, 2018). Despite all efforts, the authors managed to obtain only 14% of the datasets without restrictions and 68% of them remained completely inaccessible.

Houtkoop and colleagues (2018) attempted to explore why researchers are so reluctant to make their research data fully accessible. Survey data from 600 authors of psychological articles indicated that the main barriers are that sharing is not a common practice in their fields, their preference to share data only upon request, their perception that sharing requires extra work, and their lack of training in sharing data.

### ***Solutions to the Problems in Transparency***

Acceptance and willingness of data sharing show an increase over the years (T. Evans, 2022; Tedersoo et al., 2021; Tenopir et al., 2015) probably due to the large number of initiatives to propagate the change. Research communities, such as the Psychological Science Accelerator (Moshontz et al., 2018), the Framework for Open and Reproducible Research Training (Azevedo et al., 2019), or the Center for Open Science<sup>14</sup> have dedicated programs to encourage and educate open science practices. A more top-down influence to increase data sharing can come from funders and publishers. While some research funders already require data sharing (Kozlov, 2022) journals join initiatives such as the Transparency and Openness Promotion Guidelines (Aalbersberg et al., n.d.; Nosek et al., 2015) that set levels of openness that journals can require from authors. Some journals award badges to articles with Open Data/Code/Materials (Kidwell et al., 2016). A number of transparency guidelines also help researchers make their move towards Open Science easier (Crüwell et al., 2019; O. Klein et al., 2018; Wagenmakers et al., 2021; M. D. Wilkinson et al., 2016). Guidelines exist for specific types of research fields<sup>15</sup>. For example, the CONSORT

---

<sup>14</sup> <https://www.cos.io/>

<sup>15</sup> For their collection, see the EQUATOR website: <http://www.equator-network.org/>

statement (Schulz et al., 2010) is tailored to randomized controlled trials, the PRISMA statement (Moher et al., 2009) targets meta-analyses. There are journals, dedicated to publishing data (e.g., *Scientific Data*) and software (e.g., *Journal of Open Source Software*). Checklists also help researchers in reporting their data availability (e.g., Aczel, Szaszi, et al., 2020). Reviewers can also promote data availability throughout the peer-review system. The Peer Reviewers' Openness Initiative (Morey et al., 2016), for example, suggests that reviewers offer in-depth review only to manuscripts where the data are openly shared or the authors explicitly declare overriding limitations to data transparency. Finally, institutes can also play an important role by mandating some levels of transparency from their employees, for example data availability statement is an achievable level from any empirical work<sup>16</sup>.

### 1.3. Issues in Statistical Practice

Doing quantitative research in life and social scientists requires a range of skills one of them is being able to conduct statistical analyses on all sorts of datasets using different types of statistical methods. These analyses are then reviewed by other researchers if the study reaches the peer review part of the publication system. Typically, these researchers do not have expert statistical training. Even biomedical journals, where statistical experts are enlisted for review, conduct statistical reviews only occasionally (Hardwicke & Goodman, 2020). As a result, it is perhaps not surprising that the traditional statistical practice in these fields received heavy criticism throughout the years.

The first major problem in this area are the misinterpretations and misuses of statistical concepts. *P* values and statistical significance, for example, is a central topic here as positive findings are most often claimed from statistically significant results. Famously, Oakes (1986) asked psychologists with at least two years of research experience to tell whether statements such as this are true or false when interpreting a  $p = .01$  result:

“You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.” (p. 80)

Sixty percent of them chose the wrong “true” answer. This so called replication delusion (Keren & Lewis, 1993) can be misleading if it makes them believe that there is no point to replications

---

<sup>16</sup> For an example, see: <https://www.ppk.elte.hu/openscience>

when  $p$  is small. It is easy to see that this reasoning is wrong if we agree that one who threw two consecutive sixes with a die should not assume to get similar results in the next throws (Gigerenzer, 2018). This question was explored among psychologists in numerous occasions and the misunderstanding seems to be persistent (for a summary see Gigerenzer, 2018).

Oakes (1986) also found a list of other misinterpretations of  $p$  value. For example, it turned out that many psychologists believe that  $p = .01$  can mean one of these are true:

- “(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).
- (2) You have found the probability of the null hypothesis being true.
- (3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
- (4) You can deduce the probability of the experimental hypothesis being true.
- (5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.” (p. 80)

In fact, these are all false.  $P$  value indicates the probability of the observed result, plus more extreme results, if the null hypothesis were true and all the underlying assumptions were met (Wasserstein & Lazar, 2016). Nothing else. Every attempt to give more meaning to the  $p$  value result is mistaken. The actual value of  $p$  depends on a number of factors other than the studied effect. For example, we obtain different  $p$  value for the same effect if we increase or decrease our sample size. Also, the interpretation of  $p$  value depends on a range of circumstantial factors. For instance, two researchers analysing the same question on the same data would have different results if one analysed all of the data together and the other one looked into the results for half of the data before analysing the rest. Similarly, if one had multiple comparisons or multiple tests, she would need to adjust the significance threshold to control for type-I error. In short, the actual  $p$  value of an analysis depends on several circumstances of the analysis that are not indicators of the effect. In practical terms, the  $p$  value indicates only whether we can reject the null hypothesis or not, but we cannot use its value to infer to the strength, probability, or reliability of the effect (Dienes, 2008). As one result, it cannot be called “very significant” or “marginally significant” as these are all misuses of the term (Pritschet et al., 2016).

Another difficulty with the use of  $p$  values occurs when their value is above the alpha threshold, usually .05. These values indicate non-significant findings, but their interpretation is very problematic. Unfortunately, the so called null-hypothesis significance testing approach (NHST, Fisher, 1925; Neyman & Pearson, 1933) is not symmetric in its use; non-significant results do not relate back to the hypothesis. With a significant  $p$  value, we are entitled to reject the null hypothesis, but we cannot claim support for it when they are non-significant (Nickerson, 2000). A result can be non-significant either because the null hypothesis is true or because the test was not sensitive enough to detect the effect. The test does not indicate which case is true.

The question is then what to do when our result turns out to be non-significant. Although the statistical framework does not allow it and the American Psychological Association's (2001) publication manual clearly warns against it, researchers still try to claim evidence for the null from non-significant results. A 2006 study found that nonsignificant effects were interpreted as claims of no effect in 60% of cases in a leading psychological journal. Our own study (Aczel, Palfi, et al., 2018) found that for the leading psychological journals this value was 72%. Another wrong solution is to find ways to transform otherwise nonsignificant  $p$  values into significant  $p$  values. This can be done in multiple ways such as adding/discarding data until the  $p$  is under .05, rounding or misreporting the value to make it look significant, or neglecting the correction required after multiple testing. These questionable research practices unfortunately exist and are well documented in the literature (Hartgerink et al., 2016; Lilienfeld & Waldman, 2017; Nuijten et al., 2016; Pritschet et al., 2016). The correct reporting of non-significant results is to state that the analysis could not reject the null hypothesis. Any further claim is speculation or misinterpretation (Dienes, 2008, 2014; S. Goodman, 2008).

A similarly misunderstood statistical concept is the confidence interval. Confidence intervals are expected to be reported as part of the results of hypothesis tests (Finch et al., 2002; L. Wilkinson, 1999) and some researchers prefer to draw inference from them rather than  $p$  values (Cumming, 2014). Although, researchers seem to interpret confidence interval results more intuitively and correctly (Fidler & Loftus, 2009; Hoekstra et al., 2012), they are still confused about its meaning. Hoekstra and colleagues (2014) showed a fictitious scenario to researchers and students about the results of a professor. The 95% confidence interval of the results ranged from 0.1 to 0.4. The task was to decide which of the listed statements are true, concerning what we learned from this result. For example:

“There is a 95 % probability that the true mean lies between 0.1 and 0.4.

We can be 95 % confident that the true mean lies between 0.1 and 0.4.

If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.” (p. 1060)

Although, all statements were incorrect, researchers believed each to be true in 30-86%. The first issue is that people, even with statistical education, interpret the confidence interval as property of the parameter, when it is only a property of the procedure. Having confidence intervals from a sample means that repeating the procedure computing the confidence intervals across a series of hypothetical datasets would yield intervals that would contain the true parameter 95% of the cases (Hoekstra et al., 2014). The second issue occurs if confidence intervals are read as the probability that the true value is within the interval. That would be incorrect (Berger & Wolpert, 1988) as it would regard this frequentist concept from a Bayesian framework (Hoenig & Heisey, 2001).

Another central statistical concept that psychological researchers use is power. Statistical power is the probability of rejecting the null hypothesis given that it is false. Experiments with higher power have higher chance of detecting an effect if it exists. Underpowered studies, however, can correctly detect only large effects. Psychologists are urged to maximize the power of their studies (Munafò et al., 2017) which can be achieved by optimising the design of the experiment or increasing the sample size of the study. The expected power is conventionally 80% but journals can expect it to be as high as 95%<sup>17</sup>. Whereas social scientists are strict about keeping the type I error (rejecting the null hypothesis when it is true) at 5%, they are more relaxed with the control of power. It's long known that psychological studies suffer from low power. Analysing psychological studies published in 1960, Cohen (1962) found that their power was incapable of detecting any other than large effects ( $r \sim 0.60$ ). For the year 1984, a similar investigation found even weaker power in the analysed studies (Sedlmeier & Gigerenzer, 1989). More recently, the median power in neuroscience was found to be around 20% (Button et al., 2013), in genetic studies 8%, and in brain imaging it was 27% (Dumas-Mallet et al., 2017). The median sample size in four psychological journals was found to be 40 (Marszalek et al., 2011). If low sample sizes are combined with the authors' and publishers' preference for positive findings, then it becomes increasingly likely that the published findings are false positives or that the reported effect sizes are overestimates. That is, when the noise is great, only fluke findings or overestimated signals

---

<sup>17</sup> <https://www.nature.com/nathumbehav/submission-guidelines/registeredreports>



can be detectable. Among other issues of current scientific practice, this pervasive low power led to the famous conclusion that most published research findings are false (Ioannidis, 2005).

One explanation behind the negligence of power in many psychological studies might be its weak understanding. When asked, the great majority of psychological researchers seem to overestimate the power of researcher designs, underestimate the sample size they would need (Bakker et al., 2016; Vankov et al., 2014) and rarely discuss power then justifying their sample size (Bakker & Wicherts, 2011; Tressoldi & Giofré, 2015). Perhaps more importantly, the concept of power is convoluted in the general understanding. The main misunderstanding of power is that it is useful for planning experiments but less for evaluating particular findings. When statistical power is calculated for a given design and expected effect size, our measure refers to all possible outcomes of the experiment. When the data are obtained, this calculation loses its practical value, as we don't have to deal with hypothetically infinite unobserved datasets; all the information is within the data we collected (Wagenmakers et al., 2015). Therefore, it is fundamentally flawed to conduct post-experimental power calculations and use them for the interpretation of the results (Hoening & Heisey, 2001). In fact, the calculated post-hoc power is a one-to-one function of  $p$  value, therefore, it is, at best, redundant (Hoening & Heisey, 2001). Although methodologists and journals warn against the use of retrospective power analysis (Gilbert & Prion, 2016; S. N. Goodman & Berlin, 1994; Jiroutek & Turner, 2017; Psychonomic Society, 2012) one can easily find examples for its use in current literature.

All these three examples, the  $p$  value, the confidence interval, and the power analysis, show that researchers would like to assign some strength to their evidence or probability to the veracity of their hypotheses. The traditional statistical framework, however, builds on objective and not subjective probability. This framework, frequentism, uses probability as a persistent long-term frequency, such as the chances in roulette wheels. By definition, this type of probability cannot be used for single events. Subjective, or Bayesian probability, in contrast, is the quantification of one's personal belief (De Finetti, 2017), such as one's belief that it will rain overnight. Whenever researchers use frequentists statistics, they are in the realm of objective probability and the  $p$  values or confidence intervals that they calculate won't tell them how likely is that the hypothesis is true or false. Every attempt to interpret them that way violates the underlying principles of frequentist statistics (Dienes, 2008).

The so-called *crisis of confidence* (Pashler & Wagenmakers, 2012) in psychology and, in general, social sciences stems not just from the apparent misuses and misinterpretations of

statistical methods but also from the realisation of a number questionable research practices (Fiedler & Schwarz, 2016; John et al., 2012). Hypothesising after the results are known (known as HARKing, Kerr, 1998), *p*-hacking (John et al., 2012), outcome reporting bias (Fanelli, 2012; Mazzola & Deuling, 2013), or data fabrication (Fanelli, 2009) are just a few examples from their long list (Hall & Martin, 2019). A common element among them is the opportunistic use of the so-called researcher degrees of freedom (Simmons et al., 2011). This kind of freedom refers to the choices that researchers have at formulating the hypothesis, designing a study, running the experiments, collecting and analysing data, or reporting the findings. The existence of this freedom reflects not researchers' preference but rather the lack of robust phenomena and the weakness of theoretical constraints in psychological science (Eronen & Bringmann, 2021). Wicherts and colleagues (2016) identified 34 degrees of freedom that researchers have, for example, choosing between different options when dealing with incomplete or missing data, choosing the estimation method or inference criteria when analysing data.

Researchers with the best skills and intentions cannot avoid making choices during their data analyses. A new approach of statistical metascience aims to explore how much these choices matter for the results and conclusions. The question would not carry weight if legitimate data handling methods and statistical analyses of the same data would lead to the same conclusions. Most explorations of the topic, however, indicate that this so-called analytical robustness does not hold in many areas of behavioural and social sciences. A good example is Botvinik-Nezer and colleagues' (2020) study in which 70 independent teams were asked to analyse the same neuroimaging dataset for 9 hypotheses about brain activity in a risky-decision task. They found no two teams that followed the same analysis workflow resulting in substantial differences in their conclusions. Similar results were found in areas of health care (Bastiaansen et al., 2020), psychology (Boehm et al., 2018; Dutilh et al., 2019; Hoogeveen et al., 2022; Schweinsberg et al., 2021; Silberzahn et al., 2018; Starns et al., 2019), economics (Huntington-Klein et al., 2021), finance (Menkveld et al., 2021), sociology (Salganik et al., 2020), and medicine (Veronese et al., 2021). These findings question whether we can safely assume that the published results in social sciences are analytically robust so that other analysts, following a similarly valid statistical path, would have not arrived at different conclusions. All these studies highlight that the multiplicity of analysis strategies (Hoffmann et al., 2021) is another important aspect of the current statistical practice that cannot be neglected if our aim is to increase the trustworthiness of social science studies.

### *Solutions to the Problems in Statistical Practice*

One hope that the situation in statistical practice will improve can come from the fact that in the last years the stakeholders of science became aware of these issues and started making steps towards some changes. Journals and funders, for example, who play an important role here, started updating their policies and introducing checklists or reporting guidelines. A good example is the *Nature Life Sciences Reporting Summary* (Campbell, 2013) checklist that makes it mandatory for submitters to answer a list of questions on statistical and methodological practices followed in the study before they submit their manuscript to the journal. *Nature Human Behaviour* advertised Bayesian analysis and put the bar fairly high for sample size estimations in registered reports<sup>18</sup>. In addition, free resources are widely available on how to improve statistical inference in ebooks (e.g., Lakens, 2022; Poldrack, 2018), teaching materials<sup>19,20</sup>, or open software (e.g., JASP, Love et al., 2019).

The increased popularity of Bayesian analysis among social scientists (Van de Schoot et al., 2017) facilitated debates and some changes in statistical practice. A fundamental difference between the classical frequentist approach and the Bayesian approach is that the former allows us to draw conclusions only about the probability of the data given the theory. In other words, the results can tell us only how likely the obtained (or more extreme) data are if we assume the theory to be true (Wasserstein & Lazar, 2016). In contrast, Bayesian statistics are concerned about the probability of the theory in light of the obtained data (Etz, 2018). Therefore, from the latter we can claim how much our data support a given theory. Dienes argues that psychologists' interest in statistics is to be able to tell how much the data should change their belief about the theory, and therefore, Bayesian statistics should be applied to psychological questions (2008). He also speculates that many issues in statistical practice might come from this motivation that frequentist statistics cannot validly satisfy (2008). The proponents of Bayesian statistics add a number of other reasons why Bayesianism could be practical for psychologists. The first one, connecting to the central characteristics of Bayesian statistics, is that within this framework the strength of the evidence can be quantified. Bayes factor (BF) indicates how much the data favours one hypothesis over another one (Dienes, 2011; Morey & Rouder, 2011). Although the underlying mathematical calculations are complex, Bayes factor is simply the odds ratio of the likelihood of data given  $H_1$  and the odds ratio of the likelihood of data given  $H_0$  (Etz & Vandekerckhove, 2018). When the

---

<sup>18</sup> <https://www.nature.com/nathumbehav/submission-guidelines/registeredreports>

<sup>19</sup> <https://osf.io/t56kg/>

<sup>20</sup> <https://jasp-stats.org/teaching-with-jasp/>

Bayes factor is 1, it means that the observed data cannot differentiate between the two models, the results are inconclusive. Farther the Bayes factor is from 1, the more the data support one or the other hypothesis. Researchers, however, cannot leave everything to algorithms when doing Bayesian analysis because in order to gain meaningful results one needs to specify the hypotheses. In other words, one needs to define the prior distribution of the hypothesis. In the eyes of many critics of Bayesianism, this aspect brings in subjectivism to the calculation but its defenders argue that specification improves inference (Rouder, Morey, Verhagen, et al., 2016) and the fact that the answer depends on the question should not be such a surprise (Rouder, Morey, & Wagenmakers, 2016).

Psychologists adopted certain thresholds for evidence strength such as Bayes factor 1-3 is anecdotal evidence, 3-10 substantial evidence, >10 strong evidence (Wetzels et al., 2011). *Nature Human Behaviour*, for example, expects Bayes factor to be at least 10 times in favour of the one hypothesis over the other in registered report sample size calculations<sup>21</sup>. What  $p$  value and confidence intervals cannot provide; Bayes factors can help the researchers claim how much their results support their theory.

Another practical advantage of Bayesian statistics in psychology is that optional stopping is not a problem for Bayesians (Rouder, 2014). Whereas in classical statistics, one cannot keep analysing the data as they come in and decide whether to collect more data, in Bayesian statistics this does not cause issues as data can be collected until they reach one of the evidence thresholds (Etz et al., 2018). A final argument in favour of this approach is that the evidence can support not just the alternative hypothesis but also the null (Dienes, 2008, 2014).

The strength of evidence in psychology can be enhanced, of course, not just by extending the statistical toolbox, but also by improving the research designs and increasing the power of the studies. One simple way to increase the power of any study is by increasing its sample size. As discussed above, the median sample size in many areas of psychology is fairly low. In the past, hard to reach populations and institutional capacities often limited researchers to obtain larger sample sizes. One positive move in this regard is the increasing number of collaborations and the presence of the so-called big team science (Coles et al., 2022; Koch & Jones, 2016). An exemplary case is the *Psychological Science Accelerator* (Moshontz et al., 2018) which is a globally distributed network of research labs, presently from 84 countries. Their aim is to accelerate

---

<sup>21</sup> <https://www.nature.com/nathumbehav/submission-guidelines/registeredreports>

knowledge accumulation by supporting and organizing large-scale data collections for applied and theoretical questions in psychology. Our lab led one of these projects, testing a moral dilemma question collecting data from 27,502 participants from 45 countries (Bago et al., 2022). Other examples, such as ManyBabies (Byers-Heinlein et al., 2020) or ManyLabs (R. A. Klein et al., 2018) projects show how hundreds of researcher can work together to provide strong answers to important questions and to decrease the homogeneity of the data coming from WEIRD (Western, Educated, Industrialised, Rich and Democratic) societies (Muthukrishna et al., 2018).

(References for the Introduction and Summary sections are at the end of the thesis)

## 2. PUBLICATION PRACTICE

### 2.1. A Billion-Dollar Donation: Estimating the Cost of Researchers' Time Spent on Peer Review<sup>22</sup>

Balazs Aczel<sup>a\*</sup>, Barnabas Szaszi<sup>a\*</sup>, Alex O. Holcombe<sup>b</sup>

<sup>a</sup>Institute of Psychology, ELTE, Eotvos Lorand University, Budapest, Hungary

<sup>b</sup>School of Psychology, University of Sydney, Sydney, Australia

**Keywords:** peer-review, academic publishers, publication system

---

<sup>22</sup> published as:

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1), 1–8.

## **Abstract**

### Background

The amount and value of researchers' peer review work is critical for academia and journal publishing. However, this labor is under-recognized, its magnitude is unknown, and alternative ways of organizing peer review labor are rarely considered.

### Methods

Using publicly available data, we provide an estimate of researchers' time and the salary-based contribution to the journal peer review system.

### Results

We found that the total time reviewers globally worked on peer reviews was over 100 million hours in 2020, equivalent to over 15 thousand years. The estimated monetary value of the time US-based reviewers spent on reviews was over 1.5 billion USD in 2020. For China-based reviewers, the estimate is over 600 million USD, and for UK-based, close to 400 million USD.

### Conclusions

By design, our results are very likely to be under-estimates as they reflect only a portion of the total number of journals worldwide. The numbers highlight the enormous amount of work and time that researchers provide to the publication system, and the importance of considering alternative ways of structuring, and paying for, peer review. We foster this process by discussing some alternative models that aim to boost the benefits of peer review, thus improving its cost-benefit ratio.

## Background

One of the main products of the academic publication system, the journal article, is a co-production of researchers and publishers. Researchers provide value not only by doing the research and writing up the results as a manuscript, but also by serving as peer reviewers. Publishers provide services of selection, screening, and dissemination of articles, including ensuring (proper) meta-data indexing in databases. Although several careful estimates are available regarding the cost of academic publishing (e.g., 1), one aspect these estimates often neglect is the cost of peer reviews (2). Our aim was to provide a timely estimation of reviewers' contribution to the publication system in terms of time and financial value and discuss the implications.

In their peer reviewer role, scientists and other researchers provide comments to improve other researchers' manuscripts and judge their quality. They offer their time and highly specialized knowledge to provide a detailed evaluation and suggestions for improvement of manuscripts. On average, a reviewer completes 4.73 reviews per year<sup>23</sup>, yet, according to Publons<sup>24</sup>, certain reviewers complete over a thousand reviews a year. This contribution takes considerable time from other academic work. In the biomedical domain alone, the time devoted to peer review in 2015 was estimated to be 63.4M hours (3).

A manuscript typically receives multiple rounds of reviews before acceptance, and each round typically involves two or more researchers as peer reviewers. Peer review work is rarely formally recognized or directly financially compensated in the journal system (exceptions include some medical journals that pay for statistical reviewers and some finance journals that pay for quick referee reports). Most universities seem to expect academics to do review work as part of their research or scholarly service mission, although we know of none with an explicit policy about how much time they should spend on it.

While peer review work is a critical element of academic publishing, we found only a single estimate of its financial value, which was from 2007. Then, when the global number of published

---

<sup>23</sup> Based on Personal communication with the Publons team.

<sup>24</sup>

[https://publons.com/researcher/?is\\_core\\_collection=1&is\\_last\\_twelve\\_months=1&order\\_by=num\\_reviews](https://publons.com/researcher/?is_core_collection=1&is_last_twelve_months=1&order_by=num_reviews)



articles was not even half of the present volume, rough estimates indicated that if reviewers were paid for their time, the bill would be on the order of £1.9bn (4).

As a facet of the research process that currently requires labor by multiple human experts, reviewing contributes to a cost disease situation for science. “Cost disease” (5) refers to the fact that while the cost of many products and services have steadily decreased over the last two hundred years, this has not happened for some for which the amount of labor time per unit has not changed. This can make some products and services increasingly expensive relative to everything else in society, as has occurred, for example, for live classical music concerts. This may also be the fate of scholarly publication, unless reviewing is made more efficient.

The fairness and efficiency of the traditional peer review system has recently become a highly-debated topic (6–7). In this paper, we extend this discussion by providing an update on the estimate of researchers’ time and the salary-based contribution to the peer-review system. We used publicly available data for our calculations. Our approximation is almost certainly an underestimate because not only do we choose conservative values of parameters, but for the total number of academic articles, we rely on a database (Dimensions) that does not purport to include every journal in the world. We discuss the implications of our estimates and identify a number of alternative models for better utilizing research time in peer review.

## **Methods and Results**

To estimate the time and the salary-based monetary value of the peer review conducted for journals in a single year, we had to estimate the number of peer reviews per year, the average time spent per review, and the hourly labor cost of academics. In case of uncertainty, we used conservative estimates for our parameters, therefore, the true values are likely to be higher.

### Coverage

The total number of articles is obviously a critical input for our calculation. Unfortunately, there appears to be no database available that includes all the academic articles published in the entire world. Ulrich’s Periodicals Database may list the largest number of journals - querying their database for “journals” or “conference proceedings” and “Refereed / Peer-reviewed” yielded 99,753 entries. However, Ulrich’s does not indicate the number of articles that these entities publish. Out of the available databases that do report the number of articles, we chose to use

Dimensions' dataset (<https://www.dimensions.ai/>) which collects and collected articles from 87,000 scholarly journals, much more than Scopus (~20,000) or Web of Science (~14,000) (8).

### *Number of peer reviews per year*

Only estimates exist for how many peer reviews associated with journals occur each year. Publons (9) estimated that the 2.9 million articles indexed in the Web of Science in 2016 required 13.7 million reviews. To calculate the number of reviews relevant to 2020, we used the formula used by Publons (9) - equation 1 below. In that formula, a review is what *one* researcher does in *one* round of a review process<sup>25</sup>. For submissions that are ultimately accepted by the journal submitted to, the Publons formula assumes that on average there are two reviews in the first round and one in the second round; for rejected articles (excluding desk rejections) the formula assumes an average of two reviews for submissions that are ultimately rejected, both in the first round. Publons' assumptions are based on their general knowledge of the industry but no specific data. Note, however, that if anything these are most likely underestimations as not all peer reviews are included in our estimation. For example, the review work done by some editors when handling a manuscript is not usually indexed in Publons, and a single written review report may be signed by several researchers.

Publons estimated the acceptance rate for peer-reviewed submissions to be 55%. That is, 45% of manuscripts that are not desk rejected are, after one or more rounds of review, ultimately rejected. Before including Publons' estimates in our calculations, we evaluated them based on other available information. The Thomson Reuters publishing company reported numbers regarding the submissions, acceptances, and rejections that occurred at their ScholarOne journal management system for the period 2005-2010 (10). In agreement with other sources (11,12), it showed that the mean acceptance rates have apparently declined (10), the proportion of submissions that are eventually accepted by the journal the manuscript was submitted at was 0.40 in 2005: 0.37 in 2010, and 0.35 in 2011 (11,12).

---

<sup>25</sup> Note, that there are cases when a single submitted review is prepared by more than one individual, but the used formula does not differentiate these cases from when a review is prepared by only one individual.

We did not find estimates of acceptance rates for the last several years, but we assume that the decline described by Thomson Reuters (10) continued to some extent, and assume that the present mean acceptance rate at journals is 0.30 then we can arrive at Publons' figures. However, for the final numbers, we also need to estimate the rate of desk rejections as well. Although the rate of desk rejections likely varies substantially across journals (e.g., 22-26% at PLOS ONE<sup>26</sup>), referenced values (13,14) and journal publisher estimates<sup>27</sup> lead us to estimate this value around 0.45.

The above estimates imply that, on average, every 100 submissions to a journal comprise 30 that are accepted after one or more rounds of peer review, 45 that are desk rejected, and 25 that are rejected after review. Thus, among submissions sent out for review, 55% (30 / (30 + 25)) are ultimately accepted. That is, the articles published represent 55% of all reviewed submissions, indicating that 45% of submissions that were reviewed were rejected. These values are undoubtedly speculative, but they are consistent with Publons' estimates.

Therefore, to estimate the number of peer reviews per year, we used Publons' (9) formula:

Equation 1:

$$\begin{aligned} &Nr\ of\ submissions_{accepted} \times Average\ Nr\ of\ reviews_{accepted} \\ &+ Nr\ of\ submissions_{rejected} \times Average\ Nr\ of\ reviews_{rejected} \end{aligned}$$

To obtain these values, we had to estimate the number of peer reviews performed for articles in 2020. For that, we used the numbers provided by the Dimensions portal ([www.dimensions.ai](http://www.dimensions.ai)). The free version as well as the subscription version of Dimensions currently provide separate numbers for articles, chapters, proceedings, preprints, monographs, and edited books. For the sake of simplicity, our estimate is confined to articles.

The total number of articles published in 2020 according to the Dimensions database is 4,701,988. Assuming that this sum reflects the 55% acceptance rate of reviewed submissions, the number of

---

<sup>26</sup> <https://journals.plos.org/plosone/s/journal-information>

<sup>27</sup> <https://www.elsevier.com/authors-update/story/publishing-tips/5-ways-you-can-ensure-your-manuscript-avoids-the-desk-reject-pile>

reviewed but rejected submissions (the 45% of all reviewed submissions) are estimated to be globally  $4,701,988/55*45 = 3,847,081$ . Based on these calculations, the total number of peer reviews for submitted articles in 2020 is  $4,701,988*3 + 3,847,081*2 = 21,800,126$ .

### *Time spent on reviews*

Several reports exist for the average time a reviewer spends when reviewing a manuscript. All of these are unfortunately based on subjective reports by reviewers rather than an objective measure. The only thing resembling an objective indication we found was in the Publons dashboard (Publons.com), which as of 6 Aug 2021 indicated that the average length of reviews in their database across all fields is approximately 390 words. This highlights that the average review likely has substantive content beyond a yes/no verdict, but this cannot be converted to a time estimate. A 2009 survey responded to by 3,597 randomly selected reviewers indicated that the reported average time spent on the last review was 6 hours (15), a 2016 survey reported that the median reviewing time is 5 hours (9). Another survey in 2008 found that the average reported time spent reviewing was 8.5 hours (16). To be noted, it is likely that the second round of reviews do not take as long as the first one. To be conservative (and considering the tendency of people to overestimate how much time they work), we will use 6 hours as the average time reviewers spend on each review.

Based on our estimate of the number of reviews and hours spent on a review, we estimate that in 2020 reviewers spent  $21,800,126 \times 6$  hours = 130,800,757 hours on reviewing. This is equivalent to 14,932 years (at 365 days a year and 24 hours of labor per day) (Figure 1).

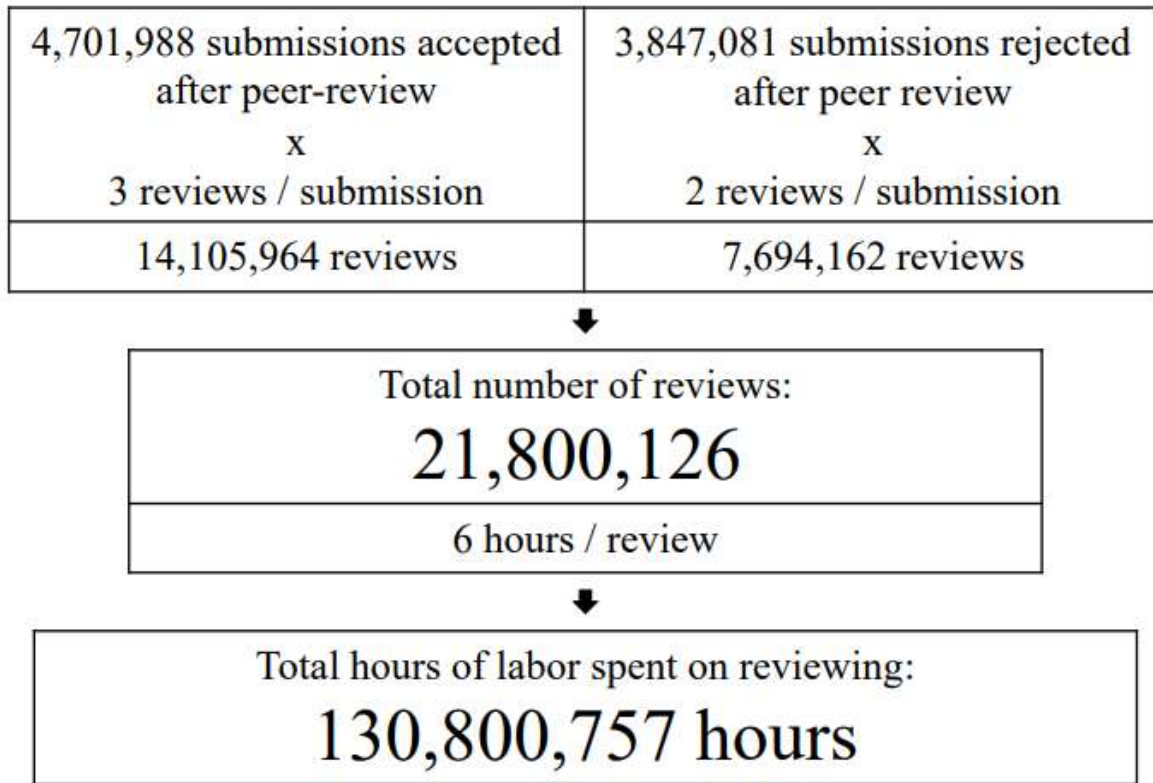


Figure 1. Overview of our calculation estimates of time spent on reviewing for scholarly articles in 2020. Number of published articles was obtained from Dimensions.AI database, all other numbers are assumptions informed by previous literature.

### *Hourly wage of reviewers*

To estimate the monetary value of the time reviewers spend on reviews, we multiplied reviewers' average hourly wage by the time they spend reviewing. Note that some scholars consider their reviewing work to be volunteer work rather than part of their professional duties (5), but here we use their wages as an estimate of the value of this time. No data seem to have been reported about the wages of journal reviewers, therefore, we require some further assumptions. We assumed that the distribution of the countries in which reviewers work is similar to the distribution of the countries in the production of articles. In other words, researchers in countries that produce more articles also perform more reviews, while countries that produce few articles also do proportionally few reviews. Given the English-language and geographically Anglophone-centered concentration of scientific journals, we suspect that people in English-speaking countries are called on as reviewers perhaps even more than is their proportion as authors (17). Because such countries have higher wages than most others, our assumption of reviewer countries being proportional to author

countries is conservative for total cost. Accordingly, we calculated the country contributions to the global article production by summing the total number of publications for all countries as listed in the Dimensions database and computing the proportion of articles produced by each country.

Based on the results of the Peer Review Survey (15) and to keep the model simple and conservative, we assumed that reviewing is conducted almost entirely by people employed by academic workplaces such as universities and research institutes and that junior and senior researchers participate in reviewing in a ratio of 1:1. Therefore, to calculate the hourly reviewer wage in a given country we used *Equation 2*:

$$\frac{\text{average annual post – doc salary} + \text{average annual full professor salary}}{2 \times \text{annual labor hours}}$$

This yields a figure of \$69.25 per hour for the U.S., \$57.21 for the UK, and \$33.26 for China (Table 1).

### ***Value of reviewing labor***

We estimated the value of reviewing by multiplying the calculated hourly reviewer wage in a country by the number of estimated reviews in that country and the time preparing one review. We calculated each country's share from the global number of reviews by using the country's proportional contribution to global production of articles. In this calculation, each article produced by international collaborations counts as one to each contributing country. This yielded that the monetary value of reviewing labor for the three countries that contributed to the most articles in 2020, is: \$USD 1.5 billion for the U.S., \$626 million for China, and \$391 million for the UK (Table 1). An Excel file including the formula used for the estimation in the present paper with interchangeable parameters is available at the OSF page of the project <https://osf.io/xk8tc/>.

Table 1

Estimating the Value of Review Labor for the US, China, and the UK for 2020

Parameter	US	China	UK
Annual postdoc salary	\$65 000	\$68 174	\$39 692
Annual professor salary	\$179 736	\$76 428	\$116 731

Annual labor hours	1 767	2 174	1367
Reviewer hourly wage	\$69.25	\$33.26	\$57.21
Articles	715 645	618 430	224 220
Contribution to global article production	16.68%	14.41%	5.22%
Reviews	3 636 031	3 141 908	1 139 106
Value of reviewing time	\$1 510 810 944	\$626 945 064	\$391 036 638

*Note.* Salary values were collected from <https://inomics.com/sites/default/files/2018-05/INOMICS%20Salary%20Report%202018.pdf> for the USA and the UK, and were downloaded on 2021.09.09. from <http://www.salaryexplorer.com/salary-survey.php?loc=44&loctype=1&job=50&jobtype=1#disabled> for China. To estimate the average full professor salary, we calculated the average of the 39 professor categories available at salaryexplorer.com. To convert the average Chinese salary to USD, we used the 2020 average exchange rates (6.90) from CNY to USD based on <https://www.macrotrends.net/2575/us-dollar-yuan-exchange-rate-historical-chart> (The calculations are available at the projects' OSF page). Note that we are concerned that the Chinese salaries may be inaccurate, based on anecdotal feedback we have received. For China, labor hours were found in <https://ourworldindata.org/working-hours>; for the USA and the UK they were retrieved from <https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS>. The numbers of articles published in 2020 for each country are from the Dimensions database. To calculate the value of reviewing time, we used the non-rounded form of the hourly wages.

## Discussion

The high price of scientific publishing receives a lot of attention, but the focus is usually on journal subscription fees, article processing charges, and associated publisher costs such as typesetting, indexing, and manuscript tracking systems (e.g., 1). The cost of peer review is typically not included. Here, we found that the total time reviewers worked on peer reviews was over 130 million hours in 2020, equivalent to almost 15 thousand years. The estimated monetary value of the time US-based reviewers spent on writing reviews was over 1.5 billion USD in 2020. For China-based reviewers, the estimate is over 600 million USD, and for UK-based, close to 400 million USD. These are only rough estimates but they help our understanding of the enormous

amount of work and time that researchers provide to the publication system. While predominantly reviewers do not get paid to conduct reviews, their time is likely paid for by universities and research institutes.

Without major reforms, it seems unlikely that reviewing will become more economical, relative to other costs associated with publishing. One reason is that while technology improvements may automate or partially automate some aspects of publishing, peer review likely cannot be automated as easily. However, reducing details that reviewers should check might soon become automated (see <https://scicrunch.org/ASWG>).

A second issue is that while there is much discussion of how to reduce other costs associated with publishing, little attention has been devoted to reducing the cost of peer review, even though it would likely be the costliest component of the system if reviewers were paid for the reviews – rather than conducting the reviews under their “salary” paid time. After a long period of above-inflation subscription journal price increases, funders have attempted to put downward pressure on prices through initiatives such as *Plan S* (18) and through funding separate publishing infrastructures (e.g., Wellcome Open Research and Gates Open Research 19,20). However, because publishers do not have to pay for peer review, putting pressure on publishers may have no effect on review labor costs. Peer review labor sticks out as a large cost that is not being addressed systematically by publishers. In another domain, research funders have worked on reducing the cost of paid grant review, for example by shortening the proposals or reducing the need for consensus meetings after individual assessments (21).

Here we will discuss two reforms to reduce the cost of peer review. The first would decrease the amount of labor needed per published article by reducing redundancy in reviews. The second would make better use of less-trained reviewers. Finally, we will briefly mention a few other reforms that may not reduce cost per review but would boost the benefits of peer review, thus improving the cost-benefit ratio.

### ***Reducing redundancy in peer review***

Many manuscripts get reviewed at multiple journals, which is a major inefficiency (e.g., 22). Because this is a multiplicative factor, it exacerbates the issue of the rising global increase in number of submissions. While improvements in the manuscript between submissions means that the reviewing process is not entirely redundant, typically at least some of the assessment being done is duplication. Based on survey data (23), we conservatively estimated that, on average, a manuscript is submitted to two journals before acceptance (including the accepting journal). In



other words, each accepted article has one rejection and resubmission behind it. Should the reviews of a previous submission be available to the journal of the new submission, reviewing time could be substantially reduced (presuming that the quality of review does not differ between journals – and it very likely does), but unfortunately this is not common practice. If we assume that the “passed on” or open reviews would reduce the requirements by one review per manuscript, then approx. 28M hours (of our 85M hour total estimate) could be saved annually. In the US alone, it would mean a savings of approx. 297M USD of work<sup>28</sup>.

Some savings of this kind have already begun. Several publishers or journals share reviews across their own journals (PLOS, Nature 24), which is sometimes known as “cascading peer review” (25). Some journals openly publish the reviews they solicit (e.g., eLife; Meta-psychology; PLOS; Research Integrity and Peer Review; for a recent review see (26)), although typically not when the manuscript is rejected (Meta-psychology is an exception, and eLife will publish the reviews after a rejected manuscript is accepted somewhere else). The Review Commons initiative allows authors to have their preprint reviewed, with those reviews used by journal publishers including EMBO and PLoS (27). Similarly, Peer Community In (peercommunityin.org) solicits reviews of preprints that can then be used by journals, including over 70 that have indicated they will consider such reviews.

A decline in the amount of research conducted, or the number of manuscripts this research results in, would reduce the amount of peer review labor needed. The number of articles being published has been growing rapidly for many decades (28,29). Some of this may be due to salami slicing (publishing thinner papers, but more of them), but this is not necessarily true - one study found that researchers’ individual publication rate has not increased (30) when normalized by the number of authors per paper, suggesting that authors are collaborating more to boost their publication count rather than publishing thinner papers. Hence, the increase in publication volume may be more a result of the steady increase in the global economy and, with it, support for researchers. Quality rather than publication quantity has, however, recently begun to be emphasized more by some funders and national evaluation schemes, and this may moderate the rate of growth in number of publications and potentially the peer review burden (31).

---

<sup>28</sup> (715,645 articles × 6h)\* \$69.25

***Improving the allocation of review labor***

## Broadening and deepening the reviewer pool

Journal editors disproportionately request reviews from senior researchers, whose time is arguably the most valuable. One reason for this is that senior researchers on average show up more often in literature searches, and also editors favor people they are familiar with, and younger researchers have had less time to become familiar to editors (32). With the same individuals tapped more and more, the proportion of requests that they can agree to falls (33), which is likely one reason that editors have to issue increasing numbers of requests to review (a contributor to increasing costs which we did not calculate). Journal peer review, therefore, takes longer and longer because the system fails to keep up with academia's changing demographics (3). Today, more women and minorities are doing academic research, and the contributions from countries such as China are growing rapidly. But many of these researchers don't show up on the radar of the senior researchers, located disproportionately in North America and Europe, who edit journals. This can be addressed by various initiatives, such as appointing more diverse editors and encouraging junior researchers to sign up to databases that editors consult when they seek reviewers (34, 35).

A more substantial increase in efficiency might come from soliciting contributions to peer review from individuals with less expertise than traditionally has been expected. Journal editors traditionally look for world experts on a topic, whose labor is particularly costly in addition to being in short supply and in high demand. But perhaps contributions to peer review shouldn't be confined only to those highly expert in a field. Evaluating a manuscript means considering multiple dimensions of the work and how it is presented. For some research areas, detailed checklists have been developed regarding all the information that should be reported in a manuscript (see [www.equator-network.org](http://www.equator-network.org)). This provides a way to divide up the reviewing labor and have some aspects where even students, after some training, can vet aspects of manuscripts. Thus, we are hopeful that after more meta-research on what is desired from peer review for particular research areas, parts of peer review can be done by people who are not experts in the very specific topic of a manuscript but can nonetheless be very capable at evaluating particular aspects of a manuscript (and as mentioned above, automation can help with some tasks).

This process could also lead to greater specialization in peer review. For example, for manuscripts that report clinical trials, some people could be trained in evaluating the blinding protocol and resulting degree of success of blinding (36), and if they had the opportunity to evaluate that particular portion of many manuscripts, they grow better at it and thus can evaluate more in a

shorter time, reducing the number of hours of labor that need be paid for. To some extent, this specialization in peer review has already begun. As reporting standards for particular kinds of research have become more widespread (e.g., Consolidated Standards of Reporting Trials (CONSORT) for clinical trials, Animal Research: Reporting of In Vivo Experiments (ARRIVE) for animal research, and Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) for systematic reviews of randomized trials<sup>29</sup>), professional staff at some publishers have begun performing some checks for compliance with these standards. For example, staff at *PLOS* check all manuscripts on human subject research for a statement regarding compliance with the Declaration of Helsinki, and clinical trials research for a CONSORT statement. These staff presumably can do this job more efficiently, and do so for a lower salary, than an academic charged with peer reviewing every word of an entire manuscript. There are also some services (e.g., RIPETA<sup>30</sup>, PUBSURE<sup>31</sup>) that automatically screen the to-be-submitted manuscripts and provide reports on potential errors and instant feedback to the authors, while other products (e.g., AuthorONE<sup>32</sup>) support publishers with automatic manuscripts screening including technical readiness checks, plagiarism checks, and checking for ethics statements.

### *Unlocking the value of reviews*

Some reforms to peer review would not reduce the cost per review, but would increase the benefits per review, improving the cost-benefit ratio. One such reform is making reviews public instead of confidential. Under the currently-dominant system of anonymised peer review, however, only the authors, other reviewers, and editor of the manuscript have the opportunity to benefit from the content of the review.

When reviews are published openly, the expert judgments and information within reviews can benefit others. One benefit is the judgments and comments made regarding the manuscript. Reviews often provide reasons for caution about certain interpretations, connections to other literature, points about the weaknesses of the study design, and what the study means from their

---

<sup>29</sup> For their collection, see <https://www.equator-network.org/>.

<sup>30</sup> <https://ripeta.com/>

<sup>31</sup> <https://pubsure.researcher.life/>

<sup>32</sup> <https://www.enago.com/authorone-publisher.htm>

particular perspective. While those comments influence the revision of the manuscript, often they either don't come through as discrete points or the revisions are made to avoid difficult issues, so that they don't need to be mentioned.

It is not uncommon for some of the points made in a review to also be applicable to other manuscripts. Some topics of research have common misconceptions that lead to certain mistakes or unfortunate choices in study design. Some of the experienced researchers that are typically called upon to do peer review can rapidly detect these issues, and pass on the “tips and tricks” that make for a rigorous study of a particular topic or that uses a particular technique. But because peer reviews are traditionally available only to the editor and authors of the reviewed study, this dissemination of knowledge happens only very slowly, much like the traditional apprenticeship system required for professions before the invention of the printing press. How much more productive would the scientific enterprise be if the information in peer reviews were unlocked? We should soon be able to get a better sense of this, as this is already being done by the journals that have begun publishing at least some of their peer reviews (e.g, *Meta-psychology*, *eLife*, *the PLOS journals*; *F1000Research*, *Royal Society Open Science*, *Annals of Anatomy*, *Nature Communications*, *PeerJ* (20)). It will be very difficult, however, to put a financial value on the benefits. Fortunately, there are also other reasons that suggest that such policies should be adopted, such as providing more information about the quality of published papers.

In some cases, performing a peer review can actually benefit the reviewer. In Publons' 2018 reviewer survey, 33% of respondents indicated that one reason (they could choose two from a list of nine) they agreed to review manuscripts was to “Keep up-to-date with the latest research trends in my field.” (p12 9). If more of such people can be matched with a manuscript, reviewing becomes more of a “win-win”, with greater benefits accruing to the reviewer than may be typical in the current system. Better matching, then, would mean an increased return on the portion of an employer's payment of a researcher's salary that pays for peer review. The initiatives that broaden the reviewer pool beyond the usual senior researchers that editors are most likely to think of may have this effect.

### **Limitations**

A limitation of the present study is that it does not quantify academic editors' labor, which is typically funded by universities, research institutes or publishers and is integral to the peer review process. At prestige journals with high rejection rates, a substantial proportion of (associate) editors' time is spent desk-rejecting articles, which could be considered wasteful, as rejected

articles are eventually published somewhere else. Which also requires additional work from authors to prepare the manuscripts and navigate different submission systems.

Additionally, our study's limitations come from the poverty of the available data. For example, today, no available database covers all scholarly journals and their articles. The rates of acceptance and rejections we used are approximate estimates. The average time spent on reviews likely strongly depends on fields and length of manuscript and we do not know how representative the number we used is of all academia. We could not calculate the cost of review for journal articles and conference papers separately, although they might differ in this regard. The nationality and salary of the reviewers are not published either, therefore, our calculations need to be treated with caution as they have to rely on broad assumptions. Nevertheless, the aim of this study was to estimate only the magnitude of the cost of peer review without the ambition to arrive at precise figures. We encourage publishers and other stakeholders to explore and openly share more information about peer review activities to foster a fairer and more efficient academic world.

#### **Availability of data and materials**

The public dataset supporting the conclusions of this article is available from the <https://app.dimensions.ai/discover/publication> webpage.

An Excel file including the formula used for the estimation in the present paper with interchangeable parameters is available at the OSF page of the project <https://osf.io/xk8tc/>.

#### **Competing interests**

**The authors declare that they have no competing interests.**

#### **Funding**

**This study was not funded.**

#### **Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

#### **Acknowledgement**

We are thankful to James Heathers and three anonymous reviewers for providing valuable feedback on an earlier version of the manuscript.

### Authors' Contributions

Conceptualization: BA and BS. Formal Analysis: BA and BS. Methodology: BA and BS. Writing - Original Draft Preparation: BA, BS, and AOH.

### References

1. Grossmann A, Brembs B. Assessing the size of the affordability problem in scholarly publishing [Internet]. PeerJ Preprints; 2019. Available from: <https://peerj.com/preprints/27809.pdf>
2. Horbach SP, Halffman W. Innovating editorial practices: academic publishers at work. *Research integrity and peer review*. 2020 Dec;5(1):1-5.
3. Kovanis M, Porcher R, Ravaud P, Trinquart L. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PloS One*. 2016;11(11):e0166387.
4. RIN. Activities, costs and funding flows in the scholarly communications system in the UK. *Res Inf Netw* [Internet]. 2008; Available from: <https://studylib.net/doc/18797972/activities-costs-and-funding-flows-report>
5. Baumol WJ, Bowen WG. Performing arts—the economic dilemma: a study of problems common to theater, opera, music and dance. *Gregg Revivals*; 1993.
6. Brainard J. The \$450 question: Should journals pay peer reviewers? *Science*. 2021;
7. Smith R. Peer reviewers—time for mass rebellion? [Internet]. *The BMJ*. 2021 [cited 2021 Mar 17]. Available from: <https://blogs.bmj.com/bmj/2021/02/01/richard-smith-peer-reviewers-time-for-mass-rebellion/>
8. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*. 2021 Jun 1;126(6):5113–42.
9. Publons. 2018 Global State of Peer Review [Internet]. 2018 [cited 2020 Sep 8]. Available from: <https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf>
10. Reuters T. Global publishing: Changes in submission trends and the impact on scholarly publishers. White Pap Thomson Reuters [Httpscholarone Commed Pdf](https://scholarone.com/med/Pdf). 2012;
11. Björk B-C. Acceptance rates of scholarly peer-reviewed journals: A literature survey. *Prof Inf* [Internet]. 2019 Jul 27 [cited 2021 Mar 9];28(4). Available from: <https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/epi.2019.jul.07>
12. Sugimoto CR, Larivière V, Ni C, Cronin B. Journal acceptance rates: a cross-disciplinary analysis of variability and relationships with journal measures. *J Informetr*. 2013;7(4):897–906.
13. Liguori EW, Tarabishy AE, Passerini K. Publishing entrepreneurship research: Strategies for success distilled from a review of over 3,500 submissions. *J Small Bus Manag*. 2021 Jan 2;59(1):1–12.
14. Shalvi S, Baas M, Handgraaf MJJ, Dreu CKWD. Write when hot — submit when not: seasonal bias in peer review or acceptance? *Learn Publ*. 2010;23(2):117–23.

15. Sense About Science. Peer review survey 2009: Full report. 2009 [cited 2020 Sep 9]; Available from: <https://senseaboutscience.org/activities/peer-review-survey-2009/>
16. Ware M. Peer review: benefits, perceptions and alternatives. London: Publishing Research Consortium.; 2008.
17. Vesper I. Peer reviewers unmasked: largest global survey reveals trends. *Nature* [Internet]. 2018 Sep 7 [cited 2021 Aug 4]; Available from: <https://www.nature.com/articles/d41586-018-06602-y>
18. Wallace N. Open-access science funders announce price transparency rules for publishers. *Science*. 2020;
19. Butler D. Wellcome Trust launches open-access publishing venture. *Nat News* [Internet]. 2016 [cited 2021 Mar 17]; Available from: <http://www.nature.com/news/wellcome-trust-launches-open-access-publishing-venture-1.20220>
20. Butler D. Gates Foundation announces open-access publishing venture. *Nat News*. 2017 Mar 30;543(7647):599.
21. Shepherd J, Frampton GK, Pickett K, Wyatt JC. Peer review of health research funding proposals: A systematic map and systematic review of innovations for effectiveness and efficiency. *PloS one*. 2018 May 11;13(5):e0196914.
22. Schriger DL, Sinha R, Schroter S, Liu PY, Altman DG. From Submission to Publication: A Retrospective Review of the Tables and Figures in a Cohort of Randomized Controlled Trials Submitted to the *British Medical Journal*. *Ann Emerg Med*. 2006 Dec 1;48(6):750-756.e21.
23. Jiang Y, Lerrigo R, Ullah A, Alagappan M, Asch SM, Goodman SN, et al. The high resource impact of reformatting requirements for scientific papers. *PLOS ONE*. 2019 Oct 30;14(10):e0223976.
24. Maunsell J. Neuroscience Peer Review Consortium. *J Neurosci*. 2008 Jan 23;28(4):787–787.
25. Barroga EF. Cascading peer review for open-access publishing. *Eur Sci Ed*. 2013;39(4):90–1.
26. Wolfram D, Wang P, Park H. Open Peer Review: The current landscape and emerging models. 2019;
27. New Policies on Preprints and Extended Scooping Protection [Internet]. *Review Commons*. [cited 2021 Aug 6]. Available from: <https://www.reviewcommons.org/blog/new-policies-on-preprints-and-extended-scooping-protection/>
28. Larsen P, Von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*. 2010;84(3):575–603.
29. de Solla Price DJ, Page T. Science since babylon. *Am J Phys*. 1961;29(12):863–4.
30. Fanelli D, Larivière V. Researchers' individual publication rate has not increased in a century. *PloS One*. 2016;11(3):e0149504.
31. Rijcke S de, Wouters PF, Rushforth AD, Franssen TP, Hammarfelt B. Evaluation practices and effects of indicator use—a literature review. *Res Eval*. 2016;25(2):161–9.
32. Garisto D. Diversifying peer review by adding junior scientists. *Nat Index*. 2020;777–84.

33. Breuning M, Backstrom J, Brannon J, Gross BI, Widmeier M. Reviewer fatigue? Why scholars decline to review their peers' work. *PS Polit Sci Polit*. 2015;48(4):595–600.
34. Kirman CR, Simon TW, Hays SM. Science peer review for the 21st century: Assessing scientific consensus for decision-making while managing conflict of interests, reviewer and process bias. *Regul Toxicol Pharmacol*. 2019;103:73–85.
35. Heinemann MK, Gottardi R, Henning PT. “Select Crowd Review”: A New, Innovative Review Modality for The Thoracic and Cardiovascular Surgeon. *Thorac Cardiovasc Surg*. 2021;
36. Bang H, Flaherty SP, Kolahi J, Park J. Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol. *Clin Res Regul Aff*. 2010 Jun 1;27(2):42–51.



## 2.2. Documenting contributions to scholarly articles using CRediT and *tenzing*<sup>33</sup>

Alex O. Holcombe<sup>1\*</sup>, Marton Kovacs<sup>2</sup>, Frederik Aust<sup>3,4</sup>, Balazs Aczel<sup>2</sup>

<sup>1</sup>School of Psychology, University of Sydney, Australia, <sup>2</sup>Institute of Psychology, ELTE, Eotvos Lorand University, Budapest, Hungary, <sup>3</sup>University of Cologne, Germany, <sup>4</sup>University of Amsterdam, Netherlands

### Abstract

Scholars traditionally receive career credit for a paper based on where in the author list they appear, but position in an author list often carries little information about what the contribution of each researcher was. “Contributorship” refers to a movement to formally document the nature of each researcher’s contribution to a project. We discuss the emerging CRediT standard for documenting contributions and describe a web-based app and R package called *tenzing* that is designed to facilitate its use. *tenzing* can make it easier for researchers on a project to plan and record their planned contributions and to document those contributions in a journal article.

---

<sup>33</sup> published as:

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and *tenzing*. *PLOS One*, *15*(12), e0244611.

## Introduction

Scholarly journal articles evolved from letters penned by individuals reporting scientific observations or experiment results. These letters listed only a single author, and it was clear that that person was claiming credit for all aspects of the work reported.

Today, over three hundred years later, most science is done by groups of people, not by lone individuals [1]. Different members of the team usually have different roles. Yet until recently, journals still operated as if there was no need to provide any information other than a list of names—the author list. Some information could be tentatively inferred from the order of names in the list, but how order is determined reflects often-unwritten practices around authorship that can be obscure to people outside a subfield and can differ substantially between labs [2].

When uncertain, people fall back on their prior beliefs. This is unfortunate for junior authors who do not have many papers to their name: when people see a list of authors with no explicit indication of who did what, they may give an outsize amount of credit to the senior author.

Fortunately, over the last few decades, many journals have begun to encourage, and some to require, that teams give some indication of who did what in the work reported by a paper. In some journals, this is done in a brief “Author Note” or “Author Information” section [e.g., 3]. Thanks to this development, researchers are more likely to get the specific recognition they deserve.

The included information would ideally be utilized by funders of scientists to allocate resources more effectively, so that teams with the right combination of skills would more often be supported. Moreover, those who hire scientists, such as universities and research institutes, should be able to assemble more effective teams for particular disciplines and projects.

Unfortunately, these potential benefits have been held back by a lack of standardization. Without a consistent vocabulary for describing what each researcher did in a project, and without a structured format for that information, it is difficult to aggregate across papers the type of contributions a researcher makes. For institutions and funders interested in supporting the right combinations of people, it is difficult to tally the sorts of contributions typically involved in different sorts of projects.

This issue is also faced by business and industry, where some solutions were devised. For commercial music for example, the recording industry uses an International Standard Musical Work Code (ISWC). This contains metadata for musical works that provide the identities of contributors and indicates whether they served the roles of, for example, composer, lyricist, or

arranger [4, 5]. A search of the associated ISWC database allows people to find the works that a musician has contributed to and what their role was in each work (<http://iswcnet.cisac.org/>).

In scientific research, roles may not be as clear cut as typical in the music industry. Nonetheless, useful distinctions can be made, such as contributions to the analysis of data versus to the drafting of a manuscript, or to the acquisition of data.

### ***CRediT***

In 2014, the first formal taxonomy was developed for scientific research—CRediT, the Contributor Role Taxonomy [6]. CRediT defines fourteen different types of contributions ([Table 1](#)), and over the last several years, it has been taken up by hundreds of journals [7] and dozens of publishers (see <http://credit.niso.org/adopters/>) and been endorsed by a number of journal editors [8].

**Table 1.** Contributor roles according to the Contributor Role Taxonomy (CRediT) [6], information available online at <http://credit.niso.org/>.

Contributor role	Description
Conceptualization	Ideas; formulation or evolution of overarching research goals and aims
Data curation	Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.
Formal analysis	Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.
Funding acquisition	Acquisition of the financial support for the project leading to this publication.

Investigation	Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
Methodology	Development or design of methodology; creation of models.
Project administration	Management and coordination responsibility for the research activity planning and execution.
Resources	Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.
Software	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
Supervision	Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.
Validation	Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
Visualization	Preparation, creation and/or presentation of the published work, specifically visualization/data presentation

Writing – original draft	Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)
Writing – review & editing	Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages

The use of CRediT not only can provide better documentation of the contributions of individual researchers, but also it enables meta-scientific research, such as into the different distribution of contributions indicated for women and men [9].

To facilitate researcher reporting of contributorship information in manuscripts and journal articles, we created *tenzing*, a web app and R package [10] for researchers and publishers. In the following, we will review how journals are currently using and reporting CRediT information. We then explain how *tenzing* can facilitate researcher and journal use of CRediT. Finally, we describe broader issues associated with CRediT contributorship that should be addressed as fields move forward with the usage of contributorship.

### ***How publishers are using CRediT***

The CRediT standard includes a specification for how to report contributorship information in the metadata that is associated with manuscript webpages (JATS-XML). But many publishers do not yet have the capability to do this. For example, it appears that none of the organizations behind preprint servers currently create CRediT metadata in JATS-XML format. In such cases, it can be useful for researchers to publish CRediT information in plain text in their manuscripts. Many journals make no mention of CRediT but ask researchers to indicate what each author did in the “Author Note” or similar section of the manuscript. Researchers can use CRediT to do this, in their preprints and in their submitted manuscripts.

An increasing number of scientific journals offer authors forms to indicate which CRediT category each author contributed to. For example, in the submission interface of *eLife*, authors encounter an array of checkboxes to indicate which category each author contributed to (Figure 1).

**\* Author Contributions** *(At least 1 is required.)*

We follow the recommendations of the ICMJE on authorship and contributorship. Please indicate the author's contributions:

- Conceptualization
- Data curation
- Formal analysis
- Funding acquisition
- Investigation
- Methodology
- Project administration
- Click here to add more detailed descriptions (optional)



Developed the binding assay; developed off-rate assays and performed measurements

- Resources
- Software
- Supervision
- Validation
- Visualization
- Writing - original draft
- Writing - review and editing

**Figure 1.** The journal *eLife*'s interface for indicating contributions when submitting a manuscript, available online at <https://elifesciences.org/inside-elifesciences/f39cfcf5/enabling-the-contributor-roles-taxonomy-for-author-contributions>.

PLOS journals provide a similar facility (Figure 2), as do over 1200 Elsevier journals (<https://www.elsevier.com/about/press-releases/corporate/elsevier-expands-credit-approach-to-authorship>).

**Country or Region\***

**Contributor Roles\***   [Instructions](#)

- Conceptualization
- Data curation
- Formal analysis
- Funding acquisition
- Investigation
- Methodology
- Project administration
- Resources
- Software
- Supervision
- Validation
- Visualization
- Writing – original draft
- Writing – review & editing

This is the corresponding author

**Figure 2.** The PLOS journals' interface for indicating contributions when submitting a manuscript. It appears when one is asked to enter information about each author.

Many authors encounter the CRediT roles for the first time when they are submitting to a journal. Or even if an author has used CRediT for a previous paper, they may be unlikely to explicitly consider these roles for a new paper until the time of journal submission. From multiple perspectives, not considering contributor roles until the time of submission is not ideal.

By the time an author submits a manuscript, the associated research project sometimes was completed months or even years before. At the time of journal submission, memory of each collaborator's contributions may be fuzzy. Ideally, authors will arrive at a consensus regarding who did what. But even if memories and records are adequate for this task, establishing such a consensus necessitates interruption of the submission of the manuscript until the submitting author hears from all the other authors and works to resolve any disagreement about various contributions, such as who contributed to the original draft of the manuscript.

Unfortunately, there is reason to believe that, when not discussed until after project completion, the rate of disagreement regarding author contributions may be high. Surveys suggest that between a third and two-thirds of researchers have been involved in authorship disagreements [11, 12, 13, 14]. In many fields, the submitting author is often the most junior author. This is typically the case when a PhD student submits her first paper, for example. Yet a student or other junior author is not in the best position to arbitrate disputes or push back on project contributors who may be overclaiming regarding their contribution [15]. For this and other reasons, there are many recommendations that authors communicate more about authorship expectations and roles, and that they should do so at the beginning of a project [16, 17, 18, 19]. This may be even more important when the manuscript is to provide not only a list of author names, but also a specification of each author's contributions.

Most authorship disputes are settled informally, but still may leave some people bitter at being excluded, or resentful that some people were included on an authorship list without any evidence they deserved it. The same likely applies to disputes over which contribution categories a researcher contributed to. It is probably best to get some agreement on these at the beginning of a project, so that researchers can proceed with some confidence around both what they are expected to do and what kind of credit they will get for it.

To facilitate project and credit attribution planning, an "authorship grid" system was described by Philippi et al. [21]. Each row of the grid is a task category or high-level responsibility associated with the project, and the columns are the researchers. At the intersection of the rows and columns, researchers indicate the more specific tasks they plan to perform, if any, in that category. This

approach is likely very useful for complex projects. For CRediT-using journals, this needs to be translated into CRediT information, which *tenzing* can facilitate.

How *tenzing* helps authors use CRediT

*tenzing* is a web app and associated R package that allows researchers to record contributorship information at any time, for eventual provision to a journal. The app is named after the mountaineer Tenzing Norgay, who together with Edmund Hillary was the first to reach the summit of Mount Everest. Norgay arguably received less credit than was appropriate given his contribution.

Here we will describe the use of *tenzing* solely in terms of the web app (<https://martonbalazskovacs.shinyapps.io/tenzing/>), although one can also use it via the underlying R package (<https://github.com/marton-balazs-kovacs/tenzing>)—full documentation for *tenzing* can be found online at <https://marton-balazs-kovacs.github.io/tenzing/>.

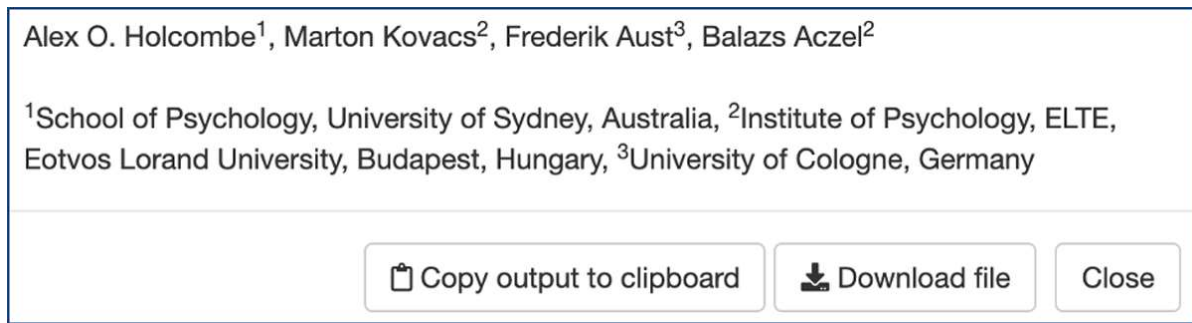
Use of *tenzing* starts with a spreadsheet template (provided as a Google Sheet, <http://bit.ly/tenzingTemplate>, but one can also use it in any spreadsheet editor, such as Excel). For a given research project, researchers make a copy of the template and then, in the rows, enter the names of their collaborators (Figure 3). One column is dedicated to each of the fourteen CRediT categories, to be checked off to indicate which categories each researcher contributed to. Because some CRediT categories are not entirely self-explanatory, one can hover the cursor over the column names to see some additional defining information.

Order in publication	Firstname	Middle name	Surname	Conceptualization	Data Curation	Formal Analysis	Funding Acquisition	Investigation	Methodology	Project Administration
1	Alex	O	Holcombe	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	Marton		Kovacs	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	Frederik		Aust	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	Balazs		Aczel	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**Figure 3.** Partial screenshot of the spreadsheet template (<http://bit.ly/tenzingTemplate>)

Around the time of the start of a project, a lead researcher may choose to send the link to the Sheet to all those involved, who can then indicate the areas they plan to contribute to. At the end of the project, or when plans change during the project, this Sheet can be revisited. Google Sheet services track the changes made in the template, thus by visiting the version history one can review the evolution of contributorship roles throughout the project.

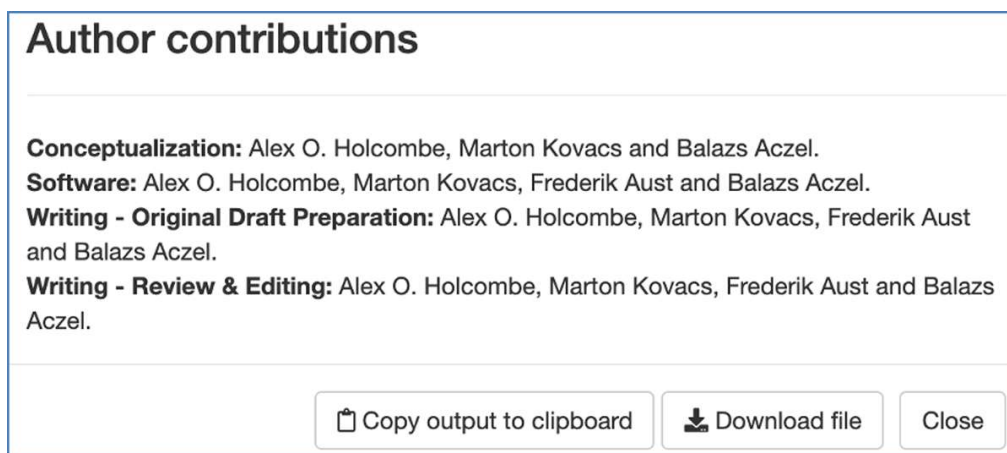




**Figure 4.** A screenshot of the author list and affiliation output screen.

When the researchers are ready to submit to a journal, they upload their filled-out spreadsheet to the *tenzing* app. They can then click a button to generate any of various outputs.

For CRediT, *tenzing* outputs a brief report in the form of a list indicating which contributor did what (Figure 5). This can be pasted into the section known at some journals as the Author Note. It is particularly appropriate for journals whose publishing platform does not support the machine-readable CRediT metadata. For example, the journal *Collabra: Psychology* encourages researchers to provide CRediT information in the “Author Contributions” section, because their publisher has not yet implemented creation of CRediT metadata in the article contents.



**Figure 5.** Screenshot of the *tenzing* window that provides a report of author contributions

The publishing platforms used by dozens of publishers can include CRediT metadata in JATS-XML-format in the journal article webpages (see <http://credit.niso.org/implementing-credit/>). *tenzing* can generate this JATS-XML information itself for users to download (Figure 6). Ideally, researchers would be able to upload this to a journal submission portal when submitting their manuscript, obviating the need to fill in arrays of checkboxes for each contributor. Unfortunately, at present no journal is capable of processing the uploaded JATS-XML, although a few publishers have privately indicated that they’re interested in adding support for this.

The Journal Article Tag Suite (JATS) is an XML format used to describe scientific literature published online. [Find out more about JATS XML](#)

```
<?xml version="1.0" encoding="UTF-8"?>
<contrib-group>
  <contrib>
    <name surname="Aczel" given-names="Balazs"/>
    <role vocab="credit" vocab-identifier="http://dictionary.casrai.org" />
    <role vocab="credit" vocab-identifier="http://dictionary.casrai.org" />
    <role vocab="credit" vocab-identifier="http://dictionary.casrai.org" />
    <role vocab="credit" vocab-identifier="http://dictionary.casrai.org" />
  </contrib>
  <contrib>
    <name surname="Aust" given-names="Frederik"/>
    <role vocab="credit" vocab-identifier="http://dictionary.casrai.org" />
  </contrib>
</contrib-group>
```

**Figure 6.** A screenshot of a portion of the JATS-XML output provided by *tenzing*.

Some researchers write manuscripts in R Markdown and use the *papaja* package [20] to generate manuscripts in APA format for submission to a journal. *tenzing* generates author metadata in YAML-format, which can be included in the R Markdown file. *papaja* then includes the CRediT information in the Author Note section of the APA-formatted manuscript.

**Tenzing** Documenting contributorship with CRediT

**1. Create your infosheet**  
First copy and then fill out this infosheet template

**2. Upload your infosheet**

Browse... No file selected

Show infosheet

**3. Download the output**

Show author contributions text

Show author list with affiliations

Show XML file (for publisher use)

Show *papaja* YAML

**How to use the application**

**1. Create your infosheet**

- Copy the infosheet template in your Google Drive File -> Make a copy
- Fill out your copy of the infosheet
- You can share it with your collaborators to make the process faster

**2. Upload your infosheet**

- Download the filled out infosheet to your computer in a .csv, .tsv or .xlsx format
  - If you use .xlsx format the contributorship information should be on the first sheet
- Click the "Browse" button and find your infosheet on your computer
  - If you want to take a look at the uploaded infosheet click "Show infosheet"

**3. Download the output**

- You can generate 3 types of outputs:
  - A human-readable report of the contributions with the "Author Contributions text"
  - The contributors affiliation page information for the manuscript with the "Annotated author list with affiliations"
  - JATS XML containing the contributions with the "XML (for publishers only)"
  - papaja* compatible YAML code of the contributor roles

[About](#)

**Figure 7.** A screenshot of the *tenzing* app. The bottom portion of both sides describes the four outputs that *tenzing* provides.

The current user interface for *tenzing* is shown in Figure 7, although its design is likely to evolve – a usability study is presently underway.

An additional output provided by *tenzing* is unrelated to CRediT: a list of the authors' names, with annotations indicating the institutions they are affiliated with, formatted to be suitable to paste into the title page of a manuscript file (Figure 4). For manuscripts with large numbers of authors, this can substantially reduce the time required to create the title page.

The current version of *tenzing* has various limitations, such as only allowing entry of one affiliation per author. Addressing this and a few other features is currently planned, with updates regarding progress available at the development site (<https://github.com/marton-balazs-kovacs/tenzing/issues>). User interface professionals have provided some suggestions, which will likely result in improvements to the app's design and usability. *tenzing* is open source [10], and researchers and other community members are invited to contribute to *tenzing* development by posting feature requests and bug reports at the Github issues page (<https://github.com/marton-balazs-kovacs/tenzing/issues>) or by contacting the corresponding author.

#### The future of CRediT

The CRediT standard was primarily designed to allow researchers to indicate what type of contribution they made. However, it also has a facility that allows one to indicate the *degree* of contribution. Specifically, one can optionally indicate whether each contributor to a particular category played a “lead”, “equal”, or “supporting” role in the associated work. It appears that most journals that use CRediT have opted not to use this feature, at least not yet. Editorial Manager, a journal platform used by thousands of journals, has integrated the degree of contribution feature but as a specific configuration, and most journals using Editorial Manager currently do not appear to have activated it.

An unresolved issue with CRediT's degree of contribution facet is how it should be used. It seems likely that if the “equal” degree is used, it must be used for multiple co-authors as it may not make sense when applied to just one. This is not currently addressed, however, by the CRediT documentation, nor are other possible constraints such as whether “equal” can be used as an intermediate indicator in cases where there are already authors with the “lead” and “supporting” labels. In addition, there is no indication to publishers of how they should indicate degrees of contribution in the machine-readable JATS-XML associated with journal articles, although Aries Systems, the creator of Editorial Manager, has done this by using the “specific-use” attribute (Caroline Webber, personal communication, 8 July 2020).

The degree of contribution under-specification is one of the issues that will likely be addressed by the group convened by the American National Information Standards Organization to formalize CRediT as an ANSI/NISO standard (<https://niso.org/press-releases/2020/04/niso-launches-work-contributor-role-taxonomy-credit-initiative>). For now, we have chosen to not yet implement the degree of contribution feature in *tenzing*.

### The future of contributorship

The number of contributors to the average scientific paper has steadily increased over the last several decades [22, 23]. In part, this has occurred because as knowledge in an area increases, specialization facilitates further advances. Some forms of research today, such as systematic reviews and meta-analyses, are based on bringing together large amounts of evidence from the literature. Library professionals contribute to some such projects with sophisticated searches of papers and databases. For other projects, technicians provide invaluable guidance regarding equipment, programmers create needed software, statisticians provide statistical advice, and informaticists create visualizations or collate information from databases. With science increasingly depending on these tasks getting done, funders need to be able to assess what sorts of projects have most benefited from specialists in order to resource science most effectively. However, people in these specialist roles are often not included in author lists, making it difficult to determine the number of specialists contributing to various projects.

One obstacle to greater inclusion of specialist contributors is the current state of journal authorship guidelines. The authorship guidelines for thousands of journals are based on the International Committee of Medical Journal Editors. These guidelines stipulate that only those who contribute to the writing or revising of a manuscript are eligible for authorship [24]. Journals should consider expanding authorship eligibility, for example by adopting the proposal of McNutt et al. [8] to eliminate the writing requirement and endorse the use of CRediT [25].

Some fields, such as genomics, already have a tradition of including groups, often known as consortia, on an author list, without enumeration of individual researcher names. This is often used to indicate those who only contributed data, which is a useful alternative to making that particular distinction with CRediT [26].

CRediT is not a good fit for all disciplines or even all projects within a discipline [27]. An ontology of roles that is both broader than those of CRediT and also more specific has been developed by the National Center for Data to Health, an initiative of the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health [28]. The scheme is called

the Contributor Role Ontology (CRO, <https://data2health.github.io/contributor-role-ontology/>), and it extends the CRediT ontology to include more than fifty roles, including “specimen collection”, “librarian”, “community engagement”, “coordination”, and “software testing” [29, 28, 30]. Given the adoption of CRediT that has already occurred, we anticipate that improvements will occur via extensions or generalizations such as CRO. The CRO scheme could be integrated into *tenzing* in the future.

If author contributions to a journal’s articles are explicitly indicated by a contributorship taxonomy such as CRediT or CRO, how should one think about the order of authorship? One might expect order to still be used for communicating the relative amount that different authors contributed, despite its limitations due to ambiguity around interpreting the meaning of first author and last author in different fields and cultures. However, note that CRediT also allows an indication of degree of contribution, beyond just how many categories a researcher contributed to. Specifically, where multiple individuals serve in the same role, the degree of contribution can optionally be specified as ‘lead’, ‘equal’, or ‘supporting’, but as described in the previous section, the proper usage of as well as the metadata for this has not yet been fully specified in the CRediT standard.

Deciding on order of authorship may get more and more difficult as the number of authors increases. Having a discussion among the researchers to decide this, without a clear decision process, may be unwieldy. Some have suggested a points system for different types of contributions. The American Psychological Association online authorship resources site for several years has included an example “scorecard” that assigns different types of contributions different numbers of points [31]. For CRediT, one such points system has been created by Mojtaba Soltanlou [32]. However, the relative value of different sorts of contributions likely differs across projects.

A critically important document for communicating contributions to scholarship is the CV. Traditionally, the extent of different authors’ contributions is communicated entirely by the order of authorship. In the future, however, we anticipate that funders or individual researchers will move to CVs that communicate the nature of the contributions made to each journal article. The Rescognito site [33] has created experimental visualizations, as did Ebersole, Adie, & Cook in a SIPS hackathon [25] with a bar graph indicating, for each CRediT category, how many papers a researcher contributed to.

Another piece of infrastructure already supporting CRediT usage is the ORCID database and metadata for identifying researchers and linking them to their papers and other scholarly

contributions [34]. Usage has grown rapidly, with over 7,000 papers a month indexed in Crossref because at least one author used ORCID [35]. The ORCID registry includes CRediT information. While *tenzing* could potentially pull author information such as name, email and affiliation from the ORCID database rather than requiring manual entry, the selection of the information to import can have complications that require user intervention (for example, one might need to include an old affiliation and not the current one). A prototype shiny app available at [https://colomb.shinyapps.io/contributorlist\\_creator/](https://colomb.shinyapps.io/contributorlist_creator/) facilitates that [36] and is now compatible with *tenzing*, as it can be used to create an infosheet one can further complete manually before uploading it into *tenzing*.

With adoption of CRediT growing rapidly, it is becoming more urgent to attend to any problems being encountered in its use or with the standard itself. The NISO effort to formalize CRediT will include a solicitation of feedback, which will be an important opportunity for the scholarly community to shape how contributorship information is recorded. We hope that the usage of CRediT facilitated by *tenzing* during the feedback period will result in a greater understanding of what about CRediT should be prioritized for refinement or change.

#### Acknowledgments

We thank the Society for the Improvement of Psychological Science (SIPS) and the participants in the 2019 SIPS Hackathon on contributorship [25] for discussion.

#### References

1. Sonnenwald DH. Scientific collaboration. *Annu. Rev. Inf. Sci. Technol.* **2008**, *41*, 643–681.
2. Patience GS, Galli F, Patience PA, Boffito DC. Intellectual contributions meriting authorship: Survey results from the top cited authors across all science categories. *PLoS ONE* **2019**, *14*, e0198117.
3. Authorship policies. (2009). *Nature*, *458*(7242), 1078. <https://doi.org/10.1038/4581078a>.
4. CISAC (4 February, 2019). CISAC launches major project to upgrade the international musical work identifier. Retrieved June 29, 2020, from <https://www.cisac.org/Newsroom/News-Releases/CISAC-launches-major-project-to-upgrade-the-international-musical-work-identifier>
5. Gilliéron, P. (2006). *Performing Rights Societies and The Digital Environment*. bepress Legal Series, 1436.
6. Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature News*, *508*(7496), 312.
7. Fennell C. (2019, 5 July). Fully agree, Elsevier is offering CRediT for close to 200 journals now, with great response from authors & editors. We're busy getting ready to expand thanks to @AriesMarketing making CRediT even easier for authors & editors. [Tweet]. Retrieved from <https://twitter.com/CatrionaFennell/status/1147119169831350272>.
8. McNutt, M. K., Bradford, M., Drazen, J. M., Hanson, B., Howard, B., Jamieson, K. H., Kiermer, V., Marcus, E., Pope, B. K., & Schekman, R. (2018). Transparency in authors'

- contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, 115(11), 2557–2560.
9. Macaluso, B., Larivière, V., Sugimoto, T., & Sugimoto, C. R. (2016). Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship: *Academic Medicine*, 91(8), 1136–1142. <https://doi.org/10.1097/ACM.0000000000001261>
  10. Kovacs, M., Aust, F., Holcombe, A. O., & Aczel, B. (2020). *tenzing: Documenting contributions to scientific scholarly output with CRediT (Version 0.1.0)*. Zenodo. <http://doi.org/10.5281/zenodo.3993411>
  11. Marušić A, Bošnjak L, Jerončić A. A systematic review of research on the meaning, ethics and practices of authorship across scholarly disciplines. *PLoS One*. 2011;6(9):e23477.
  12. Nylenna M, Fagerbakk F, Kierulf PJBME. Authorship: attitudes and practice among Norwegian researchers. *BMC Med Ethics*. 2014;15(1):53.
  13. Okonta P, Rossouw T. Prevalence of scientific misconduct among a group of researchers in Nigeria. *Dev World Bioeth*. 2013;13(3):149–57.
  14. Bhopal R, Rankin J, McColl E, Thomas L, Kaner E, Stacy R, Pearson P, Vernon B, Rodgers H. The vexed question of authorship: views of researchers in a British medical faculty. *BMJ*. 1997;314(7086):1009.
  15. Fisk C: Credit where it's due: the law and norms of attribution. *Duke Law School Faculty Scholarship Series 2006:Paper 39*.
  - 16.
  17. Albert T, Wager E. How to handle authorship disputes: a guide for new researchers. In: *Committee on Publication Ethics*; 2003.
  18. Wilcox LJ. Authorship: the coin of the realm, the source of complaints. *JAMA*. 1998;280(3):216–7.
  19. Hanson SMH. Collaborative research and authorship credit: beginning guidelines. *Nurs Res*. 1988;37(1):49–51.
  20. Bozeman, B., & Youtie, J. (2016). Trouble in paradise: Problems in academic research co-authoring. *Science and Engineering Ethics*, 22, 1717–1743. <https://doi.org/10.1007/s11948-015-9722-5>.
  21. Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
  22. Phillippi JC, Likis FE, Tilden EL. Authorship grids: Practical tools to facilitate collaboration and ethical publication. *Res Nurs Health*. 2018;41:195–208.
  23. Regalado, A. (1995). Multiauthor papers on the rise. *Science*, 268 (5207) (1995), p. 25.
  24. Fanelli, D., & Larivière, V. (2016). Researchers' Individual Publication Rate Has Not Increased in a Century. *PLOS ONE*, 11(3), e0149504. <https://doi.org/10.1371/journal.pone.0149504>
  25. International Committee of Medical Journal Editors. Defining the Role of Authors and Contributors. Updated December 2018. Available online: <http://www.icmje.org/icmje-recommendations.pdf> (accessed on 20 June 2019).
  26. Holcombe, A.O., Vazire, S., Chartier C., Ebersole, C., Giner-Sorolla, R., Haroz, S., Moreau, D., Primbs, M., Ling, M., Werner, K., Schnyder, N., Adie, J., Crook, Z., Smout, C., Ribeiro, G, Tangen, J., Aczel, B., Thibault, R., Searston, R., Van 't Veer, A., Schmalz, X. (2019). [Conference session]. Replace journals' writing-based authorship guidelines with a contributorship model. 2019 Annual Meeting of the Society for the Improvement of Psychological Science.
  27. Fontanarosa P, Bauchner H, Flanagan A. Authorship and Team Science. *Journal of the American Medical Association*. 2017;318(24):2433–2437. doi:10.1001/jama.2017.19341.
  28. Gadd E (2020). CRediT Check – Should we welcome tools to differentiate the contributions made to academic papers? [blog post].

- <https://blogs.lse.ac.uk/impactofsocialsciences/2020/01/20/credit-check-should-we-welcome-tools-to-differentiate-the-contributions-made-to-academic-papers/>
29. Vasilevsky, N. A., Hosseini, M., Teplitzky, S., Ilik, V., Mohammadi, E., Schneider, J., Kern, B., Colomb, J., Edmunds, S. C., Gutzman, K., Himmelstein, D. S., White, M., Smith, B., O’Keefe, L., Haendel, M., & Holmes, K. L. (2020). Is authorship sufficient for today’s collaborative research? A call for contributor roles. *Accountability in Research*, 1–21. <https://doi.org/10.1080/08989621.2020.1779591>
  30. Vasilevsky NA (2019). Introducing the Contribution Role Ontology: Developing a Sustainable Community-driven Approach to Attribution. [blog post] <https://www.forcel1.org/blog/introducing-contribution-role-ontology-developing-sustainable-community-driven-approach>
  31. Ilik, V., Conlon, M., Triggs, G., White, M., Javed, M., Brush, M., Gutzman, K., Essaid, S., Friedman, P., Porter, S., Szomszor, M., Haendel, M. A., Eichmann, D., & Holmes, K. L. (2018). OpenVIVO: Transparency in Scholarship. *Frontiers in Research Metrics and Analytics*, 2. <https://doi.org/10.3389/frma.2017.00012>
  32. [Author unknown]. Authorship determination scorecard. Available: <https://www.apa.org/science/leadership/students/authorship-determination-scorecard.pdf>. Retrieved 20 December 2020.
  33. Soltanlou M. Authorship\_credit.xlsx. Available: <https://osf.io/zd84r/>. Retrieved 20 December 2020.
  34. Wynne RCD (2019). “Got a DOI? Claim and Give Some CRediT” [https://figshare.com/articles/Got\\_a\\_DOI\\_Claim\\_and\\_Give\\_Some\\_CRediT\\_/9733595/1](https://figshare.com/articles/Got_a_DOI_Claim_and_Give_Some_CRediT_/9733595/1)
  35. Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25(4), 259-264.
  36. Wynne RCD (2020, June 24). Who uses ORCID IDs anyway? [post]. *LinkedIn*. <https://www.linkedin.com/pulse/who-uses-orcid-ids-anyway-richard-wynne/>
  37. Colomb J, Vasilevsky N (2020, January 23). open-science-promoters/contributor\_manager: MVP: from 1 orcid to 1 yaml (Version v.0.1-alpha). *Zenodo*. <http://doi.org/10.5281/zenodo.3625804>



## 2.3. Researchers working from home: Benefits and challenges<sup>34</sup>

Balazs Aczel<sup>1\*</sup>, Marton Kovacs<sup>1,2</sup>, Tanja van der Lippe<sup>3</sup>, Barnabas Szaszi<sup>1</sup>

<sup>1</sup>Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>2</sup>Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>3</sup>Department of Sociology, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The Netherlands

### Abstract

The flexibility allowed by the mobilization of technology disintegrated the traditional work-life boundary for most professionals. Whether working from home is the key or impediment to academics' efficiency and work-life balance became a daunting question for both scientists and their employers. The recent pandemic brought into focus the merits and challenges of working from home on a level of personal experience. Using a convenient sampling, we surveyed 704 academics while working from home and found that the pandemic lockdown decreased the work efficiency for almost half of the researchers but around a quarter of them were more efficient during this time compared to the time before. Based on the gathered personal experience, 70% of the researchers think that in the future they would be similarly or more efficient than before if they could spend more of their work-time at home. They indicated that in the office they are better at sharing thoughts with colleagues, keeping in touch with their team, and collecting data, whereas at home they are better at working on their manuscript, reading the literature, and analyzing their data. Taking well-being also into account, 66% of them would find it ideal to work more from home in the future than they did before the lockdown. These results draw attention to how working from home is becoming a major element of researchers' life and that we have to learn more about its influencer factors and coping tactics in order to optimize its arrangements.

**Keywords:** Working From Home, Telecommuting, work-life conflict, efficiency, COVID-19

---

<sup>34</sup> published as:

Aczel, B., Kovacs, M., Van Der Lippe, T., & Szaszi, B. (2021). Researchers working from home: Benefits and challenges. *PLOS One*, 16(3), e0249127.

## Introduction

Fleeing from the Great Plague that reached Cambridge in 1665, Newton retreated to his countryside home where he continued working for the next year and a half. During this time, he developed his theories on calculus, optics, and the law of gravitation - fundamentally changing the path of science for centuries. Newton himself described this period as the most productive time of his life (1). Is working from home indeed the key to efficiency for scientists also in modern times? A solution for working without disturbance by colleagues and being able to manage a work-life balance? What personal and professional factors influence the relation between productivity and working from home? These are the main questions that the present paper aims to tackle. The Covid-19 pandemic provides a unique opportunity to analyze the implications of working from home in great detail.

Working away from the traditional office is increasingly an option in today's world. The phenomenon has been studied under numerous, partially overlapping terms, such as telecommuting, telework, virtual office, remote work, location independent working, home office. In this paper, we will use 'working from home' (WFH), a term that typically covers working from any location other than the dedicated area provided by the employer.

The practice of WFH and its effect on job efficiency and well-being are reasonably well explored outside of academia (2,3). Internet access and the increase of personal IT infrastructure made WFH a growing trend throughout the last decades (4). In 2015, over 12% of EU workers (5) and near one-quarter of US employees (6) worked at least partly from home. A recent survey conducted among 27,500 millennials and Gen Z-s indicated that their majority would like to work remotely more frequently (7). The literature suggests that people working from home need flexibility for different reasons. Home-working is a typical solution for those who need to look after dependent children (8) but many employees just seek a better work-life balance (7) and the comfort of an alternative work environment (9).

Non-academic areas report work-efficiency benefits for WFH but they also show some downsides of this arrangement. A good example is the broad-scale experiment in which call center employees were randomly assigned to work from home or in the office for nine months (10). A 13% work

performance increase was found in the working from home group. These workers also reported improved work satisfaction. Still, after the experiment, 50% of them preferred to go back to the office mainly because of feeling isolated at home.

Home-working has several straightforward positive aspects, such as not having to commute, easier management of household responsibilities (11) and family demands (12), along with increased autonomy over time use (13,14), and fewer interruptions (15,16). Personal comfort is often listed as an advantage of the home environment (e.g., 15), though setting up a home office comes with physical and infrastructural demands (17). People working from home consistently report greater job motivation and satisfaction (4,11,18,19) which is probably due to the greater work-related control and work-life flexibility (20). A longitudinal nationally representative sample of 30,000 households in the UK revealed that homeworking is positively related with leisure time satisfaction (21), suggesting that people working from home can allocate more time for leisure activities.

Often-mentioned negative aspects of WFH include being disconnected from co-workers, experiencing isolation due to the physical and social distance to team members (22,23). Also, home-working employees reported more difficulties with switching off and they worked beyond their formal working hours (4). Working from home is especially difficult for those with small children (24), but intrusion from other family members, neighbours, and friends were also found to be major challenges of WFH (e.g., 17). Moreover, being away from the office may also create a lack of visibility and increases teleworkers' fear that being out of sight limits opportunities for promotion, rewards, and positive performance reviews (25).

Importantly, increased freedom imposes higher demands on workers to control not just the environment, but themselves too. WFH comes with the need to develop work-life boundary control tactics (26) and to be skilled at self-discipline, self-motivation, and good time management (27). Increased flexibility can easily lead to multitasking and work-family role blurring (28). Table 1 provides non-comprehensive lists of mostly positive and mostly negative consequences of WFH, based on the literature reviewed here.

Mostly positive	Mostly negative
Less commuting	Isolation from colleagues
More control over time	Less defined work-life boundaries
More autonomy	Higher need for self-discipline
Less office-related distractions	Reliance on private infrastructure
More comfortable environment	Communication difficulties with colleagues
More flexibility with domestic tasks	

Table 1. Positive and negative consequences of WFH.

Compared to the private sector, our knowledge is scarce about how academics experience working from home. Researchers in higher education institutes work in very similar arrangements. Typically, they are expected to personally attend their workplace, if not for teaching or supervision, then for meetings or to confer with colleagues. In the remaining worktime, they work in their lab or, if allowed, they may choose to do some of their tasks remotely. Along with the benefits on productivity when working from home, academics have already experienced some of its drawbacks at the start of the popularity of personal computers. As Snizek observed in the '80s, "(f)aculty who work long hours at home using their microcomputers indicate feelings of isolation and often lament the loss of collegial feedback and reinforcement" (page 622, 29).

Until now, the academics whose WFH experience had been given attention were mostly those participating in online distance education (e.g., 30,31). They experienced increased autonomy, flexibility in workday schedule, the elimination of unwanted distractions (32), along with high levels of work productivity and satisfaction (33), but they also observed inadequate communication and the lack of opportunities for skill development (34). The Covid-19 pandemic provided an opportunity to study the WFH experience of a greater spectrum of academics, since at one point most of them had to do all their work from home.

We have only fragmented knowledge about the moderators of WFH success. We know that control over time is limited by the domestic tasks one has while working from home. The view that women's work is more influenced by family obligations than men's is consistently shown in the literature (e.g., 35–37). Sullivan and Lewis (38) argued that women who work from home are able to fulfil their domestic role better and manage their family duties more to their satisfaction, but that comes at the expense of higher perceived work–family conflict (see also 39). Not surprisingly, during the COVID-19 pandemic, female scientists suffered a greater disruption than men in their academic productivity and time spent on research, most likely due to demands of childcare (40,41).

In summary, until recently, the effect of WFH on academics' life and productivity received limited attention. However, during the recent pandemic lockdown, scientists, on an unprecedented scale, had to find solutions to continue their research from home. The situation unavoidably brought into focus the merits and challenges of WFH on a level of personal experience. Institutions were compelled to support WFH arrangements by adequate regulations, services, and infrastructure. Some researchers and institutions might have found benefits in the new arrangements and may wish to continue WFH in some form; for others WFH brought disproportionately larger challenges. The present study aims to facilitate the systematic exploration and support of researchers' efficiency and work-life balance when working from home.

## **Materials and Methods**

Our study procedure and analysis plan were preregistered at <https://osf.io/jg5bz> (all deviations from the plan are listed in the Supporting Information). The survey included questions on research work efficiency, work-life balance, demographics, professional and personal background information. The study protocol has been approved by the Institutional Review Board from Eotvos Lorand University, Hungary (approval number: 2020/131). The Transparency Report of the study, the complete text of the questionnaire items and the instructions are shared at our OSF repository: <https://osf.io/v97fy/>.

## **Sampling**

As the objective of this study was to gain insight about researchers' experience of WFH, we aimed to increase the size and diversity of our sample rather than ascertaining the representativeness of our sample. Therefore, we distributed our online survey link among researchers in professional newsletters, university mailing lists, on social media, and by sending group-emails to authors (additional details about sampling are in the Supporting Information). As a result of the nature of

our sampling strategy, it is not known how many researchers have seen our participation request. Additionally, we did not collect the country of residence of the respondents. Responses analyzed in this study were collected between 2020-04-24 and 2020-07-13. Overall, 858 individuals started the survey and 154 were excluded because they did not continue the survey beyond the first question. As a result, 704 respondents were included in the analysis.

### **Procedure**

We sent the questionnaire individually to each of the respondents through the Qualtrics Mailer service. Written informed consent and access to the preregistration of the research was provided to every respondent before starting the survey. Then, respondents who agreed to participate in the study could fill out the questionnaire. To encourage participation, we offered that upon completion they can enter a lottery to win a 100 USD voucher.

### **Materials**

This is a general description of the survey items. The full survey with the display logic and exact phrasing of the items is transported from Qualtrics and uploaded to the projects' OSF page: <https://osf.io/8ze2g/>.

### ***Efficiency of research work***

The respondents were asked to compare the efficiency of their research work during the lockdown to their work before the lockdown. They were also asked to use their present and previous experience to indicate whether working more from home in the future would change the efficiency of their research work compared to the time before the lockdown. For both questions, they could choose among three options: “less efficient”; “more efficient”, and “similarly efficient”.

### ***Comparing working from home to working in the office***

Participants were asked to compare working from home to working from the office. For this question they could indicate their preference on a 7-point dimension (1: At home; 7: In the office), along 15 efficiency or well-being related aspects of research work (e.g., working on the manuscript, maintaining work-life balance). These aspects were collected in a pilot study conducted with 55 researchers who were asked to indicate in free text responses the areas in which their work benefits/suffers when working from home. More details of the pilot study are provided in the Supporting Information.

***Actual and ideal time spent working from home***

To study the actual and ideal time spent working from home, researcher were asked to indicate on a 0-100% scale (1) what percentage of their work time they spent working from home before the pandemic and (2) how much would be ideal for them working from home in the future concerning both research efficiency and work-life balance.

***Feasibility of working more from home***

With simple Yes/No options, we asked the respondents to indicate whether they think that working more from home would be feasible considering all their other duties (education, administration, etc.) and the given circumstances at home (infrastructure, level of disturbance).

***Background information***

Background questions were asked by providing preset lists concerning their academic position (e.g., full professor), area of research (e.g., social sciences), type of workplace (e.g., purely research institute), gender, age group, living situation (e.g., single-parent with non-adult child(ren)), and the age and the number of their children.

The respondents were also asked to select one of the offered options to indicate: whether or not they worked more from home during the coronavirus lockdown than before; whether it is possible for them to collect data remotely; whether they have education duties at work; if their research requires intensive team-work; whether their home office is fully equipped; whether their partner was also working from home during the pandemic; how far their office is from home; whether they had to do home-schooling during the pandemic; whether there was someone else looking after their child(ren) during their work from home in lockdown. When the question did not apply to them, they could select the 'NA' option as well.

**Data preprocessing and Analyses**

All the data preprocessing and analyses were conducted in R (42), with the use of the tidyverse packages (43). Before the analysis of the survey responses, we read all the free-text comments to ascertain that they do not contain personal information and they are in line with the respondent's answers. We found that for 5 items the respondents' comments contradicted their survey choices (e.g., whether they have children), therefore, we excluded the responses of the corresponding items from further analyses (see Supporting Information). Following the preregistration, we only conducted descriptive statistics of the survey results.

## Results

### Background information

The summary of the key demographic information of the 704 complete responses is presented in Table 2. A full summary of all the collected background information of the respondents are available in the Supporting Information.

Table 2. Number and Proportion of Respondents in Each Demographic Category

Background information question	Subgroup	Number of responses	Proportion of the subgroup
Gender	Female	356	50.57
Gender	Male	338	48.01
Gender	Prefer not to say	9	1.28
Gender	Other	1	0.14
Academic position	full professor	209	29.69
Academic position	associate professor	172	24.43
Academic position	assistant professor	126	17.90
Academic position	PhD student	72	10.23
Academic position	postdoc	72	10.23
Academic position	non-academic researcher	38	5.40
Academic position	research assistant	14	1.99
Academic position	not applicable	1	0.14



**Efficiency of research work**

The results showed that 94% (n = 662) of the surveyed researchers worked more from home during the COVID-19 lockdown compared to the time before. Of these researchers, 47% found that due to working more from home their research became, in general, less efficient, 23% found it more efficient, and 30% found no difference compared to working before the lockdown. Within this database, we also explored the effect of the lockdown on the efficiency of people living with children (n = 290). Here, we found that 58% of them experienced that due to working more from home their research became, in general, less efficient, 20% found it more efficient, and 22% found no difference compared to working before the lockdown. Of those researchers who live with children, we found that 71% of the 21 single parents and 57% of the 269 partnered parents found working less efficient when working from home compared to the time before the lockdown.

When asking about how working more from home would affect the efficiency of their research after the lockdown, of those who have not already been working from home full time (n = 684), 29% assumed that it could make their research, in general, less efficient, 29% said that it would be more efficient, and 41% assumed no difference compared to the time before the lockdown (Fig 1).

Focusing on the efficiency of the subgroup of people who live with children (n = 295), we found that for 32% their research work would be less efficient, for 30% it would be no different, and for 38% it would be more efficient to work from home after the lockdown, compared to the time before the lockdown.

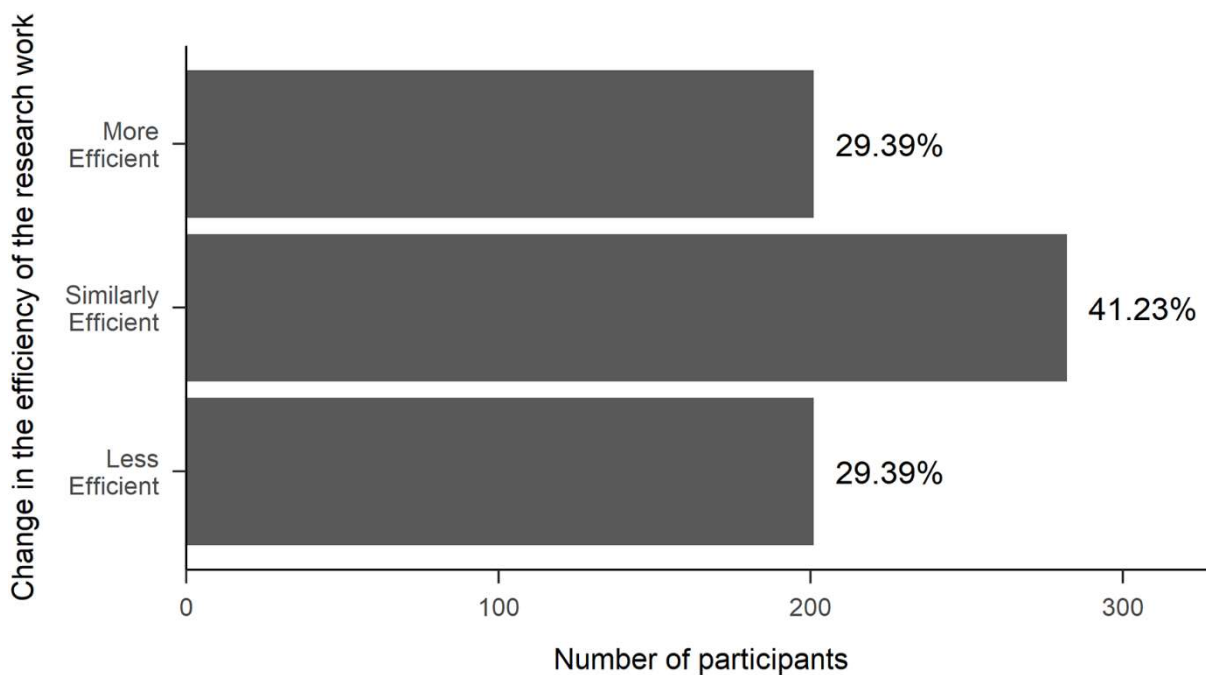


Fig 1. Percentages of the responses ( $N = 684$ ) given to the three answer options when asked how working more from home would affect the efficiency of their research after the lockdown.

### **Comparing working from home to working in the office**

When comparing working from home to working in the office in general, people found that they can better achieve certain aspects of the research in one place than the other. They indicated that in the office they are better at sharing thoughts with colleagues, keeping in touch with their team, and collecting data, whereas at home they are better at working on their manuscript, reading the literature, and analyzing their data (Fig 2).

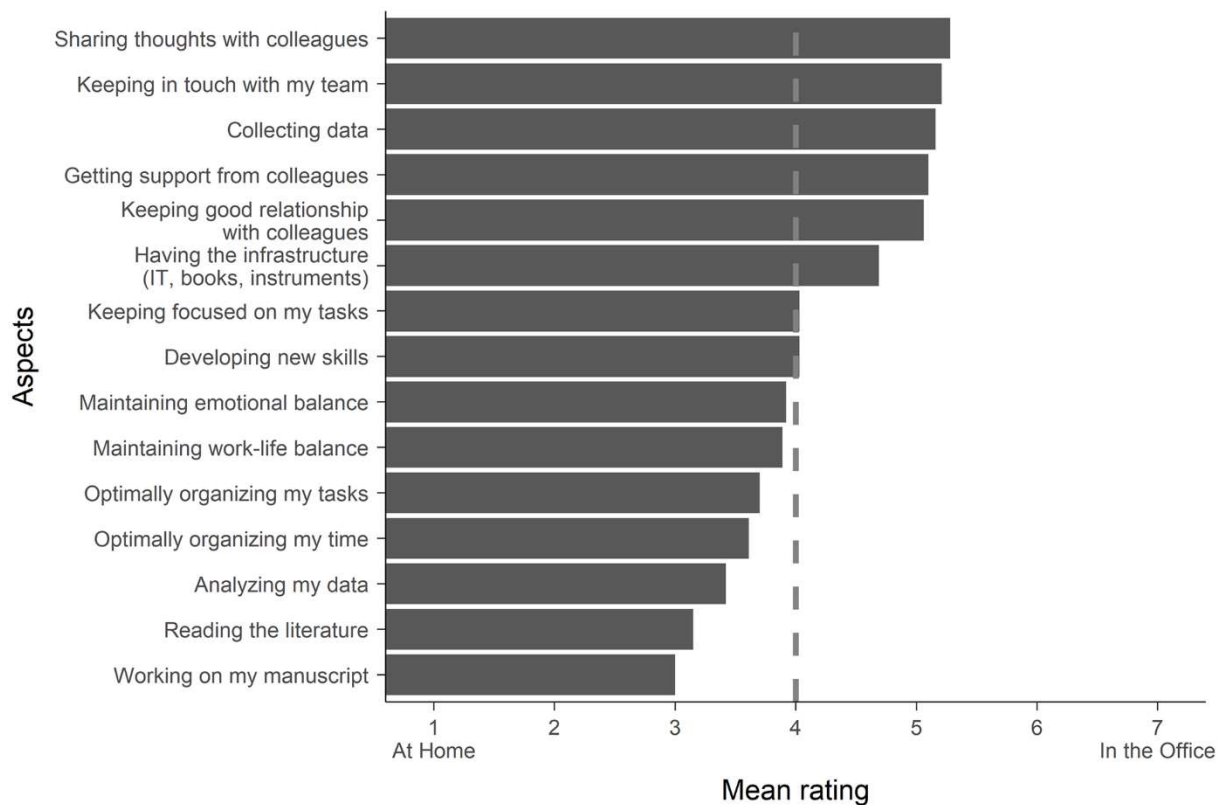


Fig 2. The comparison ( $N = 703$ ) of working at home and in the office concerning how the different aspects of research and work-life balance can be achieved. The bars represent response averages of the given aspects.

### Actual and ideal time spent working from home

We also asked the researchers how much of their work time they spent working from home in the past, and how much it would be ideal for them to work from home in the future concerning both research efficiency and well-being. Fig 3 shows the distribution of percentages of time working from home in the past and in an ideal future. Comparing these values for each researcher, we found that 66% of them want to work more from home in the future than they did before the lockdown, whereas 16% of them want to work less from home, and 18% of them want to spend the same percentage of their work time at home in the future as before. (These latter calculations were not preregistered.)

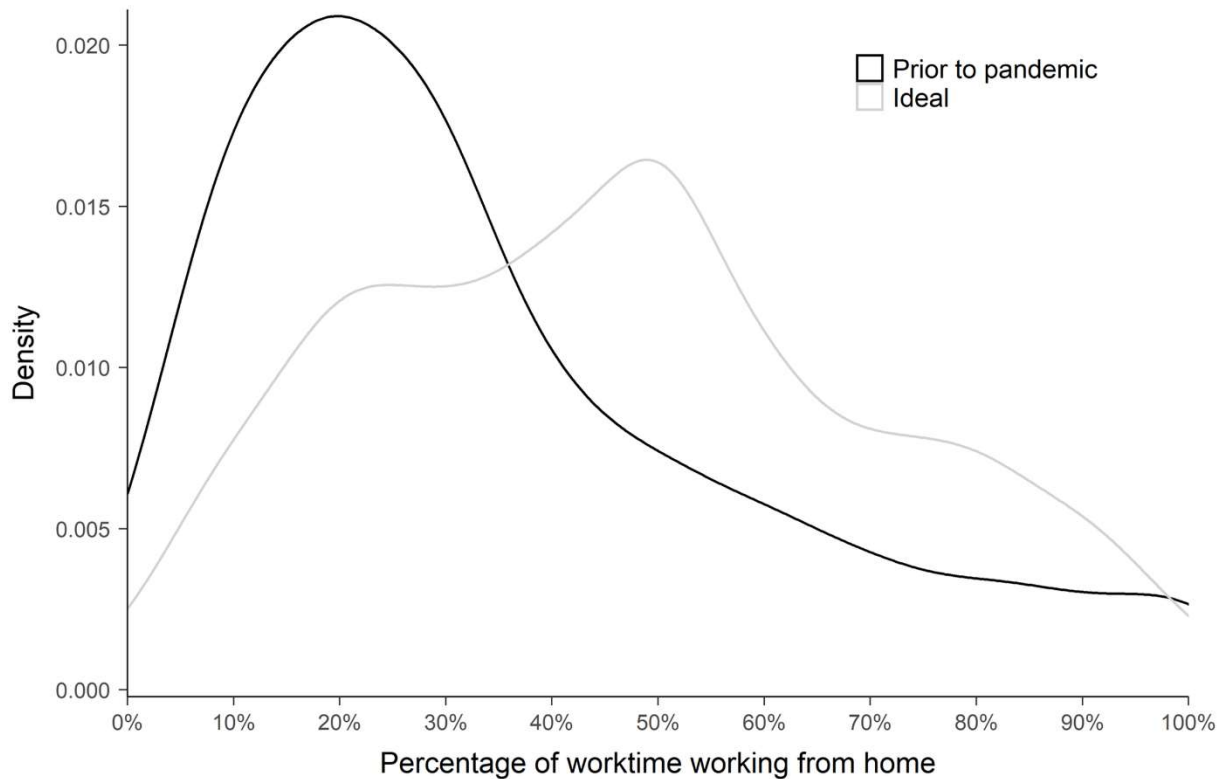


Fig 3. The density distributions of the responses ( $N = 704$ ) when asked how much of their worktime they worked from home before the pandemic lockdown and how much they would find ideal to work from home in the future.

### Feasibility of working more from home

Taken all their other duties (education, administration, etc.) and provided circumstances at home (infrastructure, level of disturbance), of researchers who would like to work more from home in the future ( $n = 461$ ), 86% think that it would be possible to do so. Even among those who have teaching duties at work ( $n = 376$ ), 84% think that more working from home would be ideal and possible.

### Discussion

Researchers' work and life have radically changed in recent times. The flexibility allowed by the mobilization of technology and the continuous access to the internet disintegrated the traditional work-life boundary. Where, when, and how we work depends more and more on our own arrangements. The recent pandemic only highlighted an already existing task: researchers' worklife has to be redefined. The key challenge in a new work-life model is to find strategies to

balance the demands of work and personal life. As a first step, the present paper explored how working from home affects researchers' efficiency and well-being.

Our results showed that while the pandemic-related lockdown decreased the work efficiency for almost half of the researchers (47%), around a quarter (23%) of them experienced that they were more efficient during this time compared to the time before. Based on personal experience, 70% of the researchers think that after the lockdown they would be similarly (41%) or more efficient (29%) than before if they could spend more of their work-time at home. The remaining 30% thought that after the lockdown their work efficiency would decrease if they worked from home, which is noticeably lower than the 47% who claimed the same for the lockdown period. From these values we speculate that some of the obstacles of their work efficiency were specific to the pandemic lockdown. Such obstacles could have been the need to learn new methods to teach online (44) or the trouble adapting to the new lifestyle (45). Furthermore, we found that working from the office and working from home support different aspects of research. Not surprisingly, activities that involve colleagues or team members are better bound to the office, but tasks that need focused attention, such as working on the manuscript or analyzing the data are better achieved from home.

A central motivation of our study was to explore what proportion of their worktime researchers would find ideal to work from home, concerning both research efficiency and work-life balance. Two thirds of the researchers indicated that it would be better to work more from home in the future. It seemed that sharing work somewhat equally between the two venues is the most preferred arrangement. A great majority (86%) of those who would like to work more from home in the future, think that it would be possible to do so. As a conclusion, both the work and non-work life of researchers would take benefits should more WFH be allowed and neither workplace duties, nor their domestic circumstances are limits of such a change. That researchers have a preference to work more from home, might be due to the fact that they are more and more pressured by their work. Finishing manuscripts, and reading literature is easier to find time for when working from home.

A main message of the results of our present survey is that although almost half of the respondents reported reduced work efficiency during the lockdown, the majority of them would prefer the

current remote work setting to some extent in the future. It is important to stress, however, that working from home is not equally advantageous for researchers. Several external and personal factors must play a role in researchers' work efficiency and work-life balance. In this analysis, we concentrated only on family status, but further dedicated studies will be required to gain a deeper understanding of the complex interaction of professional, institutional, personal, and domestic factors in this matter. While our study could only initiate the exploration of academics' WFH benefits and challenges, we can already discuss a few relevant aspects regarding the work-life interface.

Our data show that researchers who live with dependent children can exploit the advantages of working from home less than those who do not have childcare duties, irrespective of the pandemic lockdown. Looking after children is clearly a main source of people's task overload and, as a result, work-family conflict (46,47). As an implication, employers should pay special respect to employees' childcare situations when defining work arrangements. It should be clear, however, that other caring responsibilities should also be respected such as looking after elderly or disabled relatives (48). Furthermore, to avoid equating non-work life with family-life, a broader diversity of life circumstances, such as those who live alone, should be taken into consideration (49).

It seems likely that after the pandemic significantly more work will be supplied from home (50). The more of the researchers' work will be done from home in the future, the greater the challenge will grow to integrate their work and non-work life. The extensive research on work-life conflict, should help us examine the issue and to develop coping strategies applicable for academics' life. The Boundary Theory (26,51,52) proved to be a useful framework to understand the work-home interface. According to this theory, individuals utilize different tactics to create and maintain an ideal level of work-home segmentation. These boundaries often serve as "mental fences" to simplify the environment into domains, such as work or home, to help us attend our roles, such as being an employee or a parent. These boundaries are more or less permeable, depending on how much the individual attending one role can be influenced by another role. Individuals differ in the degree to which they prefer and are able to segment their roles, but each boundary crossing requires a cognitive "leap" between these categories (53). The source of conflict is the demands of the different roles and responsibilities competing for one's physical and mental resources. Working from home can easily blur the boundary between work and non-work domains. The conflict caused

by the intrusion of the home world to one's work time, just as well the intrusion of work tasks to one's personal life are definite sources of weakened ability to concentrate on one's tasks (54), exhaustion (55), and negative job satisfaction (56).

What can researchers do to mitigate this challenge? Various tactics have been identified for controlling one's borders between work and non-work. One can separate the two domains by temporal, physical, behavioral, and communicative segmentation (26). Professionals often have preferences and self-developed tactics for boundary management. People who prefer tighter boundary management apply strong segmentation between work and home (57,58). For instance, they don't do domestic tasks in worktime (temporal segmentation), close their door when working from home (physical segmentation), don't read work emails at weekends (behavioral segmentation), or negotiate strict boundary rules with family members (communicative segmentation). People on the other on one side of the segmentation-integration continuum, might not mind, or cannot avoid, ad-hoc boundary-crossings and integrate the two domains by letting private space and time be mixed with their work.

Researchers, just like other workers, need to develop new arrangements and skills to cope with the disintegration of the traditional work-life boundaries. To know how research and education institutes could best support this change would require a comprehensive exploration of the factors in researchers' WFH life. There is probably no one-size-fits-all approach to promote employees' efficiency and well-being. Life circumstances often limit how much control people can have over their work-life boundaries when working from home (59). Our results strongly indicate that some can boost work efficiency and wellbeing when working from home, others need external solutions, such as the office, to provide boundaries between their life domains. Until we gain comprehensive insight about the topic, individuals are probably the best judges of their own situation and of what arrangements may be beneficial for them in different times (60). The more autonomy the employers provide to researchers in distributing their work between the office and home (while not lowering their expectations), the more they let them optimize this arrangement to their circumstances.

Our study has several limitations: to investigate how factors such as research domain, seniority, or geographic location contribute to WFH efficiency and well-being would have needed a much greater sample. Moreover, the country of residence of the respondents was not collected in our survey and this factor could potentially alter the perception of WFH due to differing social and infrastructural factors. Whereas the world-wide lockdown has provided a general experience to WFH to academics, the special circumstances just as well biased their judgment of the arrangement. With this exploratory research, we could only scratch the surface of the topic, the reader can probably generate a number of testable hypotheses that would be relevant to the topic but we could not analyze in this exploration.

Newton working in lockdown became the idealized image of the home-working scientist. Unquestionably, he was a genius, but his success probably needed a fortunate work-life boundary. Should he had noisy neighbours, or taunting domestic duties, he might have achieved much less while working from home. With this paper, we aim to draw attention to how WFH is becoming a major element of researchers' life and that we have to be prepared for this change. We hope that personal experience or the topic's relevance to the future of science will invite researchers to continue this work.

### **Acknowledgments**

We would like to thank Szonja Horvath, Matyas Sarudi, and Zsuzsa Szekely for their help with reviewing the free text responses.

**Supporting Information:** <https://doi.org/10.1371/journal.pone.0249127.s001>



## References

1. Westfall RS. Newton's Marvelous Years of Discovery and Their Aftermath: Myth versus Manuscript. *Isis*. 1980;71(1):109–21.
2. Charalampous M, Grant CA, Tramontano C, Michailidis E. Systematically reviewing remote e-workers' well-being at work: a multidimensional approach. *Eur J Work Organ Psychol*. 2019;28(1):51–73.
3. Van der Lippe T, Lippényi Z. Beyond formal access: Organizational context, working from home, and work–family conflict of men and women in European workplaces. *Soc Indic Res*. 2018;1–20.
4. Felstead A, Henseke G. Assessing the growth of remote working and its consequences for effort, well-being and work-life balance. *New Technol Work Employ*. 2017;32(3):195–212.
5. Parent-Thirion A, Biletta I, Cabrita J, Vargas O, Vermeylen G, Wilczynska A, et al. Sixth European working conditions survey: Overview report. Eurofound (European Foundation for the Improvement of Living and Working ...; 2016.
6. US Department of Labor B of LS. American time use survey—2015 results. 2016;
7. Deloitte. The Deloitte Global Millennial Survey 2020: Millennials and Gen Zs hold the key to creating a “better normal” [Internet]. Deloitte Touche Tohmatsu; 2020 [cited 2020 Jul 5]. Available from: <https://www2.deloitte.com/global/en/pages/about-deloitte/articles/millennialsurvey.html>
8. Vilhelmson B, Thulin E. Who and where are the flexible workers? Exploring the current diffusion of telework in Sweden. *New Technol Work Employ*. 2016;31(1):77–96.
9. Tremblay D-G. Balancing work and family with telework? Organizational issues and challenges for women and managers. *Women Manag Rev*. 2002;
10. Bloom N, Liang J, Roberts J, Ying ZJ. Does working from home work? Evidence from a Chinese experiment. *Q J Econ*. 2015;130(1):165–218.
11. Wheatley D. Employee satisfaction and use of flexible working arrangements: *Work Employ Soc*. 2017;31(4):567–85.
12. Singley SG, Hynes K. Transitions to parenthood: Work-family policies, gender, and the couple context. *Gend Soc*. 2005;19(3):376–97.
13. Gajendran RS, Harrison DA. The good, the bad, and the unknown about telecommuting: meta-analysis of psychological mediators and individual consequences. *J Appl Psychol*. 2007;92(6):1524–41.
14. Kossek EE, Thompson RJ. Workplace flexibility: Integrating employer and employee perspectives to close the research–practice implementation gap. *Oxf Handb Work Fam*. 2016;255.
15. Kurland NB, Bailey DE. The advantages and challenges of working here, there, anywhere, and anytime. *Organ Dyn*. 1999;28(2):53–68.
16. Korbelt JO, Stegle O. Effects of the COVID-19 pandemic on life scientists. *Genome Biol*. 2020;21(113).
17. Gurstein P. Planning for telework and home-based employment: Reconsidering the home/work separation. *J Plan Educ Res*. 1996;15(3):212–24.

18. Binder M, Coad A. How satisfied are the self-employed? A life domain view. *J Happiness Stud.* 2016;17(4):1409–33.
19. Hill EJ, Ferris M, Mårtinson V. Does it matter where you work? A comparison of how three work venues (traditional office, virtual office, and home office) influence aspects of work and personal/family life. *J Vocat Behav.* 2003;63(2):220–41.
20. Baruch Y. The status of research on teleworking and an agenda for future research. *Int J Manag Rev.* 2001;3(2):113–29.
21. Reuschke D. The subjective well-being of homeworkers across life domains. *Environ Plan Econ Space.* 2019;51(6):1326–49.
22. Fonner KL, Roloff ME. Testing the Connectivity Paradox: Linking Teleworkers' Communication Media Use to Social Presence, Stress from Interruptions, and Organizational Identification. *Commun Monogr.* 2012 Jun 1;79(2):205–31.
23. Pinsonneault A, Boisvert M. The impacts of telecommuting on organizations and individuals: A review of the literature. In: *Telecommuting and virtual offices: Issues and opportunities.* IGI Global; 2001. p. 163–85.
24. McCloskey DW, Igarria M. Does "out of sight" mean "out of mind"? An empirical investigation of the career advancement prospects of telecommuters. *Inf Resour Manag J IRMJ.* 2003;16(2):19–34.
25. Cooper CD, Kurland NB. Telecommuting, professional isolation, and employee development in public and private organizations. *J Organ Behav Int J Ind Occup Organ Psychol Behav.* 2002;23(4):511–32.
26. Kreiner GE, Hollensbe EC, Sheep ML. Balancing borders and bridges: Negotiating the work-home interface via boundary work tactics. *Acad Manage J.* 2009;52(4):704–30.
27. Richardson J, McKenna S. Reordering Spatial and Social Relations: A Case Study of Professional and Managerial Flexworkers. *Br J Manag.* 2014;25(4):724–36.
28. Glavin P, Schieman S. Work–Family Role Blurring and Work–Family Conflict: The Moderating Influence of Job Resources and Job Demands. *Work Occup.* 2012;39(1):71–98.
29. Snizek WE. Some Observations on the Effects of Microcomputers on the Productivity of University Scientists. *Knowledge.* 1987;8(4):612–24.
30. Kanuka H, Jugdev K, Heller R, West D. The rise of the teleworker: false promises and responsive solutions. *High Educ.* 2008;56(2):149–65.
31. Ng CF. Academics Telecommuting in Open and Distance Education Universities: Issues, challenges and opportunities. *Int Rev Res Open Distrib Learn [Internet].* 2006 Sep 28 [cited 2020 Jul 15];7(2). Available from: <http://www.irrodl.org/index.php/irrodl/article/view/300>
32. Schulte M. Distance Faculty Experiences: A Personal Perspective of Benefits and Detriments of Telecommuting. *J Contin High Educ.* 2015;63(1):63–6.
33. Tustin DH. Telecommuting academics within an open distance education environment of South Africa: More content, productive, and healthy? *Int Rev Res Open Distance Learn.* 2014;15(3):186–215.
34. Dolan VLB. The isolation of online adjunct faculty and its impact on their performance. *Int Rev Res Open Distrib Learn.* 2011 Feb 28;12(2):62–77.

35. van der Lippe T. Dutch workers and time pressure: household and workplace characteristics. *Work Employ Soc.* 2007;21(4):693–711.
36. van der Horst M, van der Lippe T, Kluwer E. Aspirations and occupational achievements of Dutch fathers and mothers. *Career Dev Int.* 2014 Jan 1;19(4):447–68.
37. van der Lippe T, Jager A, Kops Y. Combination Pressure: The Paid Work: Family Balance of Men and Women in European Countries. *Acta Sociol.* 2006;49(3):303–19.
38. Sullivan C, Lewis S. Home-based telework, gender, and the synchronization of work and family: perspectives of teleworkers and their co-residents. *Gend Work Organ.* 2001;8(2):123–45.
39. Hilbrecht M, Shaw SM, Johnson LC, Andrey J. ‘I’m home for the kids’: contradictory implications for work–life balance of teleworking mothers. *Gend Work Organ.* 2008;15(5):454–76.
40. Frederickson M. COVID-19’s gendered impact on academic productivity [Internet]. GitHub. 2020 [cited 2020 Jul 15]. Available from: <https://github.com/drfreder/pandemic-pub-bias>
41. Myers KR, Tham WY, Yin Y, Cohodes N, Thursby JG, Thursby MC, et al. Unequal effects of the COVID-19 pandemic on scientists. *Nat Hum Behav.* 2020 Jul 15;1–4.
42. Team RC. R: A Language and Environment for Statistical Computing. *Dim Ca533.* 2018;1(1358):34.
43. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. *J Open Source Softw.* 2019;4(43):1686.
44. Johnson N, Veletsianos G, Seaman J. US Faculty and Administrators’ Experiences and Approaches in the Early Weeks of the COVID-19 Pandemic. *Online Learn.* 2020;24(2):6–21.
45. Ghaffarizadeh SA, Ghaffarizadeh SA, Behbahani AH, Mehdizadeh M, Olechowski A. Life and work of researchers trapped in the COVID-19 pandemic vicious cycle. *bioRxiv.* 2021;
46. Adkins CL, Premeaux SF. Spending time: The impact of hours worked on work–family conflict. *J Vocat Behav.* 2012;80(2):380–9.
47. Premeaux SF, Adkins CL, Mossholder KW. Balancing work and family: a field study of multi-dimensional, multi-role work–family conflict. *J Organ Behav Int J Ind Occup Organ Psychol Behav.* 2007;28(6):705–27.
48. Eikhof DR, Warhurst C, Haunschild A. What work? What life? What balance? Critical reflections on the work-life balance debate. *Empl Relat.* 2007;29(4):325–33.
49. Valcour M. Work-based resources as moderators of the relationship between work hours and satisfaction with work–family balance. *J Appl Psychol.* 2007;92(6):1512–23.
50. Barrero JM, Bloom N, Davis SJ. Why Working From Home Will Stick. *Univ Chic Becker Friedman Inst Econ Work Pap.* 2020;(2020–174).
51. Ashforth BE, Kreiner GE, Fugate M. All in a day’s work: Boundaries and micro role transitions. *Acad Manage Rev.* 2000;25(3):472–91.
52. Clark SC. Work/family border theory: A new theory of work/family balance. *Hum Relat.* 2000;53(6):747–70.
53. Zerubavel E. *The fine line.* University of Chicago Press; 1993.

54. Kim S, Hollensbe E. Work interrupted: a closer look at work boundary permeability. *Manag Res Rev.* 2017 Jan 1;40(12):1280–97.
55. Golden TD. Altering the effects of work and family conflict on exhaustion: Telework during traditional and nontraditional work hours. *J Bus Psychol.* 2012;27(3):255–69.
56. Carlson DS, Grzywacz JG, Kacmar KM. The relationship of schedule flexibility and outcomes via the work-family interface. *J Manag Psychol.* 2010;25(4):330–55.
57. Kossek EE, a Noe R, DeMarr BJ. Work-family role synthesis: Individual and organizational determinants. *Int J Confl Manag.* 1999;10(2):102–29.
58. Nippert-Eng C. Calendars and keys: The classification of “home” and “work.” In: *Sociological Forum.* Springer; 1996. p. 563–82.
59. Kossek EE, Ruderman MN, Braddy PW, Hannum KM. Work–nonwork boundary management profiles: A person-centered approach. *J Vocat Behav.* 2012;81(1):112–28.
60. Troup C, Rose J. Working from home: Do formal or informal telework arrangements provide better work–family outcomes? *Community Work Fam.* 2012;15(4):471–86.

### 3. TRANSPARENCY

#### 3.1. A Consensus-Based Transparency Checklist<sup>35</sup>

B. Aczel<sup>1\*</sup>, B. Szaszi<sup>1</sup>, A. Sarafoglou<sup>2</sup>, Z. Kekecs<sup>1</sup>, Š. Kucharský<sup>2</sup>, D. Benjamin<sup>3</sup>, C. D. Chambers<sup>4</sup>, A. Fisher<sup>2</sup>, A. Gelman<sup>5</sup>, M. A. Gernsbacher<sup>6</sup>, J. P. Ioannidis<sup>7</sup>, E. Johnson<sup>5</sup>, K. Jonas<sup>8</sup>, S. Kousta<sup>9</sup>, S. O. Lilienfeld<sup>10,11</sup>, D. S. Lindsay<sup>12</sup>, C. C. Morey<sup>4</sup>, M. Munafò<sup>13</sup>, B. R. Newell<sup>14</sup>, H. Pashler<sup>15</sup>, D. R. Shanks<sup>16</sup>, D. J. Simons<sup>17</sup>, J. M. Wicherts<sup>18</sup>, D. Albarracín<sup>17</sup>, N. D. Anderson<sup>19</sup>, J. Antonakis<sup>20</sup>, H. Arkes<sup>21</sup>, M. D. Back<sup>22</sup>, G. C. Banks<sup>23</sup>, C. Beevers<sup>24</sup>, A. A. Bennett<sup>25</sup>, W. Bleidorn<sup>26</sup>, T. W. Boyer<sup>27</sup>, C. Cacciari<sup>28</sup>, A. S. Carter<sup>29</sup>, J. Cesario<sup>30</sup>, C. Clifton<sup>31</sup>, R.M. Conroy<sup>33</sup>, M. Cortese<sup>34</sup>, F. Cosci<sup>35</sup>, N. Cowan<sup>36</sup>, J. Crawford<sup>37</sup>, E. A. Crone<sup>38</sup>, J. Curtin<sup>6</sup>, R. Engle<sup>39</sup>, S. Farrell<sup>40</sup>, P. Fearon<sup>16</sup>, M. Fichman<sup>41</sup>, W. Frankenhuis<sup>42</sup>, A. M. Freund<sup>43</sup>, M. G. Gaskell<sup>44</sup>, R. Giner-Sorolla<sup>45</sup>, D. P. Green<sup>5</sup>, R. L. Greene<sup>46</sup>, L. L. Harlow<sup>47</sup>, F. Hoces de la Guardia<sup>48</sup>, D. Isaacowitz<sup>49</sup>, J. Kolodner<sup>50</sup>, D. Lieberman<sup>51</sup>, G. D. Logan<sup>52</sup>, W. B. Mendes<sup>53</sup>, L. Moersdorf<sup>43</sup>, B. Nyhan<sup>54</sup>, J. Pollack<sup>55</sup>, C. Sullivan<sup>56</sup>, S. Vazire<sup>26</sup>, E.-J. Wagenmakers<sup>2</sup>

<sup>1</sup>ELTE, Eotvos Lorand University.

<sup>2</sup>University of Amsterdam.

<sup>3</sup>University of Southern California.

<sup>4</sup>Cardiff University.

<sup>5</sup>Columbia University.

<sup>6</sup>University of Wisconsin-Madison.

<sup>7</sup>Stanford University.

<sup>8</sup>Maastricht University.

<sup>9</sup>Nature Human Behaviour, Springer Nature.

<sup>10</sup>Emory University.

<sup>11</sup>University of Melbourne.

<sup>12</sup>University of Victoria.

<sup>13</sup>University of Bristol.

<sup>14</sup>University of New South Wales.

<sup>29</sup>University of Massachusetts, Boston.

<sup>30</sup>Michigan State University.

<sup>31</sup>University of Massachusetts Amherst.

<sup>33</sup>Royal College of Surgeons in Ireland.

<sup>34</sup>University of Nebraska Omaha.

<sup>35</sup>University of Florence.

<sup>36</sup>University of Missouri.

<sup>37</sup>The College of New Jersey.

<sup>38</sup>Leiden University.

<sup>39</sup>Georgia Institute of Technology.

<sup>40</sup>University of Western Australia.

<sup>41</sup>Carnegie Mellon University.

<sup>42</sup>Radboud University.

<sup>43</sup>University of Zurich.

---

<sup>35</sup> published as:

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4(1), 4–6.

- <sup>15</sup>University of California San Diego.                      <sup>44</sup>University of York.
- <sup>16</sup>University College London.                                <sup>45</sup>University of Kent.
- <sup>17</sup>University of Illinois.                                        <sup>46</sup>Case Western Reserve University.
- <sup>18</sup>Tilburg University.     <sup>47</sup>University of Rhode Island.
- <sup>19</sup>Rotman Research Institute, Baycrest.                   <sup>49</sup>Northeastern University.
- <sup>20</sup>University of Lausanne.                                   <sup>50</sup>Boston College.
- <sup>21</sup>Ohio State University.                                    <sup>51</sup>University of Miami.
- <sup>22</sup>University of Münster.                                     <sup>52</sup>Vanderbilt University.
- <sup>23</sup>University of North Carolina at Charlotte.           <sup>53</sup>University of California, San Francisco.
- <sup>24</sup>University of Texas at Austin.                         <sup>54</sup>University of Michigan.
- <sup>25</sup>Old Dominion University.                               <sup>55</sup>North Carolina State University.
- <sup>26</sup>University of California Davis.
- <sup>27</sup>Georgia Southern University.                           \*
- <sup>28</sup>University of Modena-Reggio Emilia.

Standfirst:

We present a consensus-based checklist to improve and document the transparency of research reports in social and behavioural research. An accompanying online application allows users to complete the form and generate a report that they can submit with their manuscript or post it to a public repository.

### Good Science Requires Transparency

Ideally, science is characterized by a “show me” norm, meaning that claims should be based on observations that are reported transparently, honestly, and completely<sup>1</sup>. When parts of the scientific process remain hidden, the trustworthiness of the associated conclusions is eroded. This erosion of trust affects the credibility not only of specific articles, but — when a lack of transparency is the norm — perhaps even entire disciplines. Transparency is required not only for evaluating and reproducing results (from the same data), but also for research synthesis and meta-analysis from the raw data and for effective replication and extension of that work. Particularly when the research is funded by public resources, transparency and openness constitute a societal obligation.

In recent years many social and behavioural scientists have expressed a lack of confidence in some past findings<sup>2</sup>, partly due to unsuccessful replications. Among the causes for this low replication rate are underspecified methods, analyses, and reporting practices. These research practices can be difficult to detect and can easily produce unjustifiably optimistic research reports. Such lack of transparency need not be intentional or deliberately deceptive. Human reasoning is vulnerable to a host of pernicious and often subtle biases, such as hindsight bias, confirmation bias, and motivated reasoning, all of which can drive researchers to unwittingly present a distorted picture of their results.

### The Practical Side of Transparency

How can scientists increase the transparency of their work? To begin with, they could adopt open research practices such as study preregistration and data sharing<sup>3-5</sup>. Many journals, institutions, and funders now encourage or require researchers to adopt these practices. Some scientific subfields have seen broad initiatives to promote transparency standards for reporting and summarizing research findings, such as START, SPIRIT, PRISMA, STROBE, CONSORT (see [www.equator-network.org](http://www.equator-network.org)). A few journals ask authors to answer checklist questions about statistical and methodological practices<sup>e.g., Nature Life Sciences Reporting Summary, 6</sup> and transparency (e.g., *Psychological Science*). Journals can signal that they value open practices by offering “badges” that acknowledge open data, code, and materials<sup>7</sup>. Endorsed by many journals, the Transparency and Openness Promotion (TOP) guidelines<sup>8</sup> promote the availability of all research items, including data, materials, and code. Authors can declare their adherence to these TOP standards by adding a transparency statement in their articles<sup>TOP Statement, 9</sup>. Collectively, these somewhat piecemeal innovations illustrate a science-wide shift toward greater transparency in research reports.

### Transparency Checklist

We provide a consensus-based, comprehensive transparency checklist that behavioural and social science researchers can use to improve and document the transparency of their research, especially for confirmatory work. The checklist reinforces the norm of transparency by identifying concrete actions that researchers can take to enhance transparency at all the major stages of the research process. Responses to the checklist items can be submitted along with a manuscript, providing

reviewers, editors, and eventually readers with critical information about the research process necessary to evaluate the robustness of a finding. Journals could adopt this checklist as a standard part of the submission process, thereby improving documentation of the transparency of the research that they publish.

We developed the checklist contents using a preregistered ‘reactive-Delphi’ expert consensus process<sup>10</sup>, with the goal of ensuring that the contents cover most of the elements relevant to transparency and accountability in behavioural research. The initial set of items was evaluated by 45 behavioural and social science journal editors-in-chief and associate editors as well as 18 open-science advocates. The Transparency Checklist was iteratively modified by deleting, adding, and rewording the items until a sufficiently high level of acceptability and consensus were reached and no strong counter arguments for single items were made (for the selection of the participants and the details of the consensus procedure see Supplementary materials). As a result, the checklist represents a consensus among these experts.

The final version of the Transparency Checklist 1.0 contains 36 items that cover four components of a study: *Preregistration; Methods; Results and Discussion; Data, Code, and Materials Availability*. For each item, authors select the appropriate answer from pre-specified options. It is important to emphasize that none of the responses on the checklist is *a priori* good or bad, and the transparency report provides researchers the opportunity to explain their choices at the end of each section.

In addition to the full checklist, we provide a shortened 12-item version (See Figure 1). By reducing the demands on researchers’ time to a minimum, the shortened list may facilitate broader adoption, especially among journals that intend to promote transparency but are reluctant to ask authors to complete a 36-item list. We created online applications for the two checklists that allow users to complete the form and generate a report that they can submit with their manuscript and/or post to a public repository (Box 1). The checklist is subject to continual improvement, and users can always access the most current version on the checklist website; access to previous versions will be provided on a subpage.



This checklist presents a consensus-based solution to a difficult task: identifying the most important steps needed for achieving transparent research in the social and behavioural sciences. Although this checklist was developed for social and behavioural researchers who conduct and report confirmatory research on primary data, other research approaches and disciplines might find value in it and adapt it to their field's needs. We believe that consensus-based solutions and user-friendly tools are necessary to achieve meaningful change in scientific practice. Without doubt, there might remain important topics the current version fails to cover; nonetheless, we trust that this version provides a useful to facilitate starting point for transparency reporting. The checklist is subject to continual improvement, we encourage researchers, funding agencies and journals to provide feedback and recommendations. We also encourage meta-researchers to assess the use of the checklist and its impact in the transparency of research.

#### Box 1. Online Applications and the Benefits of the Transparency Checklist

##### ***Online Applications for the Checklist***

<http://www.shinyapps.org/apps/TransparencyChecklist/> for the complete, 36-item version

<http://www.shinyapps.org/apps/ShortTransparencyChecklist/> for the shortened, 12-item version

##### ***Benefits of the Checklist***

- The checklist can help authors improve the transparency of their work before submission.
- Disclosed checklist responses can help editors, reviewers, and readers gain insight into the transparency of the submitted studies.
- Guidelines built on the checklist can be used for educational purposes and to raise the standards of social and behavioural sciences, as well as other scientific disciplines, regarding transparency and credibility.
- Funding agencies can use a version of this checklist to improve the research culture and accelerate scientific progress.

(1) Prior to analyzing the complete data set, a time-stamped preregistration was posted in an independent, third-party registry for the data analysis plan.  Yes  No  N/A

(2) The study was preregistered...

Please select an option

**The preregistration fully describes...**

(3) the intended statistical analysis for each research question (this may require, for example, information about the sidedness of the tests, inference criteria, corrections for multiple testing, model selection criteria, prior distributions etc.).  Yes  No  N/A

**The manuscript fully describes...**

(4) the rationale for the sample size used (e.g., an a priori power analysis).  Yes  No  N/A

(5) the study design, procedures, and materials to allow independent replication.  Yes  No  N/A

(6) the measures of interest (e.g., friendliness) and their operationalizations (e.g., a questionnaire measuring friendliness).  Yes  No  N/A

(7) any changes to the preregistration (such as changes in eligibility criteria, group membership cutoffs, or experimental procedures)?  Yes  No  N/A

**The manuscript...**

(8) distinguishes explicitly between "confirmatory" (i.e., prespecified) and "exploratory" (i.e., not prespecified) analyses.  Yes  No  N/A

**The following have been made publicly available...**

(9) the (processed) data, on which the analyses of the manuscript were based.  Yes  No  N/A

(10) all code and software (that is not copyright protected).  Yes  No  N/A

(11) all instructions, stimuli, and test materials (that are not copyright protected).  Yes  No  N/A

(12) The manuscript includes a statement concerning the availability and location of all research items, including data, materials, and code relevant to your study.  Yes  No  N/A

Fig 1: The Shortened Transparency Checklist 1.0

*Note.* After each section, the researchers can add a free text if they find that further explanation of their response is needed. The full version of the checklist can be reached at:

<http://www.shinyapps.org/apps/TransparencyChecklist/>

## References

1. Merton, R. *The Sociology of Science: Theoretical and Empirical Investigations*. (University of Chicago Press, 1973). doi:<http://dx.doi.org/10.1063/1.3128814>
2. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nat. News* **533**, 452–454 (2016).
3. Chambers, C. D. Registered reports: a new publishing initiative at Cortex. *Cortex* **49**, 609–610 (2013).
4. Gernsbacher, M. A. Writing empirical articles: Transparency, reproducibility, clarity, and memorability. *Adv. Methods Pract. Psychol. Sci.* **1**, 403–414 (2018).
5. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
6. Campbell, P. Announcement: Reducing our irreproducibility. *Nature* **496**, 398 (2013).
7. Kidwell, M. C. *et al.* Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biol.* **14**, e1002456 (2016).
8. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
9. Aalbersberg, Ij. J. *et al.* Making Science Transparent By Default; Introducing the TOP Statement. doi:[10.31219/osf.io/sm78t](https://doi.org/10.31219/osf.io/sm78t)
10. McKenna, H. P. The Delphi technique: a worthwhile research approach for nursing? *J. Adv. Nurs.* **19**, 1221–1225 (1994).

## Competing interests

Stavroula Kousta (SK) is Chief Editor of the journal Nature Human Behaviour. SK has recused herself from any aspect of decision-making on this manuscript and played no part in the assignment of this manuscript to in-house editors or peer reviewers. She was also separated and blinded from the editorial process from submission inception to decision. The other authors declared no competing interests.

## Data and materials availability

All anonymized raw and processed data as well as the survey materials are publicly shared on the Open Science Framework page of the project: <https://osf.io/v5p2r/>. Our methodology and data-analysis plan were preregistered prior to the project. The preregistration document can be accessed at: <https://osf.io/v5p2r/registrations>.

## **Funding**

This research was not funded.

## **Acknowledgements**

We thank Felix Schönbrodt and Andrei Tamas Foldes for their technical help with the application.

## **Author contributions**

B.A., B.S., A.S., Z.K., and E-J.W. conceptualized the project, conducted the survey study, analysed the data and drafted the initial version of the manuscript. Š.K. developed and designed the online application. D.B., C.D.C., A.F., A.G., M.A.G., J.P.I., E.J., K.J., S.K., S.O.L., D.S.L., C.C.M., M.M., B.R.N., H.P., D.R.S., D.J.S., and J.M.W. took part in the preparation and conclusion of the checklist items. D.A., N.D.A., J.A., H.A., M.D.B., G.C.B., C.B., A.A.B., W.B., T.W.B., C.C., A.S.C., J.C., C. Clifton, R.M.C., M.C., F.C., N.C., J. Crawford, E.A.C., J. Curtin, R.E., S.F., P.F., M.F., W.F., A.M.F., M.G.G., R.G-S., D.P.G., R.L.G., L.L.H., F.H.G., D.I., J.K., D.L., G.D.L., W.B.M., L.M., B.N., J.P., C.S., and S.V. evaluated the checklist items. All authors were involved in reviewing and editing the final version of the manuscript.

Supplementary Methods are available from:

<https://www.nature.com/articles/s41562-019-0772-6#Sec5>

## 3.2. A survey on how preregistration affects the research workflow: Better science but more work<sup>36</sup>

Alexandra Sarafoglou<sup>1</sup>, Marton Kovacs<sup>2,3</sup>, Bence Bakos<sup>3</sup>, Eric-Jan Wagenmakers<sup>1</sup>, Balazs Aczel<sup>3</sup>

<sup>1</sup> Department of Psychology, University of Amsterdam, The Netherlands

<sup>2</sup> Doctoral School of Psychology, ELTE Eotvos Lorand University, Hungary

<sup>3</sup> Institute of Psychology, ELTE Eotvos Lorand University, Hungary

---

<sup>36</sup> published as:

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7), 211997

### Abstract

The preregistration of research protocols and analysis plans is a main reform innovation to counteract confirmation bias in the social and behavioral sciences. While theoretical reasons to preregister are frequently discussed in the literature, the individually experienced advantages and disadvantages of this method remain largely unexplored. The goal of this exploratory study was to identify the perceived benefits and challenges of preregistration from the researcher's perspective. To this aim, we surveyed 355 researchers, 299 of whom had used preregistration in their own work. The researchers indicated the experienced or expected effects of preregistration on their workflow. The results show that experiences and expectations are mostly positive. Researchers in our sample believe that implementing preregistration improves or is likely to improve the quality of their projects. Criticism of preregistration is primarily related to the increase in work-related stress and the overall duration of the project. While the benefits outweighed the challenges for the majority of researchers with preregistration experience, this was not the case for the majority of researchers without preregistration experience. The experienced advantages and disadvantages identified in our survey could inform future efforts to improve preregistration and thus help the methodology gain greater acceptance in the scientific community.

*Keywords:* Open Science, Meta-Science, Replication Crisis

A physicist had a horseshoe hanging on the door of his laboratory. His colleagues were surprised and asked whether he believed that it would bring luck to his experiments. He answered: "No, I don't believe in superstitions. But I have been told that it works even if you don't believe in it."

---

Jones (1973, p. 14)

Over the past decade, the social sciences have undergone a methodological metamorphosis. In order to increase the quality and credibility of confirmatory empirical research, both journals and researchers have adopted a series of methodological reform measures (Spellman, 2015; Spellman, Gilbert, & Corker, 2018). Among these reform measures, preregistration is arguably the most consequential. The preregistration of empirical studies entails the specification of the research design, the hypotheses, and the analysis plan before data is collected and analyzed. Preregistration protects the confirmatory status of the reported results by preventing biases –such

as confirmation bias and hindsight bias— from contaminating the statistical analysis (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Munafò et al., 2017).

The concept of preregistration is not new; as early as 1878, Peirce (1878, p. 476) established three rules to guarantee that a hypothesis leads to a probable result, the first rule being that a hypothesis should be explicitly stated before data are collected to test its truth. In some research areas, such as medical clinical trials, preregistration has long become scientific routine. For instance, in the world's highest impact journal, the *New England Journal of Medicine*, the registration of clinical trials is a prerequisite for publication. A recent interdisciplinary study by Malički, Aalbersberg, Bouter, Mulligan, and ter Riet (2022) shows that while preregistration receives the least support by researchers in a catalogue of responsible research practices, as many as 39% of researchers within the health sciences agreed with the statement that all studies should be preregistered (compared to 17% of researchers in other fields).<sup>37</sup>

In the last ten years, preregistration has also found its way into psychological science. In fact, preregistration has become so widespread that some believe that it is on its way to becoming the norm (Nosek & Lindsay, 2018). The number of preregistrations has increased at “unprecedented and accelerating rates” (Nosek & Lindsay, 2018, p. 19). For instance, a recent survey among researchers in the Netherlands, found that 38.9% of researchers in the social and behavioral sciences had preregistered a study before (Gopalakrishna et al., 2021). Online repositories have been created to store preregistrations (e.g., the Open Science Framework (OSF; <https://osf.io>) and AsPredicted.org), and several journals recognize preregistered studies with badges (Kidwell et al., 2016). In addition, over 300 journals now offer the Registered Reports format as a submission option, allowing authors to integrate preregistration with the peer-review process (Chambers, 2013; Nosek & Lakens, 2014; <https://osf.io/rr/>).

In the course of its rapid spread, however, the effectiveness of preregistration has been repeatedly questioned. When discussing ways to combat the crisis of confidence, critics have argued that too heavy an emphasis is being placed on methodological reforms (e.g., Fiedler, 2018; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Szollosi et al., 2020). Preregistration was not designed to improve the theoretical foundation of studies. Instead it was proposed to limit the degrees of freedom researchers have in designing and executing studies, and

---

<sup>37</sup> The catalog also included, for instance, the statement that authors should report the availability of all data, materials, and codes (83% agreement across all fields) and the statement that journals should encourage the submission of replication studies (61% agreement across all fields).

analyzing the results. For that reason, critics argue that strong theory development, more so than methodological reforms, would advance psychological science in the long term. That is, if predictions were derived from weak theories, even the application of the most rigorous methods will not produce reliable scientific results. For instance, if theories do not adequately define the conditions under which a particular phenomenon is observed, it remains unclear whether a non-significant result constitutes evidence against the theory or whether the chosen operationalizations were inappropriate (Oberauer & Lewandowsky, 2019). Thus, instead of focusing primarily on the prevention of questionable research practices, the discussion on how to improve psychological science should be dominated by topics such as theory development, good experimental designs, and the proper statistical modelling of theoretical predictions (Fiedler, 2018; Oberauer & Lewandowsky, 2019; Szollosi et al., 2020; Szollosi & Donkin, 2021).

In defence of preregistration, van 't Veer and Giner-Sorolla (2016) argued that while preregistration might not *directly* improve theory development, preregistration will help shift the research focus away from the evaluation of a consistent and statistically significant pattern of results and toward the assessment of theory and methods. In addition, van 't Veer and Giner-Sorolla (2016) argue that preregistration may lead to positive side-effects that improve the overall quality of the scientific product. For instance, since all team members need to approve and scrutinize the hypotheses, methods, and analyses before data collection, study preregistration would improve the collaboration within the team and therefore yield more carefully thought-out research plans. However, it is still unclear whether or to what extent researchers actually perceive preregistered studies to be of higher quality than non-preregistered studies. On the one hand, Alister, Vickers-Jones, Sewell, and Ballard (2021) found that researchers reported that they would be more confident that a finding would replicate when the original authors had adhered to open science practices such as preregistration. On the other hand, a study by Field et al. (2020) found only ambiguous evidence that researchers trust in preregistered empirical findings more than non-preregistered ones.

It has been argued that the scrutiny associated with preregistration might even harm certain aspects of the research workflow. For instance, preregistration can be effortful and time-consuming (e.g., Nosek & Lindsay, 2018; van 't Veer & Giner-Sorolla, 2016). Open research practices were also found to have a small but statistically significant association with work pressure (Gopalakrishna et al., 2021). As recognized by Nosek et al. (2019) “[p]reregistration requires research planning and it is hard, especially contingency planning. It takes practice to make design and analysis decisions in the abstract, and it takes experience to learn what contingencies



are most important to anticipate. This might lead researchers to shy away from preregistration for worries about imperfection” (p. 817). Note that other researchers have claimed the exact opposite, namely that preregistration is easy (Wagenmakers & Dutilh, 2016) and that the Registered Report format saves time (Field et al., 2020).

To date there does not exist an empirical assessment about the experiences and expectations that researchers have concerning the impact of preregistration on their workflow. This study seeks to chart the perceived benefits and drawbacks of preregistration we may learn what motivates researchers to adopt this practice and possibly also what prevents researchers from adopting it. At the same time, researchers’ past experiences with preregistration may be informative for pragmatic would-be adopters. This study concerns two groups of researchers: those who published both preregistered studies and non-preregistered studies and those who only published non-preregistered studies.

## Disclosures

### Data, Materials, and Preregistration

The current study was preregistered on the Open Science Framework; in our project folder, readers can access the preregistration, as well as all materials for both the pilot and the main survey, the contact database used for the main survey, the anonymized raw and processed data (including relevant documentation), and the R code to conduct all analyses (including all figures; see Table 1 for an overview of URLs for the different resources). In our datasets, identifying information such as names and affiliations of the respondents were removed. Any deviations from the preregistration are mentioned in this manuscript. Note that we removed email addresses from the contact database for privacy reasons.

### Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Table 1

*Overview of URLs to this Study’s Materials Available on the Open Science Framework.*

Resource	URL
Project page	<a href="https://osf.io/jcdvb/">https://osf.io/jcdvb/</a>
Preregistration of main study	<a href="https://osf.io/qezv5/">https://osf.io/qezv5/</a>
Preregistration of pilot study	<a href="https://osf.io/g3fv7/">https://osf.io/g3fv7/</a>
Data and analysis code	<a href="https://osf.io/5ytpk/">https://osf.io/5ytpk/</a>

Surveys	<a href="https://osf.io/dzybn/">https://osf.io/dzybn/</a>
Ethics documents	<a href="https://osf.io/atgb7/">https://osf.io/atgb7/</a>

---

### **Ethical Approval and Participant Compensation**

The study was approved by the local ethics board of the University of Amsterdam (registration number: 2019-PML-11423) and of the Eotvos Lorand University (registration number: 2019/17). All participants were treated in accordance with the Declaration of Helsinki. Researchers who participated in the survey were given the opportunity to enter a raffle for a voucher from a webshop of their choice.

## **Methods**

### **Pilot Study and Creating Materials**

Before conducting the main survey, we conducted a pilot study to determine the aspects of the research workflow that are most affected by preregistration. For this pilot study we contacted 176 researchers from our database (described in the following sections) and asked them how their preregistered studies differed from their non-preregistered studies in terms of workflow, data management, and scientific quality. Respondents were asked to list both advantages and disadvantages in a free-text format. In total, we received answers from 49 researchers. The answers were then categorized by three of the authors (A.S., B.A., and M.K.). In total, nine aspects of the research process were identified as being especially impacted by preregistration. These aspects of the research process were then included as items in the main survey.

### **Participants**

The researchers in the preregistration group were recruited based on a contact database of published preregistered studies. Initially, we created a collection of 711 research articles in which the authors referred to a preregistered analysis plan. This collection of studies consisted of 404 preregistered and published articles that were part of the bibliographical collection of published preregistered articles from the Center of Open Science (COS), 128 articles mentioned in van den Akker et al. (2021) which originated from a database of articles with open science badges by Kambouris et al. (2020), 22 articles based on a collection from Schäfer and Schwarz (2019), and 157 articles based on a non-systematic collection of the present authors. From this initial collection of articles, we then excluded non-empirical studies (e.g., meta-analyses), Registered Reports, articles that did not include a URL to their preregistration, articles whose preregistration has been published on platforms other than the OSF (e.g., AsPredicted.org), and duplicates. This left a final sample of 487 articles from which we extracted the email-addresses of the corresponding authors.

## Sampling Plan

No sample size target was specified for the preregistration group; we contacted all authors from our contact database. For the non-preregistration group, we preregistered that data would be collected until we reached a sample size as large as at least 90% of the sample size from the preregistration group. As will be discussed in the section “Sample Characteristics”, we were unable to reach that goal.

## Materials

The survey was generated using the online survey software Qualtrics (Qualtrics, 2021). The items in the main survey were based on the results of the pilot study and a discussion among the authors. The survey included questions about (1) the nine aspects of the research process that were identified in the pilot study; (2) the respondents’ general opinion about preregistration; and (3) the respondents’ research background. Respondents from the preregistration group were instructed to relate the questions to their own *experience* (i.e., “Please indicate below how you believe preregistration has affected your work.”), whereas researchers from the non-preregistration group were instructed to indicate their *expectations* about preregistration (e.g., “Please indicate below how you believe preregistration would affect your work.”). Finally, respondents also had the opportunity to give feedback on the survey and provide us with free-text on the topic of preregistration.

Table 2

*Nine aspects of the research process included in the survey as presented to the preregistration group. Respondents were asked to on the following 1 to 7 scales, how they believed preregistration has affected their work. Researchers in the non-preregistration group were asked how they believed preregistration would affect each aspect.*

the	Response Anchors of the 7-Point Rating Scales Due to Preregistration,	
(1)	(7)	
Analysis Plan	<i>got less thought-through</i>	<i>got more thought-through</i>
Research Hypothesis	<i>got less thought-through</i>	<i>got more thought-through</i>
Experimental Design	<i>got less thought-through</i>	<i>got more thought-through</i>
Preparatory Work (e.g., pilot or simulation studies)	<i>got worse</i>	<i>improved</i>

Data Management	<i>got less thought-through</i>	<i>got more thought-through</i>
Project Workflow	<i>got less thought-through</i>	<i>got more thought-through</i>
Collaboration in the Team	<i>got worse</i>	<i>got better</i>
Work-related Stress	<i>was increased</i>	<i>was reduced</i>
Total Project Duration	<i>was longer</i>	<i>was shorter</i>

---

**Nine Aspects of Research Process.** Respondents were asked to indicate whether preregistration has benefited or harmed (preregistration group) or would benefit or harm (non-preregistration group) the nine aspects of the research process listed in Table 2. For each question, respondents could also select the options *I do not know* and *Not applicable*.

**Opinion About Preregistration.** Three items asked respondents about their general opinion concerning preregistration. The first item asked about whether respondents thought preregistration has made it easier (preregistration group) or would make it easier (non-preregistration group) to avoid questionable research practices. The item was answered using a 7-point Likert scale from 1 (*Very Strongly Disagree*) to 7 (*Very Strongly Agree*). The second item asked how often respondents would consider preregistration in their future work. The item was answered using a 7-point Likert scale from 1 (*Always*) to 7 (*Never*). The third item asked about whether respondents would recommend preregistration to other researchers in their field. The item was answered using a 7-point Likert scale from 1 (*Very Strongly Disagree*) to 7 (*Very Strongly Agree*). For items one and three, respondents could also select the options *I do not know* and *Not applicable*.

**Respondents' Research Background.** Two items asked respondents about their research background. The first item asked respondents to categorize their main research approach into either (1) hypothesis testing, (2) estimation, (3) modelling/simulations, (4) qualitative research, or (5) other. The second item asked respondents to write down their specific research background (e.g., developmental psychology) as free text.

## Procedure

Responses from the preregistration group were elicited by contacting all authors in our database (including the ones who participated in the pilot survey). Then, for each author in the preregistration group we contacted up to five authors who published a non-preregistered empirical study in the same journal, volume, and issue. When we did not reach the desired sample size for the non-preregistration group, we proceeded to contact authors who had published in previous

issues of the journals. This procedure was repeated several times and stopped when we had invited almost 2,000 authors to our study. The decision to discontinue data collection deviates from our preregistered sampling plan but was motivated by the limitations of time and resources.

In the main survey, respondents were first asked to indicate if they had ever (1) preregistered a study that was not published; (2) preregistered a study that was published; (3) published a study that was neither preregistered nor a Registered Report; (4) created a Registered Report that was not published; or (5) published a Registered Report. Based on their answers, the respondents were assigned to groups. Respondents were assigned to the preregistration group if they had published both preregistered and non-preregistered studies (i.e., they answered “yes” to both option 2 and 3). Respondents were assigned to the non-preregistration group if they had published exclusively non-preregistered studies (i.e., answered “yes” to option 3 and “no” to all other options). In accordance with the preregistration plan, we only analyze and report data from these two groups.

Respondents then answered the remaining survey items and one intermediate attention check item (i.e.,  $2 + 2 = ?$ ). The survey items and the attention check were presented in fixed order to the participants. The median amount of time respondents took to fill out the questionnaire was 3 minutes and 18 seconds.

### **Data Exclusions**

As preregistered, we excluded respondents if (1) they were assigned neither to the preregistered group nor to the non-preregistered group ( $n = 99$ ); (2) they did not answer all questions in the survey ( $n = 23$ ); (3) they failed the attention check ( $n = 18$ ); (4) they indicated in the comment section that they could not provide adequate responses or they did not accept the informed consent form ( $n = 0$ ).<sup>38</sup> In total, we received 495 responses to our survey. After exclusion, 355 responses remained for the analysis. Of these, 299 responses came from the preregistration group and 56 responses came from the non-preregistration group.

---

<sup>38</sup> Note that exclusion criterion (1) also pertains to respondents who indicated that their experience with preregistration related solely to Registered Reports (i.e., they responded “yes” to options 4 or 5, but “no” to all other options). We decided to exclude these respondents ( $n = 2$ ) since we suspected that secondary benefits of the Registered Reports format might be influenced in large part by the extensive review process.

## Analysis

This is an exploratory study and therefore we present our results mainly through descriptive statistics. For the questions relating to nine aspects of the research process, we report both the means and 95% confidence intervals (Figure 1). Note that the presence of confidence intervals deviates from our preregistration, which stated that no inferential procedure was going to be used.<sup>39</sup> For the questions on the respondents' opinion on preregistration, we visualize the frequency distributions of the survey responses (Figure 2). We preregistered the intention to compare, both within the preregistration group and non-preregistration group, the answers of those who choose hypothesis testing as their empirical approach to the answers of those who choose a different approach (i.e., estimation, modelling/simulations, qualitative research, or other). Due to low response rate in the non-preregistration group we could execute the intended comparison only within the preregistration group (as the sample size in the non-preregistration group was simply too small). We present the results of this comparison in Appendix B. To foreshadow the results, the answers from the hypothesis testing group did not differ notably from those of the other group. For our analyses, we excluded responses that indicated *I do not know* and *Not applicable*. Finally, we compared the responses of the preregistration and non-registration group with respondents who reported having experience with preregistration but were not (yet) able to publish the studies they preregistered. This comparison was not preregistered but was suggested by the relatively high number of respondents that could not be assigned to either the preregistration or the non-registration group ( $n = 99$ ). The results, reported in Appendix C, show that the perceptions of researchers with unpublished preregistrations fall in between those with published preregistrations and the group without preregistration experience.

## Results

### Sample Characteristics

We first sent 487 e-mail invitations to our contact database of researchers with experience in preregistration (see the Method section for a description). Out of these 487 e-mails, 30 bounced (i.e., there was an automatic failure to deliver the e-mail, for instance, because an address was no longer active), yielding a total of 457 successfully delivered requests. Removing incomplete

---

<sup>39</sup> Since we had not made any predictions about our data, we did not preregister inferential procedures, but found it informative to display the statistical uncertainty associated with the mean ratings.

surveys and respondents who failed the attention check left a total sample of 299 respondents who had experience with preregistration (i.e., a response rate of 65.43%).

Next, we invited a total of 1,999 researchers who had published only non-preregistered studies. Out of these 1,999 e-mails, 146 bounced, yielding a total of 1,853 successfully delivered requests. The response rate for the non-preregistration group was lower than anticipated; receiving 56 responses from 1,999 authors yields a response rate of only 2.80%. Due to this low response rate, we were unable to reach the preregistered target sample size, that is, for the non-preregistration group we only reached 18.7% of the number of responses from the preregistration group instead of the preregistered target of 90%.

Most respondents had a background in psychological science. Specifically, out of the 389 reported research backgrounds (some respondents reported more than one), 112 could be classified as social psychology (28.79%), 104 as experimental and cognitive psychology (26.74%), 36 as developmental and educational psychology (9.25%), 32 as personality psychology (8.23%), 17 as neurophysiology and physiological psychology (4.37%), 15 as applied psychology (3.86%), 12 as clinical psychology (3.08%), and 4 as methodology and statistics (1.03%). The remaining 57 responses (14.7%) could not be categorized into one of the areas above (e.g., anesthesiology).

Out of the combined total of 355 respondents, 291 respondents indicated that hypothesis testing was their primary research approach, 21 indicated estimation, 25 indicated modelling/simulations, 3 indicated qualitative research, and 15 respondents indicated other approaches.

**Nine Aspects of Research Process.** Figure 1 illustrates how preregistration was perceived to influence the nine different aspects of the research process. The specific breakdown of the answers to the individual questions is shown in Table 3. Overall, both groups have a positive opinion on how preregistration influenced or would influence the different aspects of the research process, with the preregistration group generally being more positive than the non-preregistration group. Specifically, respondents were most positive about the benefits of preregistration regarding the analysis plan, the hypotheses, and the study design. For two aspects, however, respondents perceived preregistration to be disadvantageous: specifically, respondents indicated that preregistration would increase both work-related stress and total project duration.

Table 3

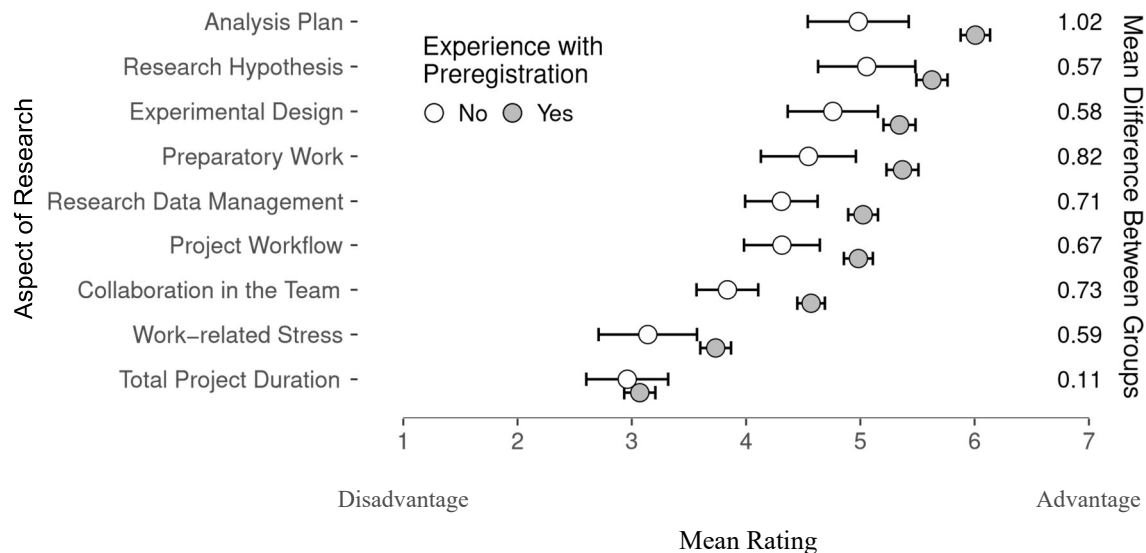
Per group, the mean ratings and 95% confidence intervals for each individual aspect on the research workflow measured on a 7-point rating scale, as well as the number of respondents answering I do not know or Not applicable on each aspect.

Aspect	Experience with preregistration	Rating	Nr. respondents	
			“I do not know”	“Not applicable”
Analysis Plan	Yes	$M = 6.01[5.88,6.14]$	0	0
	No	$M = 4.98[4.54,5.42]$	1	0
Research Hypothesis	Yes	$M = 5.63[5.49,5.77]$	1	1
	No	$M = 5.06[4.63,5.49]$	2	0
Experimental Design	Yes	$M = 5.34[5.20,5.48]$	1	3
	No	$M = 4.76[4.37,5.15]$	1	1
Preparatory Work	Yes	$M = 5.37[5.23,5.51]$	2	4
	No	$M = 4.55[4.14,4.96]$	1	0
Research Data Management	Yes	$M = 5.02[4.89,5.15]$	2	4
	No	$M = 4.31[3.99,4.63]$	1	0
Project Workflow	Yes	$M = 4.98[4.85,5.11]$	5	2
	No	$M = 4.31[3.98,4.64]$	5	0
Collaboration in the Team	Yes	$M = 4.57[4.45,4.69]$	5	4
	No	$M = 3.84[3.57,4.11]$	6	1
Work-related Stress	Yes	$M = 3.73[3.59,3.87]$	5	1
	No	$M = 3.14[2.71,3.57]$	6	0
Total Project Duration	Yes	$M = 3.07[2.93,3.21]$	11	1
	No	$M = 2.96[2.60,3.32]$	6	0

Note. Square brackets indicate the 95 % confidence interval for the ratings.  $N = 299$  for preregistration group,  $N = 56$  for non-preregistration group.

The preregistration group and the non-preregistration group differed mostly in their opinion on how preregistration influences the analysis plan and preparatory work. Although both groups reported that preregistration would benefit these aspects, respondents with preregistration experience were more enthusiastic. That is, the preregistration group reported that preregistration had made the analysis plan more thought-through ( $M = 6.01[5.88,6.14]$  versus  $M = 4.98[4.54,5.42]$ ) and that preregistration improved the preparatory work of the project ( $M = 5.37[5.23,5.51]$  versus  $M = 4.55[4.14,4.96]$ ).





*Figure 1.* Respondents' opinion on how preregistration influenced different aspects of the research process. Grey dots represent the mean ratings from respondents who have experience with preregistration and white dots represent the mean ratings from respondents who have no experience with preregistration. The square skewers represent 95% confidence intervals. Ratings above and below 4 indicate that preregistration helped and harmed a certain research aspect, respectively.

In four aspects of the research process, that is, research hypothesis, experimental design, work-related stress, and total project duration, the groups showed the smallest differences of opinion. Whereas both groups perceived preregistration to benefit the experimental design ( $M = 5.34[5.20, 5.48]$  in the preregistration group versus  $M = 4.76[4.37, 5.15]$  in the non-preregistration group) and the research hypothesis ( $M = 5.63[5.49, 5.77]$  in the preregistration group versus  $M = 5.06[4.63, 5.49]$  in the non-preregistration group), preregistration was perceived to be a disadvantage with respect to work-related stress ( $M = 3.73[3.59, 3.87]$  in the preregistration group versus  $M = 3.14[2.71, 3.57]$  in the non-preregistration group) and total project duration ( $M = 3.07[2.93, 3.21]$  in the preregistration group versus  $M = 2.96[2.60, 3.32]$  in the non-preregistration group).

One aspect in which both groups gave qualitative different answers based on the group means was the influence of preregistration on the collaboration in the team. While respondents in the preregistration group indicated that it had improve the collaboration in the team ( $M = 4.57[4.45, 4.69]$ ), respondents in the non-preregistration group indicated that it would be a slight disadvantage ( $M = 3.84[3.57, 4.11]$ ).

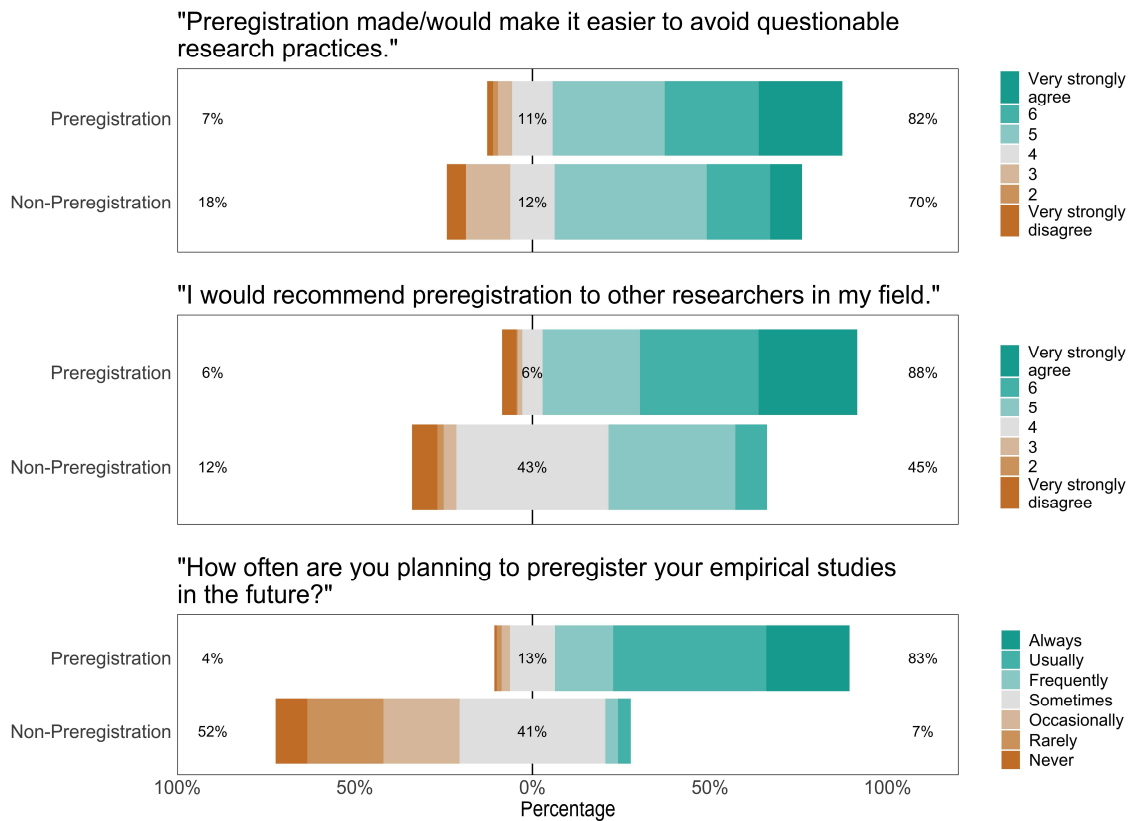
**Opinion About Preregistration.** Figure 2 summarizes the general opinion about preregistration among respondents. The vast majority of respondents in the preregistration group

had a positive overall opinion about the practice. 82% of respondents agreed with the statement that compared to their non-preregistered work, preregistration had helped avoid questionable research practices. For this statement, no researcher responded with *Not applicable* and one researcher responded with *I do not know*. A quarter of respondents (23.5%; 70 of 298) reported to *very strongly agree* with this statement, which may suggest that other researchers have at least some reservations that preregistration is the ultimate solution to preventing questionable research practices.

In addition, 88% of respondents would recommend the practice to other researchers in their field. No researchers indicated *I do not know* or *Not applicable* to this statement. Finally, 83% of the respondents in the preregistration group would consider preregistration in their future work. The results are somewhat more ambiguous in the group of respondents without preregistration experience. Although 70% agreed with the statement that preregistration would make it easier to avoid questionable research practices (with only 9%, that is, 5 of 56, indicating to *very strongly agree* with the statement), only 45% would recommend the practice to other researchers in their field. No researchers in the non-preregistration group indicated *I do not know* or *Not applicable* to these statements. Preregistration is also not seen as desirable for future research projects: only 7% in the non-preregistration group would consider this practice in their future work.

### **Constraints on Generality**

The present study surveyed researchers who have experience with preregistering studies and those who did not. Our sample consisted exclusively of researchers in the field of psychology, presumably from differing career stages. The biggest concern regarding generalizability is that our sample was subject to self-selection. Since participation in the survey was voluntary, researchers who already had a strong opinion about preregistration might have been more likely than others to participate.



*Figure 2.* Respondents’ general opinion about preregistration. The top bar represents answers from respondents who have experience with preregistration, and the bottom bar represents answers from respondents who have no experience with preregistration. For each survey question, the number to the left of the data bar (in brown/orange) indicates the percentage who (slightly or strongly) disagreed or who would recommend preregistration occasionally or less frequently. The number in the center of the data bar (in grey) indicates the percentage who responded with “neither agree or disagree” or “neutral”. The number to the right of the data bar (in green/blue) indicates the percentage who (slightly or strongly) agreed or who would recommend preregistration frequently or more.

Since the proportion of respondents in the preregistration group was relatively high with 65.43%, we assume that our sample therefore reflects the population of these researchers relatively well. Therefore, we expect the results from respondents in the preregistration group to generalize to other researchers within the field of psychology who have experience with preregistration.

The results from the non-preregistration group, on the other hand, might generalize poorly to other researchers in the field since the proportion of respondents in the non-preregistration group was very low (2.80%). In the field of meta-science, low response rates are no exception: Field et al. (2020), for instance, achieved a response rate of 6%, Malički et al. (2022) a response rate of 4.9%. Gopalakrishna et al. (2021), on the other hand, achieved an exceptional high response rate of over 21%. The low response rate in our study suggests that for the non-preregistration group

self-selection might have had a stronger effect on the results. That is, it may be that predominantly researchers with strong opinions about preregistration responded to this survey, rather than those who felt neutral about the practice. However, it should be noted that despite the low response rates in the non-preregistration group the general response pattern (that is, the ranking of the research aspects) is consistent in both groups. This systematicity might indicate that we were not dealing with a select subgroup, or at least that the opinions of the select subgroup do not differ much from researchers with preregistration experience.

### **Discussion**

In the last decade, preregistration has been advocated as a tool to prevent researchers' biases and expectations from contaminating the statistical analyses. It has also been argued that preregistration may have secondary effects on the research process. The current study sought to unveil these expectations and experiences.

Our results suggest that researchers find preregistration to benefit their work in most aspects of the research process. Researchers in our sample reported that preregistration improved the theoretical aspects of the project (e.g., the generation of the research hypothesis, the research design, and the analysis plan) as well as practical aspects of the project (e.g., the design and execution of pilot or simulation studies, and the general project workflow). However, disadvantages of preregistration also became apparent; researchers reported that preregistering a study had increased or was expected to increase the total project duration and the work-related stress.

The increase in time and effort to publish a preregistered study had been acknowledged in the literature (e.g., Nosek & Lindsay, 2018; van 't Veer & Giner-Sorolla, 2016). However, some statements made previously on the influence of preregistration on workrelated stress contradict our findings. For instance, Frankenhuis and Nettle (2018, p.441) write: "From hearsay and our own experience, we think that scholars find it relaxing not to have to make [...] critical decisions after having seen the data, accompanied by a lingering sense of guilt, while cognizant of some of their biases and frustratingly unaware of others."

Although researchers with preregistration experience reported that this practice increased the total project duration and work-related stress, the vast majority of this group also indicated that they would recommend the practice to other researchers in their field, and continue to use it for their own research projects. As one respondent mentioned in the free-text comments: "Pre-Reg

improves quality, which causes more work, as it should be". For researchers without preregistration experience, the equation does not seem to add up: the majority of this group would not recommend the practice to their peers, or consider this practice for themselves in the future.

We identified three limitations of the study. The first limitation is that our survey was based on self-report and therefore cannot demonstrate the extent to which the perceived secondary effects of preregistration correspond to its actual secondary effects. To answer this question, workflows and manuscripts from preregistered and non-preregistered studies would need to be evaluated by independent researchers. To avoid potential sample bias, this could be done in an experimental setting: research teams could be randomly assigned to the preregistration group or the non-preregistration group and be instructed to design and conduct a study to answer the same research question. An appropriate setting for such an experiment would be, for instance, a multi-lab project conducting conceptual replications.

The second limitation concerns the low response rate and small sample size of the non-preregistration group. One explanation for this could be that, of the researchers who do not have experience with preregistration, only those who already have strong opinions about the practice are inclined to answer a preregistration survey. For researchers who are neutral about preregistration, a survey on this topic may simply not be interesting enough.

Perhaps the researchers were also averse to the way we approached them, perhaps our invitation email was worded too strongly in favor of preregistration (our invitation letters can be accessed at <https://osf.io/t376k/>), or it was off-putting that the survey was signed by known proponents of preregistration (i.e., the email was signed by all co-authors and sent from B. A.'s private email account). In fact, the meta-scientific survey study by Gopalakrishna et al. (2021) which had a remarkably high response rate of 21.1% had the data collection conducted by an international market research company.

The last limitation concerns to the wording of the items in this survey. In the current study, respondents in the preregistration group were asked about their experiences with their previous research projects, whereas respondents in the non-preregistration group were asked about their expectations for future research. We opted for this phrasing as we intended to capture the actual effects of preregistration on workflow in the preregistration group, which might arguably be less subject to bias than expected secondary effects. However, this wording may have reduced comparability between the two groups. Future research might therefore consider asking respondents in the preregistration group additionally about their expectations for future projects.

How can researchers benefit from the secondary effects of preregistration? Whether or not preregistration improves the secondary aspects of the research process depends largely on the quality of the preregistration document. That is, the thoroughness of the preregistration protocol determines how carefully researchers need to think about the study design and analysis plan. A high-quality preregistration document features detailed information about the experimental conditions, the materials and stimuli used, and a comprehensive analysis plan (preferably featuring a mock data set and analysis code). To ensure that preregistration protocols meet these quality standards without considerable extra effort, researchers can fall back on a range of checklists, guidelines, and preregistration templates. Preregistration templates for the standard experimental framework can be found, for instance, on the websites [aspredicted.org](https://aspredicted.org) or on the Open Science Framework (<https://osf.io/zab38/>). The number of preregistration templates and tutorials for other research areas and more complex methods is increasing and includes cognitive modeling (Crüwell & Evans, 2019), secondary data analysis of pre-existing data (Mertens & Krypotos, 2019; Van den Akker et al., 2021), studies using experience sampling methods (Kirtley, Lafit, Achterhof, Hiekkaranta, & MyinGermeys, 2021), and qualitative research (Haven & van Grootel, 2019; Haven et al., 2020). Finally, the recently developed Transparency Checklist is a quick way to check whether the preregistration and the accompanying paper comply with the current transparency standards (Aczel et al., 2020).

Some researchers might also prefer alternative methods to preregistration. One of these alternatives that allows for more flexibility while still safeguarding the confirmatory status of the research is analysis blinding (MacCoun, 2020; MacCoun & Perlmutter, 2015; MacCoun & Perlmutter, 2018; Dutilh, Sarafoglou, & Wagenmakers, 2019). With analysis blinding, researchers are in principle not required to write a preregistration document. Instead, they collect their experiment data as usual and develop their analysis plan based on an altered version of the data in which the effect of interest is hidden (e.g., by shuffling the outcome variable). Another alternative would be to minimize bias by trying to map out the uncertainty in the analyses with various statistical practices (Wagenmakers et al., 2021). For instance, researchers could explore the entire universe of outcomes through multiverse analyses (in which all theoretically sensible data-preprocessing steps are explored; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) or multi-analysts approaches (in which multiple analysis teams answer the same research question based on the same dataset; e.g., The MARP Team, 2022; Silberzahn & Uhlmann, 2015).

Our survey shows that researchers see preregistration as beneficial to their research workflow and the overall quality of their work. We consider this to be a welcome byproduct of the

practice: one ensures the confirmatory status of one's analyses and experiences an improvement in practical aspects of one's workflow. However, this does not mean preregistration is the preferred means of improving workflow; other methods are probably better suited for this purpose. For instance, the recently proposed theory construction methodology by Borsboom, van der Maas, Dalege, Kievit, and Haig (2021) was developed to assist researchers in identifying and linking empirical phenomena, in constructing and mathematically representing theories, and evaluating these theories. As such, this methodology could likewise improve the quality of the analysis plan, research hypothesis, preparatory work, and experimental design, presumably to a greater extent than preregistration can. Similarly, we expect that the Registered Report format, which entails close scrutiny and revision of theory, experimental design, and analysis plan by independent scholars, could achieve greater secondary benefits than preregistration alone.

Researchers who have strong reservations about preregistration, whether conceptual or practical, are unlikely to be persuaded by the experiences of their peers. However, those who are still undecided whether the practice is worth trying may be convinced by its practical advantages. To them we say: try preregistration and form your own opinion about its possible advantages and disadvantages.

In order for preregistration to truly become the norm in psychology, it is necessary for journals, institutions, and funding agencies to provide sufficient incentives for researchers. In addition, we believe that the research culture still needs to evolve: in terms of making preregistration considered good research practice in individual labs, but also in terms of making sure that studies that cannot be preregistered are not stigmatized. Some of the negative experiences that have been made with preregistration could possibly be reduced with methodological advancements. For instance, combining preregistration with analysis blinding might increase the adherence to analysis plans. Better-structured templates could improve the efficiency of the method, and more precise instructions could increase the accuracy of preregistration, thereby also increasing its effectiveness.

**Concluding Remarks.** The aim of this study was to obtain an overview of the experienced and expected advantages and disadvantages of the practice of preregistration. Our survey shows that relying on intuition alone when developing open research practices might not be enough. Only if we know how the conceptual advantage of preregistration weighs against the individual experienced benefits and challenges can we find suitable means to improve the methodology so that it finds wider acceptance among researchers.

**Author Contributions**

Contributorship was documented with CRediT taxonomy using tenzing (Holcombe, Kovacs, Aust, & Aczel, 2020).

**Alexandra Sarafoglou:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, and Writing - original draft.

**Marton Kovacz:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, and Writing review & editing.

**Bence Bakos:** Data curation and Formal analysis.

**Eric-Jan Wagenmakers:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, and Writing - original draft. **Balazs Aczel:** Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, and Writing - original draft.

**Conflicts of Interest**

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

**Acknowledgements**

We thank Zsuzsa Szekely for her assistance in the project.

**Funding**

This research was supported by a talent grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568), as well as a Vici grant from the NWO to EJW (016.Vici.170.083).



## References

- Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, v., Benjamin, D., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4, 4–6.
- Alistair, M., Vickers-Jones, R., Sewell, D. K., & Ballard, T. (2021). How do we choose our giants? Perceptions of replicability in psychological science. *Advances in Methods and Practices in Psychological Science*, 4, 1–21.
- Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, 16, 756–766.
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at *Cortex*. *Cortex*, 49, 609–610.
- Crüwell, S., & Evans, N. J. (2019). Preregistration in complex contexts: A preregistration template for the application of cognitive models. *Manuscript submitted for publication*. Retrieved from <https://psyarxiv.com/2hykx/>
- Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, 198, S5745–S5772.
- Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, 13, 433–438.
- Field, S. M., Wagenmakers, E.-J., Kiers, H. A., Hoekstra, R., Ernst, A. F., & van Ravenzwaaij, D. (2020). The effect of preregistration on trust in empirical research findings: Results of a Registered Report. *Royal Society Open Science*, 7, 181351.
- Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science*, 13, 439–447.
- Gopalakrishna, G., Wicherts, J., Vink, G., Stoop, I., van den Akker, O., ter Riet, G., & Bouter, L. (2021). Prevalence of responsible research practices and their potential explanatory factors: A survey among academic researchers in The Netherlands. *Manuscript submitted for publication*. Retrieved from <https://osf.io/preprints/metaarxiv/xsn94/>
- Haven, T. L., Errington, T. M., Gleditsch, K. S., van Grootel, L., Jacobs, A. M., Kern, F. G., ... Mokkink, L. B. (2020). Preregistering qualitative research: A Delphi study. *International Journal of Qualitative Methods*, 19, 1609406920976417.
- Haven, T. L., & van Grootel, L. (2019). Preregistering qualitative research. *Accountability in Research*, 26(3), 229–244.
- Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, 15, e0244611.

- Jones, R. (1973). The theory of practical joking – its relevance to physics. In E. Mendoza (Ed.), *A random walk in science: An anthology compiled by the late R L Weber (1914–1997)* (p. 14). Bristol: Institute of Physics Publishing.
- Kambouris, S., Singleton Thorn, F., Van den Akker, O., De Jonge, M., Ruffer, F., Head, A., & Fidler, F. (2020). *Database of articles with Open Science badges: 2020-02-21 snapshot*. Retrieved from <https://osf.io/q46r5>
- Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., ... Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLOS Biology*, *14*, e1002456.
- Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experiencesampling methods. *Advances in Methods and Practices in Psychological Science*, *4*, 1–16.
- MacCoun, R. (2020). Blinding to remove biases in science and society. In R. Hertwig & C. Engel (Eds.), *Deliberate ignorance: Choosing not to know* (pp. 51–64). Cambridge: MIT Press.
- MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth: More fields should, like particle physics, adopt blind analysis to thwart bias. *Nature*, *526*, 187–190.
- MacCoun, R., & Perlmutter, S. (2018). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–322). John Wiley and Sons.
- Malički, M., Aalbersberg, I. J., Bouter, L., Mulligan, A., & ter Riet, G. (2022). Transparency in conducting and reporting research: A survey of authors, reviewers, and editors across scholarly disciplines. *Manuscript submitted for publication*. Retrieved from <https://doi.org/10.21203/rs.3.rs-1296644/v1>
- Mertens, G., & Kryptos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*, 338–352.
- Munafò, M., Nosek, B. A., Bishop, D., Button, K., Chambers, C., Du Sert, N., ... Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229.
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in cognitive sciences*, *23*, 815–818.
- Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.
- Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*, 19–21.
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618.
- Peirce, C. S. (1878). The probability of induction. *Popular Science Monthly*, *12*, 705–718.
- Qualtrics. (2021). *Online survey software Qualtrics*. Retrieved from <https://www.qualtrics.com>

- Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813.
- Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature, 526*, 189.
- Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*, 886–899.
- Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. In J. Wixted & E.-J. Wagenmakers (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience (4th ed.)*, Volume 5: *Methodology* (pp. 297–322). New York: Wiley.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712.
- Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science, 16*, 717–724.
- Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences, 24*, 94–95.
- The MARP Team. (2022). A many-analysts approach to the relation between religiosity and wellbeing. *Manuscript submitted for publication*. doi: 10.31234/osf.io/pbfye
- van den Akker, O., van Assen, M., Bakker, M., Enting, M., de Jonge, M., ... Wicherts, J. (2021). Selective hypothesis reporting - preregistration. *Open Science Framework*. Retrieved from <https://osf.io/z4awv>
- Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., ... Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology, 5*, 2–19.
- van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12. doi: <https://doi.org/10.1016/j.jesp.2016.03.004>
- Wagenmakers, E.-J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer, 29*.
- Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., ... Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour, 5*, 1473–1480.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632–638.

## Appendix A

## Summary of Free-Text Comments

In our survey, respondents both completed the questionnaire and had the opportunity to provide comments on preregistration in an open-ended format. This section summarizes these comments. For this purpose, the authors A.S. and M.K. have divided the comments into different topics and evaluated whether they were positive, negative, or neutral statements. Comments on other topics than preregistration (e.g., comments on the survey) are not here. The full list of comments is available in our online repository at <https://osf.io/5ytpk/>. We would like to emphasize that the results should be interpreted with caution. The comments evaluated below are based on only a fraction of the respondents. Therefore, the overview given here is not necessarily representative of the opinions in our sample.

78 researchers provided us with free-text comments on preregistration. These comments highlighted both the advantages and disadvantages of preregistration: 20 comments were exclusively positive, 22 comments were negative, and 36 comments were mixed. The comments could be categorized roughly into five topics. The topics were (1) the additional workload of preregistration (mentioned by  $n = 24$  respondents); (2) the effectiveness of preregistration in solving the crisis of confidence (mentioned by  $n = 19$ ); (3) the impact of preregistration on one's career (mentioned by  $n = 16$  respondents); (4) how preregistration might contribute to inequality and stigmatization in different research areas (mentioned by  $n = 13$  respondents); (5) and the difficulties in the compliance with the preregistration protocol (mentioned by  $n = 11$  respondents).

**Additional workload of preregistration: harder, but worthwhile?**

Proponents of preregistration argue that despite the additional workload preregistration cases, it is still “worthwhile” (e.g., Nosek & Lindsay, 2018). But do researchers agree with that statement? Not necessarily. From the  $n = 24$  respondents who mentioned the additional workload,  $n = 11$  respondents believed that preregistration was harder and worthwhile while seven respondents believed that it was harder, but not worthwhile—six respondents mentioned the increased workload without any further judgement. For respondents who thought preregistration was hard, but worthwhile, the added benefit of improved overall quality outweighed the added workload or was perceived as necessary consequence (e.g., “Pre-Reg improves quality, which causes more work, as it should be”). Others recognized the theoretical value of preregistration, but did not see the benefits translating into practice. For instance, one respondent wrote: “I think preregistration is great in theory, but in practice it serves only to increase the red tape and time

until publication. In today's hyper-competitive publish-or-perish job market, it amounts to time wasted". The added time it takes to write a preregistration even seems to scare researchers from trying out the practice: "I understand the importance of [preregistration], but the amount of time and effort needed to preregister is probably the biggest reason I have avoided it in the past".

### **Effectiveness of preregistration in solving the crisis of confidence**

19 respondents mentioned that preregistration improved the credibility of their results and the overall quality of their work. Seven respondents, however, questioned whether preregistration was a suitable tool to address the crisis of confidence. Besides the need for theory development and exploratory research, lack of methodological knowledge, and possibilities to cheat the system (by creating multiple preregistration documents) were mentioned. In addition, multiple respondents criticized the incentive structure in science, which is designed to reward research output and thus discourages the adoption of preregistration (e.g., "[U]nless we rid science from the publication for-profit industry and educate our universities not to use the incentive structure that still very much determines who gets hired and who gets promoted based on where researchers publish rather than what they publish, I am afraid we have left the big elephant in the room untouched."; "[T]he speed at which our institutions expect us to pump through graduate students often means that pre-reg cannot happen for their work [...]").

### **Influence of preregistration on the career**

16 respondents reported how preregistration influenced their career. Two respondents indicated that embracing open science practices helped their career, for instance, by giving them an advantage during the hiring process. With respect to research output, five respondents reported that publishing preregistered studies was easier while six respondents reported that it was harder. The main arguments as to why preregistered articles were easier to publish was that the respondents felt that a preregistration was expected by the journals, or they described that the "in principle acceptance" granted for Registered Reports made the publication process easier. On the other hand, respondents also described how reviewers or editors rejected papers if authors did not adhere to their preregistered plan, or that they pushed them towards rewriting their manuscripts to present polished narratives (e.g., "[R]eviewers sometimes have even criticized that I report non-significant results"; "[I] often encounter editors who still seem to want my team to change a priori aspects of manuscripts to better fit with a *we knew it all along* or in the context of competing hypothesis situations, favor the hypothesis that was ultimately supported by the data").

### **Inequality and stigmatization**

In our survey, 13 respondents addressed disadvantages preregistration can have in research fields outside of psychology and for descriptive and exploratory study designs. As mentioned by some respondents, when working in fields outside of psychology (e.g., animal research) or when the research area has interfaces with industry, preregistration is relatively unknown which makes preregistered studies harder to publish (e.g., “[...] My field (animal research) is substantially behind the curve. To date, of the preregistered studies I have attempted to publish, no reviewer has commented on the preregistration as a positive aspect of the study [...]. Rather, the reviewers who have mentioned it have used the preregistration to point out deviations (which we take care to explicitly point out in the methods) and thus has led to more challenges with publication rather than fewer. I am of the opinion that if I had submitted identical studies without preregistration, they would have been easier to

publish. [...]”)

In addition, respondents perceived that preregistration went to the detriment of descriptive and exploratory research. For instance, one respondent argues that confirmatory and preregistered experimental studies are currently perceived as “the gold standard [...] which leaves behind other kinds of exploratory and descriptive studies.” Another respondent argues that psychology “needs a clearer distinction between confirmatory and exploratory work, and wider recognition of the value of exploratory, descriptive research that can form the basis for well-specified hypotheses”. Lastly, five respondents critiqued that preregistration causes stigmatization for studies that have not been preregistered. In their comments, respondents critiqued that the reviewers often prematurely condemn a non-preregistered study, without considering its individual peculiarities. As suggested by one of the respondents, the scientific community should place more emphasis on positive reinforcement rather than harsh judgement (e.g., “I am still in favor of pre-registration and open science and I plan to pre-register the studies that I lead. At the same time, I wish that the movement was more moderate and based more on positive reinforcement”).

### **Problems with data exploration and compliance with the preregistration protocol**

11 respondents commented that preregistration would limit creativity, that it discourages researchers to explore the data and that adherence with the preregistration protocol was problematic, especially for early career researchers “who are still learning as they go”, or when working with complex models (e.g., “In my work it’s hard or sometimes impossible to know how

the data should be analysed before seeing its structure, distribution, etc etc and there is no way of accounting for every possibility in the prereg.”).

## Appendix B

## Hypothesis Testing and Exploratory Research

The following section takes a closer look at the responses within the preregistration group. Specifically, we were interested in whether a researcher's empirical approach influences perceptions of preregistration, for instance, in that researchers who primarily test hypotheses (i.e., focusing mainly on the existence of an effect) view preregistration as more beneficial than researchers with other empirical approaches. Such alternative approaches include parameter estimation (focusing mainly on the size of an effect), qualitative research (focusing mainly on understanding an effect), or modeling/simulations (focusing mainly on development of statistical methods).

Within the preregistration group, 250 respondents indicated that hypothesis testing was their main empirical approach while 49 respondents indicated that their main empirical approach was a different one (e.g., estimation, modeling/simulations, qualitative research, other).

Figure B1 illustrates how preregistration was perceived to influence the nine different aspects of the research process. Overall, both groups have a positive opinion on how preregistration influenced the different aspects research process. The pattern resembles that of the preregistration group in general, with the analysis plan benefiting the most from preregistration while the total project duration and work-related stress have been negatively affected by the practice. Respondents who do hypothesis-testing seemed to be somewhat more negative than respondents with a different empirical approach. The biggest difference in opinion was regarding work-related stress. Here, the hypothesis-testing group perceived preregistration to be a disadvantage ( $M = 3.67[3.52,3.81]$ ), while respondents with a different empirical approach were neutral ( $M = 4.08[3.77,4.40]$ ).

Figure B2 illustrates the general opinion about preregistration among the respondents. The two groups do not show meaningful differences in opinion. In both groups, more than 75% agreed with the statement that compared to their non-preregistered work preregistration helped them avoid questionable research practices and more than 85% would



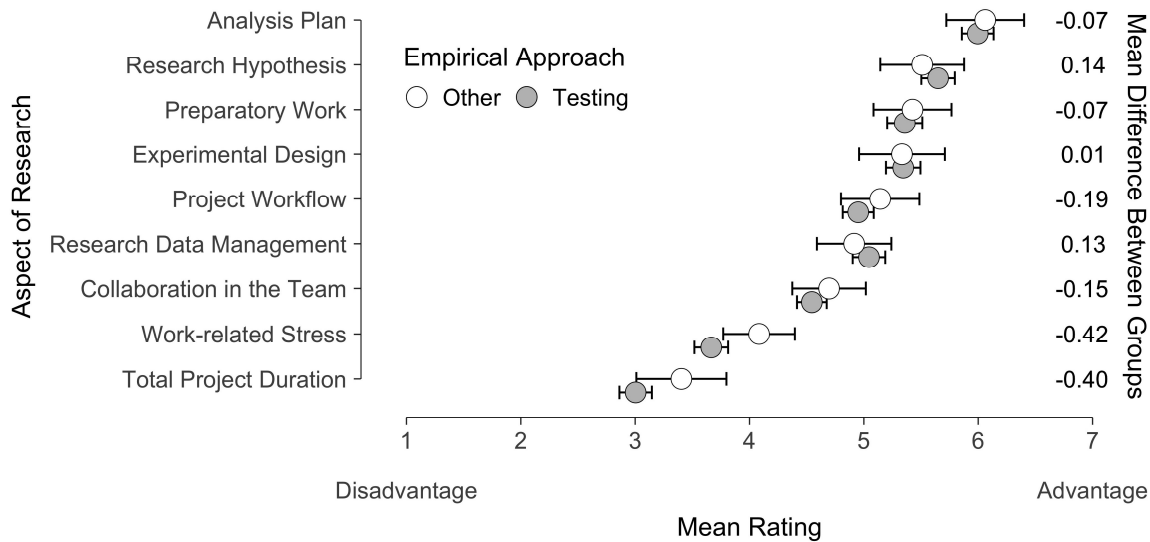


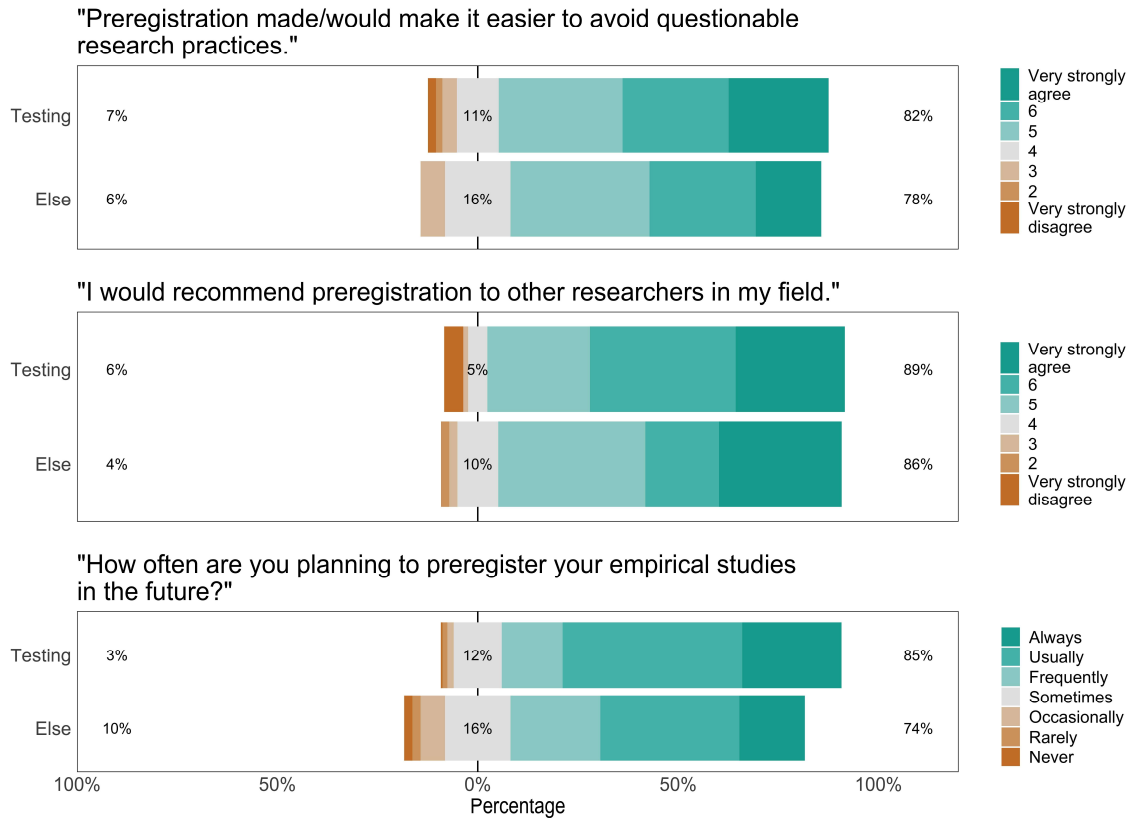
Figure B1. Respondents’ opinion on how preregistration influenced different aspects of the research process. Grey dots represent the mean ratings from the respondents who indicated that their empirical approach was hypothesis testing and white dots represent the mean ratings from respondents who indicated a different empirical approach. The square skewers represent 95% confidence intervals. Ratings above and below 4 indicate that preregistration helped or harmed a certain research aspect, respectively.

recommend the practice to other researchers in their field. Finally, over 85% of the respondents who do hypothesis-testing would consider preregistration in their future work and 73% percent of the respondents with a different empirical approach would consider it in their future work.

### Appendix C

#### Published versus Unpublished Preregistrations

In our main results, all respondents in the preregistration group had at least one positive experience with preregistration in that they successfully published at least one preregistered article. In this section we explore the attitudes of researchers who have not (yet) been able to publish the studies they preregistered. Specifically, we were interested to explore if this group experienced preregistration as particularly frustrating or whether they perceive the practice as positive as researchers who have successfully published a preregistration. This comparison was not preregistered.



*Figure B2.* Respondents’ general opinion about preregistration. The top bar represents answers from respondents whose main empirical approach was hypothesis-testing, the bottom bar represents answers from respondents whose main empirical approach was different. For each survey question, the number to the left of the data bar (in brown/orange) indicates the percentage who (slightly or strongly) disagreed or who would recommend preregistration occasionally or less frequently. The number in the center of the data bar (in grey) indicates the percentage who responded with “neither agree or disagree” or “neutral”. The number to the right of the data bar (in green/blue) indicates the percentage who (slightly or strongly) agreed or who would recommend preregistration frequently or more.

From the 99 respondents who were assigned neither to the preregistration group nor to the non-preregistration group, 63 reported having experience with preregistration but have not published one (yet). Excluding the respondents who have experience with Registered Reports, this left a sample of 55 respondents (henceforth denoted as unpublished preregistration group). Note that from these data it is not possible to deduce why the researchers could not publish their preregistered studies. Their experiences could be based on ongoing studies, or perhaps on studies that were difficult to publish.

Table C1

*For the 55 respondents in the unpublished-preregistration group, the Table shows the mean ratings and 95% confidence intervals for each individual aspect on the research workflow*

measured on a 7-point rating scale, as well as the number of respondents answering I do not know or Not applicable on each aspect.

Aspect	Rating	Nr. respondents	
		"I do not know"	"Not applicable"
Analysis Plan	$M = 5.56[5.21, 5.91]$	0	0
Research Hypothesis	$M = 5.44[5.10, 5.78]$	0	0
Preparatory Work	$M = 5.02[4.65, 5.39]$	1	0
Experimental Design	$M = 4.98[4.65, 5.31]$	0	3
Research Data Management	$M = 4.96[4.63, 5.29]$	0	1
Project Workflow	$M = 4.94[4.63, 5.25]$	0	1
Collaboration in the Team	$M = 4.40[4.14, 4.66]$	1	2
Work-related Stress	$M = 3.32[3.05, 3.59]$	2	0
Total Project Duration	$M = 3.14[2.73, 3.55]$	4	0

*Note.* Square brackets indicate the 95 % confidence interval for the ratings.

Figure C1 shows how respondents rated the effects of preregistration on the nine different aspects of the research process. Table C1 shows a more detailed overview of their responses. As in our previous results, respondents in the unpublished-preregistration group (dark grey dots) have a positive opinion on how preregistration influences the different aspects of the research process. The response pattern in this group resembles that of our main sample, depicted with white dots and light grey dots. The figure suggests that the opinions of respondents in the unpublished-preregistration group lie between those who have published preregistrations and those who have no preregistration experience. Concerning the aspects 'research data management', 'project workflow', and 'collaboration in the team', the group seems closer to the opinions of the preregistration group. In the aspect 'work-related stress', however, the group has a more negative attitude, similar to the non-preregistration group.

Figure C2 illustrates the general opinion about preregistration among the respondents. Again, the opinions of respondents who have only unpublished preregistration experience lie between those who have published preregistrations and those who have no preregistration experience. More than 69% agreed with the statement that preregistration would help them avoid questionable research practices and 80% would recommend the practice to other researchers in their field. Unlike respondents in the non-preregistration group, the majority of respondents in the unpublished-preregistration group plans to use preregistration in future projects (7% versus 65%, respectively).

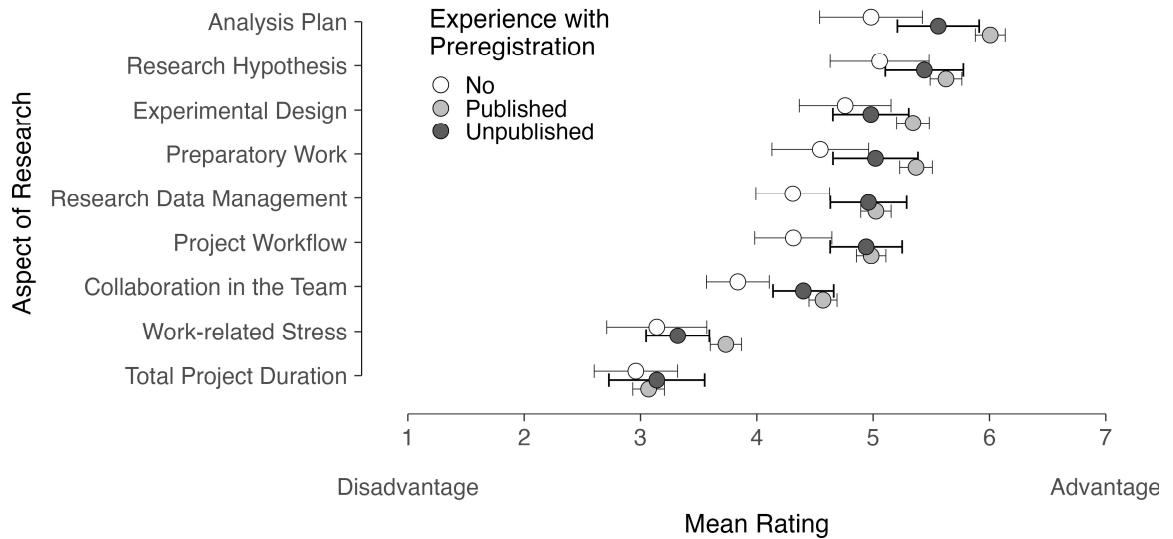


Figure C1. Respondents’ opinion on how preregistration influenced different aspects of the research process. Dark grey dots represent the mean ratings from the  $n = 55$  respondents who have experience solely with unpublished preregistrations. White dots represent the mean ratings from the  $n = 56$  respondents who have no experience with preregistration, light grey dots represent the mean ratings from  $n = 299$  respondents who have published a preregistration. The square skewers represent 95% confidence intervals. Ratings above and below 4 indicate that preregistration helped or harmed a certain research aspect, respectively.

Overall, respondents in the unpublished-preregistration group do not seem to feel frustrated by the process of preregistration. At the same time, this group is somewhat less enthusiastic about the practice than the respondents who have already published a preregistered study.



*Figure C2.* Respondents’ general opinion about preregistration. The top bar represents answers from respondents who have published a preregistration, the middle bar represents answers from respondents who have experience with unpublished preregistrations, and the bottom bar represents answers from respondents who have no experience with preregistration. For each survey question, the number to the left of the data bar (in brown/orange) indicates the percentage who (slightly or strongly) disagreed or who would recommend preregistration occasionally or less frequently. The number in the center of the data bar (in grey) indicates the percentage who responded with “neither agree or disagree” or “neutral”. The number to the right of the data bar (in green/blue) indicates the percentage who (slightly or strongly) agreed or who would recommend preregistration frequently or more.

### 3.3. Seven Steps Toward More Transparency in Statistical Practice<sup>40</sup>

Eric-Jan Wagenmakers<sup>\*1</sup>, Alexandra Sarafoglou<sup>1</sup>, Sil Aarts<sup>2</sup>, Casper Albers<sup>3</sup>, Johannes Algermissen<sup>4</sup>, Štepan Bahník<sup>5</sup>, Noah van Dongen<sup>1</sup>, Rink Hoekstra<sup>6</sup>, David Moreau<sup>7</sup>, Don van Ravenzwaaij<sup>8</sup>, Aljaž Sluga<sup>9</sup>, Franziska Stanke<sup>10</sup>, Jorge Tendeiro<sup>11, 8</sup>, and Balazs Aczel<sup>12</sup>

<sup>1</sup>Department of Psychology, University of Amsterdam, The Netherlands; <sup>2</sup>School for Public Health and Primary Care, Maastricht University, The Netherlands <sup>3</sup>Heymans Institute of Psychological Research, University of Groningen, The Netherlands <sup>4</sup>Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands <sup>5</sup>Faculty of Business Administration, Prague University of Economics, Czech Republic <sup>6</sup>Department of Educational Science, University of Groningen, The Netherlands; <sup>7</sup>School of Psychology and Centre for Brain Research, The University of Auckland, New Zealand; <sup>8</sup>Department of Psychology, University of Groningen, The Netherlands; <sup>9</sup>Rotterdam School of Management, Erasmus University Rotterdam, The Netherlands; <sup>10</sup>Department of Psychology, University of Münster, Germany; <sup>11</sup>Office of Research and Academia-Government-Community Collaboration Education and Research Center for Artificial Intelligence and Data Innovation Hiroshima University ; <sup>12</sup>Institute of Psychology, ELTE Eotvos Lorand University, Hungary

#### Abstract

We argue that statistical practice in the social and behavioral sciences benefits from transparency, a fair acknowledgement of uncertainty, and openness to alternative interpretations. To promote such a practice, we recommend seven concrete statistical procedures: (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We discuss their benefits and limitations, and provide guidelines for adoption. Each of the seven procedures finds inspiration in Merton's ethos of science as reflected in the norms of communalism, universalism, disinterestedness, and organized skepticism. We believe that these ethical considerations –and their statistical consequences– establish common ground among data analysts, despite continuing disagreements about the foundations of statistical inference.

---

<sup>40</sup> Published as:

Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical practice. *Nature Human Behaviour*, 5(11), 1473–1480.

## INTRODUCTION

A superficial assessment of the published literature suggests that statisticians rarely agree on anything. Different schools –mostly frequentists, likelihoodists, and Bayesians– have fought one another tooth and nail for decades, debating the meaning of “probability”, arguing about the role of prior knowledge, disputing the value of objective vs. subjective analyses, and disagreeing about the primary goal of inference itself: whether researchers should control error rates, update beliefs, or make coherent decisions. Fundamental disagreement exists not only between the different statistical schools, but is also present within the same school. For instance, within the frequentist school there is the perennial debate between those who seek to test hypotheses through  $p$ -values and those who emphasize estimation through confidence intervals; and within the Bayesian school, Jack Good’s claim that there are 46,656 varieties of Bayesians may prove an underestimate (<sup>1</sup>; but see<sup>2</sup>).

The disagreement also manifests itself in practical application, whenever multiple statisticians and practitioners of statistics find themselves independently analyzing the same data set. Specifically, recent “multiple-analyst” articles show that statisticians rarely used the same analysis, and often drew different conclusions, even for the exact same data set and research question<sup>3-7</sup>. Deep disagreement is also exhibited by contradictory guidelines on  $p$ -values (e.g.,<sup>8-13</sup>). Should practitioners avoid the phrase “statistically significant”? Should they lower the  $p$ -value thresholds, or justify them, or abandon  $p$ -values altogether? And if  $p$ -values are abandoned, what should replace them? With statisticians fighting over these fundamental issues, users of applied statistics may be forgiven for adopting a wait-and-see attitude and carrying on as usual.

In this article, we claim that besides the numerous disputes and outstanding arguments, statisticians might agree on a set of scientific norms. We bring these norms to the fore, as we believe that they have considerable relevance for the practice of statistics in the social and behavioural sciences. The norms which we believe should guide statistical practice are communalism, universalism, disinterestedness, and organized skepticism, which are the four scientific norms proposed by Merton (1973)<sup>14</sup> (originally published in 1942; see the textbox for a detailed overview of the Mertonian norms).

In general, when Mertonian norms are carried over to the field of statistics, general themes include the need to be transparent, to acknowledge uncertainty, and to be open to alternative interpretations. As such, the Mertonian norms, although proposed over half a century ago,

embody the current aspirations to increase the transparency and reproducibility of science. Critically, the principles behind the Mertonian norms can be translated into concrete statistical practices. A non-exhaustive list of these practices include (1) visualizing data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; (7) sharing data and code. We believe that most statisticians would generally endorse these practices<sup>15</sup>, barring reasonable exceptions (e.g., privacy concerns, severe restrictions of time and money). In this article, we will explain these practices in more detail, including their benefits, limitations and guidelines.

### Box 1

#### Merton's Ethos of Science

Merton<sup>14</sup> proposed that scientific ethos is characterized by the following four norms:

1. Communalism. "The substantive findings of science are a product of social collaboration and are assigned to the community. (...) Property rights in science are whittled down to a bare minimum by the rationale of the scientific ethic. (...) The institutional conception of science as part of the public domain is linked with the imperative for communication of findings. Secrecy is the antithesis of this norm; full and open communication its enactment."<sup>14</sup>, pp. 273–274
2. Universalism. "truth-claims, whatever their source, are to be subjected to *preestablished impersonal criteria*: consonant with observation and with previously confirmed knowledge. The acceptance or rejection of claims entering the lists of science is not to depend on the personal or social attributes of their protagonist; his race, nationality, religion, class, and personal qualities are as such irrelevant."<sup>14</sup>, p. 270; italics in original
3. Disinterestedness. "Science, as is the case with professions in general, includes disinterestedness as a basic institutional element. (...) A passion for knowledge, idle curiosity, altruistic concern with the benefit to humanity (...) have been attributed to the scientist."<sup>14</sup>, pp. 275-276
4. Organized Skepticism. This "involves a latent questioning of certain bases of established routine, authority, vested procedures and the realm of the "sacred" generally. (...) Science which asks questions of fact concerning every phase of nature and society comes into psychological, not *logical*, conflict with other attitudes toward these same data which have been crystallized and frequently ritualized by other institutions. Most institutions demand unqualified faith; but the institution of science makes skepticism a virtue."<sup>14</sup>, p. 264–265; italics in original



## VISUALIZING DATA

### 1.1 *Description*

By visualizing data, researchers can graphically represent key aspects of the observed data as well as important properties of the statistical model applied.

### 1.2 *Benefits and Examples*

Data visualization is important in all phases of the statistical workflow. In exploratory data analysis, data visualization helps researchers formulate new theories and hypotheses<sup>16</sup>. In model assessment, data visualization supports the detection of model misfit and guides the development of appropriate statistical models (e.g., <sup>17-21</sup>). Finally, once the analysis is complete, visualization of data and model fit is arguably the most effective way to communicate the main findings to a scientific audience<sup>22</sup>.

For an example of how data visualization facilitated the development of a new hypothesis, consider the famous “map of the distribution of deaths from cholera” created by London anaesthetist Dr. John Snow during the cholera outbreak in Soho, London in September 1854. In order to trace the source of the outbreak, Dr. Snow created a dot map that displayed the homes of the deceased as well as the water pumps in the neighborhood (Figure 1). The scatter of the data showed that the deaths clustered around a particular water pump in Broad Street, suggesting that the disease was waterborne instead of airborne<sup>23</sup>. Upon Dr. Snow’s request, the pump was disabled by removing its handle, which immediately ended the neighbourhood epidemic. It was discovered later that the well belonging to the pump was contaminated with sewage, which caused the outbreak in the neighborhood.

For an example of how data visualization can reveal model misspecification, consider



Figure 1: Recreation of Dr. Snow’s map of the distribution of deaths from cholera. In this map, the points represent the homes of the deceased and the crosses represent the water pumps in the neighborhood. The contaminated water pump that triggered the cholera epidemic in the neighborhood is located on Broad Street. Reprinted with permission from *Pioneer maps of health and disease in England* (p. 174), by E. W. Gilbert, 1958, The Royal Geographical Society (with the Institute of British Geographers).

Anscombe’s quartet<sup>24</sup> shown in Figure 2. The four scatter plots all have identical summary statistics (i.e., means, standard deviations, and Pearson correlation coefficient). By visually inspecting the panels, it becomes obvious that the bivariate relation is fundamentally different for each panel (see also<sup>25</sup>).

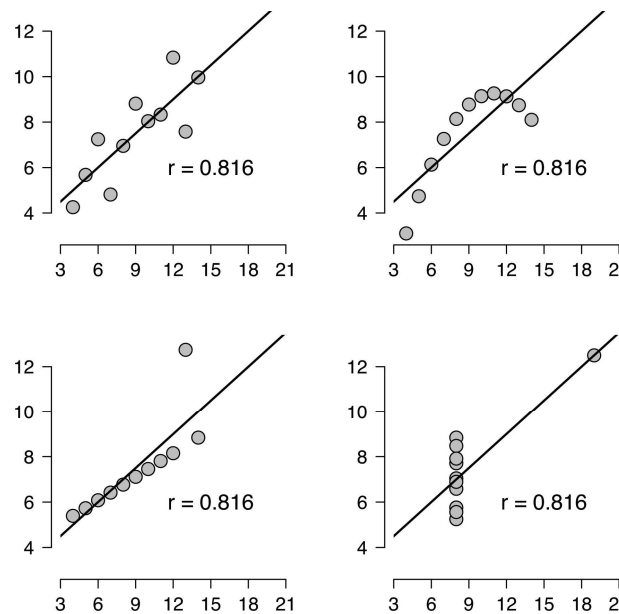


Figure 2: Anscombe’s quartet emphasizes the importance of data visualization to detect model misspecification. Although the four data sets are equivalent in terms of their summary statistics, the Pearson correlation is only valid for the data set in the upper left panel.

### **1.3 *Current Status***

Since William Playfair (1759–1823) invented the first statistical graphs –such as line graphs and bar charts<sup>26</sup>–, data visualization has become an essential part of science. Today, graphs are part of most statistical software packages and have become an indispensable tool to perform certain analyses (e.g., principal component analysis, or prior and posterior predictive checks), or for handling big data sets (e.g., through cluster analysis<sup>27</sup>). Technology now allows us to go beyond static visualizations and display the dynamic aspects of the data, for instance, by using the software packages R Shiny<sup>28</sup> or iNZight<sup>29</sup>.

### **1.4 *Limitations***

Despite the obvious benefits, data visualization also offers the opportunity to mislead, for instance, when displaying spurious patterns by either expanding the scale to minimize variation, or by minimizing the scale to accentuate differences (e.g.,<sup>30–32</sup>).

Furthermore, the informativeness of a graph often depends on the design capabilities of the researcher and how much thought they put into what information should be communicated. Scientists without programming experience often find themselves constrained by the options offered in standard graphics software. However, the example of Anscombe’s quartet shows that even the simplest plots can be highly informative.

### **1.5 *Guidelines***

There are no uniform guidelines as to when and which graphical representations should be used. There is, however, a fundamental principle of good statistical graphics due to Tufte<sup>33, p.92</sup>: “Above all else show the data” (i.e., minimize non-data elements). In general, scientists should aim to create a graph that is as clean, informative, and as complete as possible. These characteristics are also emphasized in the ASA Ethical Guidelines<sup>34</sup>. The guidelines mention that to ensure the integrity of data and methods, the ethical statistician “[i]n publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics” (p. 3).

Beyond that, guidelines depend on the individual aspects of the data (e.g., complexity of the data and experimental design) and context (cf.<sup>35</sup>); here we refer the interested reader to the numerous manuals describing good practices in graphical representation of statistical information (e.g.,<sup>33;36–41</sup>).

## QUANTIFYING INFERENCE UNCERTAINTY

### 1.6 *Description*

By reporting the precision with which model parameters are estimated, the analyst communicates the inevitable uncertainty that accompanies any inference from a finite sample.

### 1.7 *Benefits and Example*

Only by assessing and reporting inferential uncertainty is it possible to make any claim about the degree to which results from the sample generalize to the population. For example, Strack et al.<sup>42</sup> studied whether participants rate cartoons to be funnier when they hold a pen with their teeth (which induces a smile) instead of holding it with their lips (which induces a pout). On a 10-point Likert scale, the authors observed a raw effect size of 0.82 units. For the interpretation of this result it is essential to know the associated inferential uncertainty. In this case, the 95% confidence interval ranges from  $-0.05$  to  $1.69$ , indicating that the data are not inconsistent with a large range of effect size estimates (including effect sizes that are negligible or negative).

### 1.8 *Current Status*

In virtually all statistics courses, students are taught to provide not only the summary of statistical tests (such as  $F$  -,  $t$ -,  $p$ -values and associated degrees of freedom), but also parameter point-estimates (e.g., regression weights, effect sizes) and their associated uncertainty (e.g., standard error, posterior distribution, confidence intervals, credible intervals). Nevertheless, there exists a gap between what is taught and what is practiced. Studies of published articles in physiology<sup>21</sup>, the social sciences<sup>43</sup>, and medicine<sup>44;45</sup> revealed that error bars, standard errors, or confidence intervals were not always presented. Also, popular metrics such as Cronbach's alpha (a measure of test score reliability) are virtually never presented with a measure of inferential uncertainty.

### 1.9 *Limitations*

We agree with Jeffreys's comment in the epigraph that there are no acceptable excuses for omitting a measure of inferential uncertainty in any report.

Although not a limitation per se, it should be noted that inferential uncertainty always needs to be quantified relative to the inferential goal: does a researcher want to generalize across people, stimuli, time points, or another dimension? The proper way of computing standard errors depends on the researcher's purpose.

### **1.10 Guidelines**

Various guidelines strongly recommend that effect size estimates are accompanied by measures of uncertainty in the form of standard errors or confidence intervals. For instance,

the publication manual of the American Psychological Association (6th ed.) states: "When point estimates (e.g., sample means or regression coefficients) are provided, always include an associated measure of variability (precision), with an indication of the specific measure used (e.g., the standard error)," (p. 34). Also, the International Committee of Medical Journal Editors<sup>46</sup> explicitly recommend to "[w]hen possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals)" (p. 17).

**Box 2 - Seven Mertonian Statistical Procedures**

This box outlines how each of the seven procedures discussed in the main manuscript fulfill the Mertonian norms. An overview is given in Table 1.

	Communalism	Universalism	Disinterestedness	Organized Skepticism
1. Visualizing Data	✓		✓	✓
2. Quantifying Inferential Uncertainty	✓		✓	✓
3. Assessing Data Preprocessing Choices	✓		✓	✓
4. Reporting Multiple Models	✓		✓	✓
5. Involving Multiple Analysts		✓	✓	✓
6. Interpreting Results Modestly			✓	✓
7. Sharing Data and Code	✓	✓	✓	✓

**1. Visualizing Data**

Well-designed visualizations show at a glance the key aspects of the data. Moreover, by giving the reader a more complete picture of the data and related statistics, visualizations can either support or weaken a conclusion drawn by the researcher, or help the reader find alternative ways of interpreting the results and analyzing the data.

**2. Quantifying Inferential Uncertainty**

Acknowledging inferential uncertainty (e.g., by presenting standard errors or confidence intervals) contributes to open communication. In addition, quantifying inferential uncertainty signals that researchers are openly acknowledging the extent to which their measurements are imprecise, especially when sample size is small. Finally, explicitly acknowledging inferential uncertainty may prompt readers to question how well the results from the sample generalize to the population.

**3. Assessing Data Preprocessing Choices**

When researchers share the results from only a single data pre-processing pipeline, they may unintentionally hide important information. If a result proves sensitive to particular pre-processing choices, this warrants skepticism and may initiate a debate on the importance and plausibility of relevant data pre-processing choices (cf.<sup>47</sup>, p. 308).

#### ***4. Reporting Multiple Models***

Similar to the previous section, reporting results from only a single model may unintentionally hide important information.

#### ***5. Involving Multiple Analysts***

The multiple-analysts approach can reveal whether different (teams of) analysts reach converging or diverging conclusions from the same data set. By including other analysts with different backgrounds and interests, the potential impact of self-interest of any single analyst is counteracted. The multiple-analysts approach also stimulates skepticism by bringing to light alternative statistical perspectives on the data.

#### ***6. Interpreting Results Modestly***

Disinterested analysts arguably have little need to exaggerate claims, impress reviewers, and downplay signs of model misfit. Analysts who facilitate organized skepticism do not attempt to suppress doubt — they are not defensive, and they do not wish to protect their work against good-faith scrutiny from their peers.

#### ***7. Sharing Data and Code***

All secrecy about data is a limitation to knowledge accumulation and violates the ethos of science. All interested researchers should have access to relevant, properly anonymized data. Importantly, sharing data allows skeptical eyes to scrutinize the results, promoting quality control.

## **ASSESSING DATA PREPROCESSING CHOICES**

### ***1.11 Description***

By assessing the impact of plausible alternative data pre-processing choices (i.e., examining the “data multiverse”<sup>48</sup>), the analyst determines the extent to which the finding under scrutiny is either fragile or sturdy.

### ***1.12 Benefits and Example***

A “data multiverse” analysis reveals the fragility or sturdiness of the finding under plausible alternative data pre-processing choices. This prevents researchers from falling prey to hindsight bias and motivated reasoning, which may lead them to unwittingly report only the pre-processing pipeline that yields the most compelling result (e.g.,<sup>49;50</sup>). But even a completely unbiased analysis will benefit from a “data multiverse” analysis, as it reveals uncertainty that would otherwise remain hidden.

For example, Steegen et al.<sup>48</sup> reexamined the results of Durante et al.<sup>51</sup>, who reported an interaction between relationship status (i.e., single or not) and menstrual cycle (i.e., fertile or not) on reported religiosity. After applying a series of 180 different data pre-processing procedures (e.g., five different ways to split women into high versus low fertility), the multiverse reanalysis showed that the resulting 180 *p*-values were distributed uniformly between 0 and 1, indicating that the reported interaction is highly fragile.

### **1.13 Current Status**

The idea of assessing sensitivity to data-preprocessing choices dates back at least to De Groot<sup>49, p. 190</sup> and Leamer<sup>47, p. 308</sup> and was revived by Simmons et al.<sup>50</sup> and by Steegen et al.<sup>48</sup>. In the field of functional magnetic resonance imaging, both Carp<sup>52</sup> and Poldrack et al.<sup>53</sup> emphasized the hidden influence of different plausible pre-processing pipelines. In psychology, recent applications are Bastiaansen et al.<sup>3</sup> and Wessel et al.<sup>54</sup>. Nevertheless, the overwhelming majority of empirical articles does not report the results of a data multiverse analysis.

### **1.14 Limitations**

A pragmatic limitation of the data multiverse lies in the extra work that it entails. Another limitation can be found in ambiguities surrounding the definition of the data multiverse. The analyst has to determine what constitutes a sufficiently representative set of pre-processing choices and whether all pre-processing choices are equally plausible, such that they should be given equal weight in the multiverse analysis. A final limitation is that it is not always clear how to interpret the results of a data multiverse analysis. Interpretation can be facilitated with certain graphical formats that cluster related pipelines (e.g., specification curves;<sup>55</sup>).

### **1.15 Guidelines**

Some specific guidelines on assessing data pre-processing choices are offered by Simmons et al.<sup>50</sup>, see Requirements for Authors, numbers 5 and 6, but it is difficult to provide general guidelines as “(···) a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgment about which options can be considered reasonable and will typically depend on the experimental design, the research question, and the researchers performing the research”<sup>48, p. 709</sup>. More general



guidelines that relate exclusively to the reporting of pre-processing choices are given in the ASA Ethical Guidelines<sup>34</sup>. These mention that to ensure the integrity of data and methods, the ethical statistician “[w]hen reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms” (p. 2).

## REPORTING MULTIPLE MODELS

### 1.16 *Description*

By assessing the impact of plausible alternative statistical models (i.e., examining the “model multiverse”), the analyst gauges the extent to which a statistical conclusion is either fragile or sturdy.

### 1.17 *Benefits and Example*

Similar to the “data multiverse” analysis discussion in the previous section, a model multiverse analysis examines the fragility or sturdiness of the finding under plausible alternative statistical modeling choices. Modeling choices comprise differences in estimators and fitting regimes, but also in model specification and variable selection. Reporting the outcomes of multiple plausible models reveals uncertainty that would remain hidden if only a single model were entertained. In addition, this practice protects analysts against hindsight bias and motivated reasoning, which may unwittingly lead them to select the single model that produces the most flattering conclusion. For example, Patel et al.<sup>56</sup> quantified the variability of results under different model specifications. They considered 13 clinical, environmental, and physiological variables as potential covariates for the association of 417 self-reported, clinical, and molecular phenotypes with all-cause mortality. Consequently, they computed  $p$ -values for  $2^{13} = 8,192$  models and examined the instability of the inference, which they call the “vibration of effects”.

### 1.18 *Current Status*

Although the idea of the model multiverse dates back at least to De Groot (1956/2014) and Leamer<sup>47</sup>, most empirical researchers still base their conclusion on only a single analysis (but see<sup>57;58</sup>).

### **1.19 Limitations**

As was the case for the construction of the data multiverse, a pragmatic limitation of the model multiverse lies in the extra work that it entails—for the analyst as well as the reader. Recent work suggests that the number of plausible models can be very large (i.e.,<sup>4;7</sup>). Also, multiverses vary in their informativeness, and readers need to assess themselves whether a multiverse features notably distinct models or just runs the essentially same model multiple times. Model spaces can be overwhelming; any single analyst will naturally be drawn towards the subset of models that they are familiar with (or, unwittingly, the subset of models that yields the result that is most flattering or most in line with prior expectations). In addition, Del Giudice et al. (in press, p. 5) argue that “By inflating the size of the analysis space, the combinatorial explosion of unjustified specifications may, ironically, exaggerate the perceived exhaustiveness and authoritativeness of the multiverse while greatly reducing the informative fraction of the multiverse. At the same time, the size of the specification space can make it harder to inspect the results for potentially relevant findings. If unchecked, multiverse-style analyses can generate analytic “black holes”: Massive analyses that swallow true effects of interest but, due to their perceived exhaustiveness and sheer size, trap whatever information is present in impenetrable displays and summaries.”

### **1.20 Guidelines**

Because the construction of the model multiverse depends on the knowledge and expertise of the analyst, it is challenging to provide general guidelines. For relatively simple regression models, however, clear guidelines do exist (e.g.,<sup>56;60</sup>). Furthermore, Simonsohn et al.<sup>55</sup> suggested a specification curve analysis, and Dragicevic et al.<sup>61</sup> suggest interactive ways of presenting the results. The ASA Ethical Guidelines<sup>34</sup> mention that to meet the responsibilities towards funders and clients, the ethical statistician “[t]o the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision” (p. 3). The ASA, however, does not mention that researchers share the same responsibility towards their scientific colleagues, although this may be implicit.

One general recommendation for constructing a comprehensive model multiverse is to collaborate with statisticians who have complementary expertise, bringing us to the next section.

## INVOLVING MULTIPLE ANALYSTS

### 1.21 *Description*

By having multiple analysts independently analyze the same data set, the researcher can decrease the impact of analyst-specific choices regarding data pre-processing and statistical modeling.

### 1.22 *Benefits and Example*

The multiple-analysts approach reveals the uncertainty that is due to the subjective choices of a single analyst and promotes the application of a wider range of statistical techniques. When the conclusions of the analysts converge, this bolsters one's confidence that the finding is robust; when the conclusions diverge, this undercuts that confidence and stimulates a closer look at the statistical reasons for the lack of consensus.

The multiple-analysts approach was used, for example, in a study by Silberzahn et al.<sup>7</sup> where 29 teams of analysts examined, using the same dataset, whether the skin tone of soccer players influences their probability of getting a red card. While most of the analysis teams reported that players with a darker skin tone have a higher probability of getting a red card, some of the teams reported null results. The analysis approach used by the teams differed widely, both with respect to data pre-processing and statistical modeling (e.g., included covariates, link functions, assumption of hierarchical structure).

### 1.23 *Current Status*

A precursor to the multiple-analysts approach concerns the 1857 "Cuneiform competition", where four scholars independently translated a previously unseen ancient Assyrian inscription (Rawlinson et al., 1857). The overlap between their translations –sent to the Royal Asian Society in sealed envelopes, and simultaneously opened and inspected by a separate committee of examiners– was striking and put to rest any doubts concerning the method used to decipher such inscriptions. The multiple-analysts approach never caught on in practice, although recent examples exist in psychology and neuroscience<sup>3–5;7;62;63</sup>

### **1.24 *Limitations***

As was the case for the construction of the data multiverse and the model multiverse, a pragmatic limitation of the multiple analyst approach lies in the extra work that it entails, specifically with respect to (1) finding knowledgeable analysts who are interested in participating; (2) documenting the data set, describing the research question, and identifying the target of statistical inference; (3) collating the initial responses from each team, and potentially coordinating a review and feedback round. While differences in opinion should be respected, there need to be ways to filter out analysis approaches that involve clear mistakes. An additional limitation concerns possible homogeneity of the analysts. For instance, all analysts involved could be rigidly educated in the same school of thought, share cultural or social biases, or just make the same mistake. In such a case, the results may create an inflated sense of certainty in the conclusion that was reached. This potential limitations can be mitigated by selecting a diverse group of analysts and incorporating feedback and revision options in the process<sup>7</sup>, a round-table discussion<sup>5</sup> or, more systematically, a Delphi approach<sup>64</sup>.

### **1.25 *Guidelines***

There are no explicit guidelines concerning the multiple-analysts approach. We propose that the optimal number of analysts to include depends on factors such as the complexity of the data, the importance of the research question (e.g., a clinical trial on the effectiveness of a new drug against COVID-19 warrants a relatively large number of analysts), and the probability that the analysts could reasonably reach a different conclusion (e.g., there may be multiple ways to interpret the research question, and there may be multiple dependent variables and predictor variables that could or could not be relevant).

When analysts are selected, care should be taken to ensure heterogeneity, diversity, and balance. Specifically, one should be mindful of the potential biasing effects of specific background knowledge, culture, education, and career stage of the analyst.

The ASA Guidelines emphasize the legitimacy and value in alternative analytic approaches, stating that “[t]he practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers (···) can arrive at different and potentially diverging judgments about the plausibility of different explanations” (p. 5).

## INTERPRETING RESULTS MODESTLY

### 1.26 *Description*

By modestly interpreting the results, the analyst explicitly acknowledges any remaining doubts concerning the importance, replicability, and generalizability of the scientific claims at hand.

### 1.27 *Benefits and Example*

Modestly presented scientific claims enable the reader to evaluate the outcomes for what they usually are: not final, but tentative results pointing in a certain direction, with considerable uncertainty surrounding their generalizability and scope. Overselling results might lead to the misallocation of public resources towards approaches that are in fact not properly validated and not ready for application in practice. Also, researchers themselves risk losing long-term credibility for short-term gains of greater attention and higher citation counts. Moreover, after having publicly committed to a bold claim, it becomes difficult to admit that one's initial assessment was wrong; in other words, overconfidence is not conducive to scientific learning.

Scientists of true modesty remain doubtful even at moments of great success. For example, when James Chadwick found experimental proof of neutrons, the discovery that earned him the Nobel prize, he communicated it modestly under the title "Possible Existence of Neutron"<sup>65</sup>.

### 1.28 *Current Status*

Tukey<sup>66</sup> already remarked that "Laying aside unethical practices, one of the most dangerous [(···) practices of data analysis (···)] is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an imprimatur." (p. 13). Almost 60 years later, an editorial in *Nature Human Behaviour* warns its readers about "conclusive narratives that leave no room for ambiguity or for conflicting or inconclusive results"<sup>67</sup>, p. 1. Similarly, Simons et al.<sup>68</sup> suggested adding a mandatory Constraints on Generality statement in the discussion section of all primary research articles in the field of psychology to prevent authors from making wildly exaggerated claims of generality. This suggests that scientific modesty is rarer than we would expect if Mertonian norms were widely adopted. There are some clear indications of a lack of modesty.

First of all, the frequency of stronger language (words like “amazing”, “ground-breaking”, “unprecedented”) seemed to have increased in the last few decades<sup>69</sup>. Secondly, dichotomization of findings (i.e., ignoring the uncertainty inherent to statistical inference) is common practice (e.g.,<sup>43</sup>; also see paragraph 4.3). Thirdly, textbooks (which are typically a reflection of current practice) on how to write papers often explicitly encourage authors to overclaim (e.g.,<sup>70;71</sup>)

### **1.29 Limitations**

Publications and grants are important for scientific survival. Coupled with the fact that journals and funders often prefer groundbreaking and unequivocal outcomes, it may be detrimental to one’s success to modestly interpret the results. The encouragement of this Mertonian practice may require change at an institutional level, although some have argued that scientists should not hide behind the system when defending their behavior<sup>72</sup>.

### **1.30 Guidelines**

There are several ways we can contribute to increasing intellectual modesty. First of all, we could encourage intellectual modesty in others’ work when we act as reviewers of papers and grant proposals<sup>73</sup>. Since a reviewer’s career is independent of how they evaluate a paper, they can make a positive review conditional on a more modest presentation of outcomes. Hoekstra and Vazire<sup>73</sup> present a list of suggestions for increasing modesty in the traditional sections of an empirical article, which can be used by authors as well. One example (p. 16) includes “Titles should not state or imply stronger claims than are justified (e.g., causal claims without strong evidence)”.

Also, the ASA Guidelines state: “[t]he ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may affect the integrity or reliability of the statistical analysis” (p. 2).

## **SHARING DATA AND CODE**

### **1.31 Description**

By sharing data and analysis code, researchers provide the basis for their scientific claims. Ideally, data and code should be shared publicly, freely, and in a manner that facilitates reuse.

### **1.32 *Benefits and Example***

Since there are many different ways of processing and analyzing data<sup>7,48</sup>, sharing code promotes reproducibility and encourages sensitivity analyses. Sharing data and code also allows other researchers to establish the validity of the original analyses, it can facilitate collaboration, but it can also serve as protection against data loss. When publishing his theory on “general intelligence”, Spearman<sup>74</sup> shared his data as an appendix to the article. A century later, this act of foresight enabled scientists to use this data set for both research and education. Because Spearman made his data publicly available, other researchers could establish the reproducibility and generalizability of the findings.

### **1.33 *Current Status***

Data sharing has never been easier. Public repositories offer free storage space for research materials, data (e.g., the Open Science Framework), and code (e.g., Github). While data sharing is not yet a general practice in most scientific fields, several recent initiatives (e.g., Open Data/Code/Materials badges,<sup>75</sup> standards (TOP Guidelines,<sup>76</sup>), journals (e.g., Scientific Data) and checklists (e.g., Transparency Checklist,<sup>77</sup>) are helping to promote this research practice. When sharing raw data is unfeasible, researchers can make aggregated data summaries available, for example, the data used to generate certain plots or covariance matrices of involved variables.

### **1.34 *Limitations***

Restrictions imposed by funders, ethics review boards in universities and other institutions, collaborators, and legal contracts may limit the extent to which data can be publicly shared. There may also be practical considerations (e.g., sharing big data), data use agreements, privacy rights, and institutional policies that can curtail sharing intentions. What remains central is to inform the readers about the accessibility of the data of the analysis. It should be noted that these limitations should not apply to the analysis code as long as code is solely reflective of the researcher’s analysis actions and is free of any data privacy issues.

### **1.35 *Guidelines***

An important principle of sharing data is that they should be Findable, Accessible, Interoperable, and Reusable (FAIR,<sup>78</sup>). Several guides are available discussing the

practical (e.g.,<sup>79</sup>) and ethical (e.g.,<sup>80</sup>) aspects of data sharing. Researchers should follow the data sharing procedures and requirements of their fields (e.g.,<sup>81;82</sup>) and indicate the accessibility of the data in the research report<sup>76;83</sup>. The ASA Ethical Guidelines<sup>34</sup> for Statistical Practice state that the ethical statistician “[p]romotes sharing of data and methods as much as possible”, and “[m]akes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators.” (p. 5).

## CONCLUDING COMMENTS

If the statistical literature is any guide, one may conclude that statisticians rarely agree with one another. For instance, the 2019 special issue in *The American Statistician* featured 43 articles on  $p$ -values, and in their editorial Wasserstein et al.<sup>13</sup> stated that “the voices in the 43 papers in this issue do not sing as one”. However, despite the continuing disagreements about the foundations of statistical inference, we believe there is nevertheless much common ground among statisticians, specifically with respect to the ethical aspects of their profession. To explore this ethical dimension more systematically, we started by considering the Mertonian norms that characterize the ethos of science and outlined a non-exhaustive list of seven concrete, teachable, and implementable practices that we believe need wider propagation.

In essence, these practices are about promoting transparency and the open acknowledgement of uncertainty. With agreement on such practices explicitly acknowledged, we believe that commonly discussed contentious issues (e.g.,  $p$ -values) may become less crucial. Indeed, in a letter to his frequentist nemesis Sir Ronald Fisher, the arch-Bayesian Sir Harold Jeffreys wrote “Your letter confirms my previous impression that it would only be once in a blue moon that we would disagree about the inference to be drawn in any particular case, and that in the exceptional cases we would both be a bit doubtful”<sup>84</sup>, p. 162. We hope that the proposed statistical practices will improve the quality of data analysis across the board, especially in applied disciplines that are perhaps unfamiliar with the ethical aspects of statistics, aspects that a statistician may take for granted. Also, instead of counting on them to be absorbed through osmosis, we believe it is important to include these ethical considerations –and their statistical consequences– explicitly in the statistics curricula. Statistical techniques other than those discussed here may also further the Mertonian ideals. We hope that this contribution provides the impetus for a deeper exploration of how data analysis in applied fields can



become more transparent, more informative, and more open about the uncertainties that inevitably arise in any statistical data analysis problem.

## **AUTHOR CONTRIBUTIONS**

Conceptualization: E.-J.W., A. Sarafoglou, and B.A.

Project Administration: B.A.

Writing - Original Draft Preparation: E.-J.W., A. Sarafoglou, C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, J.T., and B.A.

Writing - Review & Editing: E.-J.W., A. Sarafoglou, S.A., C.A., J.A., Š.B., N.v.D., R.H., D.M., D.v.R., A. Sluga, F.S., J.T., and B.A.

## **ACKNOWLEDGEMENTS**

We are grateful to Nicole Lazar for her comments on a draft version. We also thank everyone who was involved in drafting the initial list of statistical procedures during the hackathon that took place at the 2019 meeting of the Society for the Improvement of Psychological Science in Rotterdam, The Netherlands. This work was supported in part by a European Research Council (ERC) grant to E.-J.W. (283876), a Netherlands Organisation for Scientific Research (NWO) grant to A. Sarafoglou (406-17-568), as well as a Dutch scientific organization Vidi grant from the NWO to D.v.R. (016.Vidi.188.001).

## References

- [1] I. J. Good. 46656 varieties of Bayesians. *The American Statistician*, 25:62–63, 1971.
- [2] B. Aczel, R. Hoekstra, A. Gelman, E.–J. Wagenmakers, I. G. Klugkist, J. N. Rouder, J. Vandekerckhove, M. D. Lee, R. D. Morey, W. Vanpaemel, Z. Dienes, and D. van Ravenzwaaij. Discussion points for Bayesian inference. *Nature Human Behaviour*, 4: 561–566, 2020. URL <https://doi.org/10.1038/s41562-019-0807-z>.
- [3] Jojanneke A Bastiaansen, Yoram K Kunkels, Frank J Blaauw, Steven M Boker, Eva Ceulemans, Meng Chen, Sy-Miin Chow, Peter de Jonge, Ando C Emerencia, Sacha Epskamp, et al. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, 137:110211, 2020.
- [4] R. Botvinik–Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannes-son, M. Kirchler, R. Iwanir, J. A. Mumford, A. Adcock, P. Avesani, B. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R. Benoit, R. Berkers, J. Bhanji, B. Biswal, S. Bobadilla–Suarez, T. Bortolini, K. Bot-tenhorn, A. Bowering, S. Braem, H. Brooks, E. Brudner, C. Calderon, J. Camilleri, J. Castellon, L. Cecchetti, E. Cieslik, Z. Cole, O. Collignon, R. Cox, W. Cun-ningham, S. Czoschke, K. Dadi, C. Davis, A. De Luca, M. Delgado, L. Demetriou, J. Dennison, X. Di, E. Dickie, E. Dobryakova, C. Donnat, J. Dukart, N. W. Duncan, J. Durnez, A. Eed, S. Eickhoff, A. Erhart, L. Fontanesi, G. M. Fricke, A. Galvan, R. Gau, S. Genon, T. Glatard, E. Glerean, J. Goeman, S. Golowin, C. González–García, K. Gorgolewski, C. Grady, M. Green, J. Guassi Moreira, O. Guest, S. Hakimi, J. P. Hamilton, R. Hancock, G. Handjaras, B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S. Huettel, M. Hughes, V. Iacovella, A. Iordan, P. Isager, A. I. Isik, A. Jahn, M. Johnson, T. John-stone, M. Joseph, A. Juliano, J. Kable, M. Kassinopoulos, C. Koba, X.–Z. Kong, T. Koscik, N. E. Kucukboyaci, B. Kuhl, S. Kupek, A. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M. Li, P. C. Lim, E. Lintz, S. Liphardt, A. Losecaat Vermeer, B. Love, M. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J. McGuire, H. Melero, A. Méndez Leal, B. Meyer, K. Meyer, P. Mihai, G. Mitsis, J. Moll, D. Nielson, G. Nilsonne, M. Not-ter, E. Olivetti, A. Onicas, P. Papale, K. Patil, J. E. Peelle, A. Pérez, D. Pishedda, J.–B. Poline, Y. Prystauka, S. Ray, P. Reuter–Lorenz, R. Reynolds, E. Ricciardi, J. Rieck, A. Rodriguez–Thompson, A. Romyn, T. Salo, G. Samanez–Larkin, E. Sanz–Morales, M. Schlichting, D. Schultz, Q. Shen, M. Sheridan, F. Shiguang, J. Silvers, K. Skagerlund, A. Smith, D. Smith, P. Sokol–Hessner, S. Steinkamp, S. Tashjian, B. Thirion, J. Thorp, G. Tinghög, L. Tisdall, S. Tompson, C. Toro–Serey, J. Torre, L. Tozzi, V. Truong, L. Turella, A. E. van’t Veer, T. Verguts, J. Vettel, S. Vijayara-jah, K. Vo, M. Wall, W. D. Weeda, S. Weis, D. White, D. Wisniewski, A. Xifra–Porxas, E. Yearling, S. Yoon, R. Yuan, K. Yuen, L. Zhang, X. Zhang, J. Zosky, T. E. Nichols, R. A. Poldrack, and T. Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88, 2020. URL <https://doi.org/10.1038/s41586-020-2314-9>.
- [5] N. van Dongen, J. B. van Doorn, Q. F. Gronau, D. van Ravenzwaaij, R. Hoekstra, M. N. Haucke, D. Lakens, C. Hennig, R. D. Morey, S. Homer, A. Gelman, J. Sprenger, and E.-J. Wagenmakers. Multiple perspectives on inference for two

- simple statistical scenarios. *The American Statistician*, 73:328–339, 2019.
- [6] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filip-pova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117:8398–8403, 2020. doi: 10.1073/pnas.1915006117.
- [7] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, Högden F., K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, Spörlein C., T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1:337–356, 2018.
- [8] V. Amrhein, S. Greenland, and B. B. McShane. Retire statistical significance. *Nature*, 567:305–307, 2019.
- [9] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T.-H. Ho, H. Hoijtink, J. H. Jones, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, M. Kirchler, D.

- Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. Moore, S. L. Morgan, M. Munafò, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2:6–10, 2018.
- [10] L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors. *What if There Were No Significance Tests?* Lawrence Erlbaum, Mahwah (NJ), 1997.
- [11] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett. Abandon statistical significance. *The American Statistician*, 73:235–245, 2019.
- [12] R. L. Wasserstein and N. A. Lazar. The ASA’s statement on  $p$ -values: Context, process, and purpose. *The American Statistician*, 70:129–133, 2016.
- [13] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician*, 73:1–19, 2019.
- [14] R. K. Merton. The normative structure of science (1942). In R. K. Merton, editor, *The Sociology of Science: Theoretical and Empirical Investigations*, pages 267–278. University of Chicago Press, Chicago, IL, 1973.
- [15] Melissa S Anderson, Brian C Martinson, and Raymond De Vries. Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics*, 2:3–14, 2007.
- [16] J. W. Tukey. *Explanatory Data Analysis*. Addison–Wesley, Reading, MA, 1977.
- [17] Andrew Gelman. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13:755–779, 2004. doi: 10.1198/106186004X11435.
- [18] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182:389–402, 2019.
- [19] Andrew Heathcote, Scott D Brown, and Eric-Jan Wagenmakers. An introduction to good practices in cognitive modeling. In *An Introduction to Model-Based Cognitive Neuroscience*, pages 25–48. Springer Verlag, 2015.
- [20] J. Kerman, A. Gelman, T. Zheng, and Y. Ding. Visualization in Bayesian data analysis. In C. Chen, W. Härdle, and A. Unwin, editors, *Handbook of Data Visualization*, pages 709–724. Springer Verlag, Berlin, 2008.
- [21] T. L. Weissgerber, N. M. Milic, S. J. Winham, and V. D. Garovic. Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, 13:e1002128, 2015.
- [22] Kieran Healy and James Moody. Data visualization in sociology. *Annual Review of Sociology*, 40:105–128, 2014.
- [23] Edmund William Gilbert. Pioneer maps of health and disease in England. *The Geographical Journal*, 124:172–183, 1958.
- [24] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27:17–21, 1973.
- [25] J. Matejka and G Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294, 2017.
- [26] W. Playfair. *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century*. Corry, London, 1786. Re-published in Wainer, H. and Spence, I. (eds.), *The Commercial and Political Atlas and*

- Statistical Breviary, 2005, Cambridge University Press.
- [27] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. John Wiley & Sons, Chichester, 2011.
- [28] W Chang, J Cheng, JJ Allaire, Y Xie, and J McPherson. shiny: Web application framework for R [Computer software]. <http://CRAN.R-project.org/package=shiny>, 2020.
- [29] iNZight Team. iNZight (Version 4.0.2.) [Computer software]. <https://inzight.nz>, 2020.
- [30] Alberto Cairo. *How Charts Lie: Getting Smarter about Visual Information*. WW Norton & Company, New York, 2019.
- [31] Andrew Gelman. Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, 20:3–7, 2011.
- [32] Howard Wainer. How to display data badly. *The American Statistician*, 38:137–147, 1984.
- [33] Edward R Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1973.
- [34] Committee on Professional Ethics of the American Statistical Association. Ethical guidelines for statistical practice, 2018. URL <https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice>.
- [35] L. Diamond and F. J. Lerch. Fading frames: Data presentation and framing effects. *Decision Sciences*, 23:1050–1071, 1992.
- [36] C. Chen, W. Härdle, and A. Unwin, editors. *Handbook of Data Visualization*. Springer Verlag, Berlin, 2008.
- [37] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.
- [38] Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. Let’s practice what we preach: Turning tables into graphs. *The American Statistician*, 56:121–130, 2002.
- [39] Riccardo Mazza. *Introduction to information visualization*. Springer Science & Business Media, London, 2009.
- [40] Claus O Wilke. *Fundamentals of data visualization: A primer on making informative and compelling figures*. O’Reilly Media, Sebastopol, CA, 2019.
- [41] Leland Wilkinson. *The Grammar of Graphics*. Springer Science & Business Media, New York, 1999.
- [42] F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768–777, 1988.
- [43] Rink Hoekstra, Sue Finch, Henk AL Kiers, and Addie Johnson. Probability as certainty: Dichotomous thinking and the misuse of  $p$ -values. *Psychonomic Bulletin & Review*, 13:1033–1037, 2006.
- [44] Richelle J Cooper, David L Schriger, and Reb JH Close. Graphical literacy: The quality of graphs in a large-circulation journal. *Annals of Emergency Medicine*, 40: 317–322, 2002.
- [45] David L Schriger, Reshmi Sinha, Sara Schroter, Pamela Y Liu, and Douglas G Altman. From submission to publication: A retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British Medical Journal. *Annals of Emergency Medicine*, 48:750–756, 2006.
- [46] International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. <http://www.icmje.org/icmje-recommendations.pdf>, 2019.

- [47]Edward E Leamer. Sensitivity analyses would help. *The American Economic Review*, 75:308–313, 1985.
- [48]S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11:702–712, 2016.
- [49]A. D. De Groot. The meaning of “significance” for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, 148:188–194, 1956/2014.
- [50]J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22:1359–1366, 2011.
- [51]Kristina M Durante, Ashley Rae, and Vladas Griskevicius. The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24:1007–1016, 2013.
- [52]J. Carp. On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6:1–13, 2012.
- [53]R. A. Poldrack, C. I. Baker, J. Durnez, K. J. Gorgolewski, P. M. Matthews, M. R. Munafò, T. E. Nichols, J.-B. Poline, E. Vul, and T. Yarkoni. Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18:115–126, 2017.
- [54]Ineke Wessel, Casper Albers, Anna Roos Eva Zandstra, and Vera Ellen Heininga. A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think task. *Memory*, 28:870–887, 2020.
- [55]U. Simonsohn, L. D. Nelson, and J. P. Simmons. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020.
- [56]Chirag J Patel, Belinda Burford, and John PA Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68:1046–1058, 2015.
- [57]Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.
- [58]Ross Levine and David Renelt. A sensitivity analysis of cross-country growth regressions. *The American Economic Review*, pages 942–963, 1992.
- [59]Marco Del Giudice, Steven W Gangestad, and W Steven. A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, in press.
- [60]Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, pages 382–401, 1999.
- [61]P. Dragicevic, Y. Jansen, A. Sarma, M. Kay, and F. Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2019.
- [62]U. Boehm, G. E. Hawkins, S. D. Brown, H. van Rijn, and E.-J. Wagenmakers. Of monkeys and men: Impatience in perceptual decision-making. *Psychonomic Bulletin & Review*, 23:738–749, 2016.
- [63]Gilles Dutilh, Jeffrey Annis, Scott D Brown, Peter Cassey, Nathan J Evans, Raoul PPP Grasman, Guy E Hawkins, Andrew Heathcote, William R Holmes, Angelos-Miltiadis Kryptos, et al. The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, 26: 1051–1069, 2019.

- [64]Shakila Thangaratinam and Charles WE Redman. The delphi technique. *The Obstetrician & Gynaecologist*, 7:120–125, 2005.
- [65]James Chadwick. Possible existence of a neutron. *Nature*, 129:312, 1932.
- [66]J. W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33: 1–67, 1962.
- [67]NHB Editorial. Tell it like it is. *Nature Human Behaviour*, 4:1, 2020.
- [68]D. J. Simons, Y. Shoda, and D. S. Lindsay. Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12:1123–1128, 2017.
- [69]Christiaan H Vinkers, Joeri K Tjeldink, and Willem M Otte. Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, 351:h6467, 2015.
- [70]Daryl J Bem. Writing the empirical journal. In M. R. Zanna and J. M. Darley, editors, *The compleat academic: A practical guide for the beginning social scientist*, pages 171–201. Lawrence Erlbaum Associates, Mahwah, NJ, 1987.
- [71]J. van Doorn, D. van den Bergh, F. Dablander, N. van Dongen, K. Derks, N. J. Evans, Q. F. Gronau, J. M. Haaf, Y. Kunisato, A. Ly, M. Marsman, A. Sarafoglou, A. Stefan, and E.-J. Wagenmakers. Strong public claims may not reflect researchers’ private convictions. *Significance*, 2021. URL <https://psyarxiv.com/pc4ad>.
- [72]T. Yarkoni No, it’s not the incentives – it’s you. <https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/>, 2018.
- [73]R. Hoekstra and S. Vazire. Intellectual humility is central to science: Some practices to aspire to. *PsyArXiv*, 2020.
- [74]C Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.
- [75]M. C. Kidwell, L. B. Lazarević, E. Baranski, T. E. Hardwicke, S. Piechowski, L.-S. Falkenberg, C. Kennett, A. Slowik, C. Sonnleitner, C. Hess–Holden, T. M. Errington, S. Fiedler, and B. A. Nosek. Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLOS Biology*, 14:e1002456, 2016.
- [76]B.A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. A. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.-J. Wagenmakers, R. Wilson, and T. Yarkoni. Promoting an open research culture. *Science*, 348:1422–1425, 2015.
- [77]B. Aczel, B. Szaszi, A. Sarafoglou, Z. Kekecs, Š. Kucharský, D. Benjamin, C. D. Chambers, A. Fisher, A. Gelman, M. A. Gernsbacher, J. P. Ioannidis, E. Johnson, K. Jonas, S. Kousta, S. O. Lilienfeld, D. S. Lindsay, C. C. Morey, M. Munafò, B. R. Newell, H. Pashler, D. R. Shanks, D. J. Simons, J. M. Wicherts, D. Albarracin, N. D. Anderson, J. Antonakis, H. Arkes, M. D. Back, G. C. Banks, C. Beevers, A. A. Bennett, W. Bleidorn, T. W. Boyer, C. Cacciari, A. S. Carter, J. Cesario, C. Clifton, R. M. Conroy, M. Cortese, F. Cosci, N. Cowan, J. Crawford, E. A. Crone, J. Curtin, R. Engle, S. Farrell, P. Fearon, M. Fichman, W. Frankenhuys, A. M. Freund, M. G. Gaskell, R. Giner-Sorolla, D. P. Green, R. L. Greene, L. L. Harlow, F. Hoces de la Guardia, D. Isaacowitz, J. Kolodner, D.

- Lieberman, G. D. Logan, W. B. Mendes, L. Moersdorf, B. Nyhan, J. Pollack, C. Sullivan, S. Vazire, and E.-J. Wagenmakers. A consensus-based transparency checklist. *Nature Human Behaviour*, 4:4–6, 2020.
- [78] M. D. Wilkinson, M. Dumontier, IJ. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzales–Beltran, A. J. G. Gray, P. Groth, C. Goble, Grethe J. S., J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca–Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.
- [79] Olivier Klein, Tom E Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsson, Wolf Vanpaemel, and Michael C Frank. A practical guide for transparency in psychological science. *Collabra: Psychology*, 4:1–15, 2018.
- [80] G. Alter and R. Gonzalez. Responsible practices for data sharing. *American Psychologist*, 73:146–156, 2018.
- [81] E.-J. Wagenmakers, S. Kucharsky, and the JASP Team, editors. *The JASP Data Library*. JASP Publishing, Amsterdam, 2020.
- [82] Darren B Taichman, Peush Sahni, Anja Pinborg, Larry Peiperl, Christine Laine, Astrid James, Sung-Tae Hong, Abraham Haileamlak, Laragh Gollogly, Fiona Godlee, Frank A Frizelle, and Fernando Flor. Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *JAMA*, 317:2491–2492, 2017.
- [83] IJsbrand Jan Aalbersberg, Tom Appleyard, Sarah Brookhart, Todd Carpenter, Michael Clarke, Stephen Curry, Josh Dahl, Alexander DeHaven, Eric Eich, Maryrose Franko, Leonard Freedman, Chris Graf, Sean Grant, Brooks Hanson, Heather Joseph, Veronique Kiermer, Bianca Kramer, Alan Kraut, Roshan Kumar Karn, Carole Lee, Aki MacFarlane, Maryann Martone, Evan Mayo-Wilson, Marcia McNutt, Meredith McPhail, David Mellor, David Moher, Alison Mudditt, Brian Nosek, Belinda Orland, Timothy Parker, Mark Parsons, Mark Patterson, Solange Santos, Carolyn Shore, Daniel Simons, Bobbie Spellman, Jeffrey Spies, Matthew Spitzer, Victoria Stodden, Sowmya Swaminathan, Deborah Sweet, Anne Tsui, and Simine Vazire. Making science transparent by default; Introducing the TOP statement. OSF Preprints, 2018. URL <https://osf.io/sm78t>.
- [84] J. H. Bennett, editor. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford, 1990.



3.3.1. *Situational factors shape moral judgments in the trolley dilemma in Eastern, Southern, and Western countries in a culturally diverse sample*<sup>41</sup>

Bence Bago<sup>1</sup>, [...] <sup>42</sup> Balazs Aczel<sup>2</sup>

<sup>1</sup>IAST, Toulouse School of Economics,

<sup>2</sup>Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

**Abstract**

The study of moral judgements often centers on moral dilemmas in which options consistent with deontological perspectives (i.e., emphasizing rules, individual rights, and duties) are in conflict with options consistent with utilitarian judgements (i.e., following the greater good based on consequences). Greene et al. (2009) showed that psychological and situational factors (e.g., the intent of the agent or the presence of physical contact between the agent and the victim) can play an important role in moral dilemma judgements (e.g., trolley problem). Our knowledge is limited concerning both the universality of these effects outside the United States and the impact of culture on the situational and psychological factors of moral judgements. Thus, we empirically tested the universality of the effects of intent and personal force on moral dilemma judgements by replicating the experiments of Greene et al. in 45 countries from all inhabited continents. We found that personal force and its interaction with intention, exert influence on moral judgements in the US and Western cultural clusters, replicating and expanding the original findings. Moreover, the personal force effect was present in all cultural clusters, suggesting it is culturally universal. The evidence for the cultural universality of the interaction effect was inconclusive in the Eastern and Southern cultural clusters (depending on exclusion criteria). We found no strong association between collectivism/individualism and moral dilemma judgements.

---

<sup>41</sup> Published as: Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., ... & Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, 6, 880-895.

<sup>42</sup>All contributing authors are listed in the original publication.

## Introduction

Moral dilemmas can be portrayed as decisions between two main conflicting moral principles: utilitarian and deontological. Utilitarian (also referred to as consequentialist) philosophies<sup>1</sup> hold that an action is morally acceptable if it maximizes well-being for the greatest number of people (in terms of saved lives, for example). On the other hand, deontological philosophy<sup>2</sup> evaluates the morality of the action based on the intrinsic nature of the action (i.e., the deontological option often reflects greater concern for the individual rights and duties<sup>3</sup>). The dilemma between these two principles plays a prominent role in law and policy-making decisions, ranging from decisions of health budget allocations<sup>4</sup> to the dilemma of self-driving vehicles<sup>5</sup>. This inherent conflict is well illustrated by the so-called trolley problem, which has long interested both philosophers and psychologists. One version of the dilemma is presented as follows<sup>6</sup>:

You are a railway controller. There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are 5 workmen. The trolley is headed straight for them and they will be killed if nothing is done. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a side track and you can save the 5 workmen on the main track. You notice that there are 2 workmen on the side track. So there will be 2 workmen who will be killed if you pull the lever and change the tracks but the 5 workmen on the main track will be saved. Is it morally acceptable for you to pull the lever?

A deontological decision-maker would argue that pulling the lever is morally unacceptable, as it would be murder (Note that deontological principles are often more complicated than this. Some of the deontological rules would allow for killing in this situation. The terms “deontological” and “utilitarian/consequentialist” are labels we use to refer to certain responses). On the other hand, utilitarianism would suggest that it is morally acceptable to pull the lever, as it would maximize the number of saved lives.

In an alternative version of the dilemma, one has to push a man off a footbridge in front of the trolley (“footbridge” scenario). This man will die but will stop the trolley, and the five people in the way of the trolley will be saved. Interestingly, people are less likely to make a decision

consistent with utilitarian perspectives in the footbridge scenario compared to the standard switch scenario (We call these “utilitarian” responses but the fact that these decisions are consistent with utilitarianism does not indicate that people gave them out of utilitarian principles; the same is true for “deontological” responses<sup>7,8</sup>). The difference between the utilitarian response rate in those scenarios became the basis of investigations of many influential cognitive theories in the field of moral judgement<sup>3,7-13</sup>. The fact that people respond differently to the two trolley dilemmas was proposed to be explained by people’s adherence to the so-called doctrine of double effect<sup>6,9</sup>. A simple version of this doctrine is that harm is permissible as an unintentional side-effect of a good result. This doctrine is the basis of many policies in several countries all around the world concerning issues such as abortion<sup>6</sup>, euthanasia<sup>14</sup>, international armed conflict regulations<sup>15,16</sup>, and even international business ethics<sup>17</sup>. According to this doctrine, it is morally impermissible to bomb civilians to win a war, even if ending the war would eventually save more lives. However, if civilians die in a bombing of a nearby weapons factory as a side-effect, the bombing is morally acceptable. The way people perceive or act on these moral rules can influence the policies that are accepted or even followed - as we can already see in the case of driverless cars, which sometimes have to decide between sacrificing their own passengers and saving one or more pedestrians<sup>5</sup>.

Greene et al.<sup>18</sup> and Cushman et al.<sup>9</sup>, however, argued that the difference in utilitarian response rates cannot simply be explained by the doctrine of double effect. Greene et al. presented evidence for the interaction of the intention of harm (i.e., harm as means or side effect; referring to the doctrine of double effect) and personal force (i.e., whether or not the agent had to use personal effort to kill the victim and save more people) on moral acceptability ratings. More concretely, people were less likely to judge sacrificing one person to save more people as morally acceptable when they had to use their personal force to kill the person *and* the death of this person was required to save more people (this is what is meant by *intending* the harm). Hence, they concluded that people are more sensitive to the doctrine of double effect when they have to use their own physical force. Despite some exceptions<sup>26,27</sup>, most of the evidence for this conclusion comes from samples of WEIRD (Western, Educated, Industrialized, Rich, Democratic<sup>23,24</sup>) societies, leaving the question open of whether these effects are psychologically universal<sup>25</sup> or culture-specific.

This study tests three cross-cultural hypotheses:

- (1) The effects of personal force on moral judgements are culturally universal.
- (2) The interactional effect of personal force and intention on moral judgements is culturally universal.
- (3) Collectivism-individualism has a moderating effect on the degree to which personal force and intention affect moral judgements in a way that their effect is stronger in more collectivistic cultures.

The first and second hypotheses, that the effects of personal force and intention on moral judgements are culturally universal, come from their relatedness to interpersonal violence. People seem to exhibit a general tendency to avoid causing violent harms (e.g., murder)<sup>19,20</sup>, and they are more likely to perceive actions as violent or harmful when they are supposed to use personal force or intention<sup>3</sup>. As a result, people are more likely to behave in a deontological way when personal force or intention is present in the dilemma. As all cultures regulate interpersonal violence,<sup>21</sup> we expected to find that both intention and personal force, as well as their interaction, have an effect on moral judgements across cultures. The literature seems to be in accordance with these hypotheses. For example, Chinese<sup>25-27</sup> and Russian<sup>28</sup> participants responded similarly to moral dilemmas as Americans and Western Europeans, and even small-scale societies tended to be susceptible to the effect of intention<sup>22,23</sup>.

Even though we anticipated that the effect of personal force and intention would emerge universally across cultures, we nonetheless expected cultural differences to moderate these effects. The effect of personal force on moral judgement has been attributed to emotional processes<sup>9,24-26</sup>, specifically social emotions (such as guilt, shame or regret)<sup>25,27</sup>. The potential use of personal force makes people feel guilt or shame before making a decision and, therefore, rating actions that use personal force as morally less acceptable. There is a convincing argument that these social emotions are universal<sup>28-30</sup>, despite some cultural variation in their intensity and the social contexts in which they are experienced<sup>28-30</sup>. It has been argued that shame and guilt are more important in interdependent, collectivistic cultures (as their function is argued to be linked to social control). People living in East Asian countries have reported experiencing these emotions more frequently and more intensely<sup>28-30</sup>. Other findings suggest that it is anxiety that mediates the effect of intention and personal force<sup>26</sup>, but anxiety (social anxiety in particular) has also been positively associated with collectivism<sup>31</sup>, pointing to the same direction. Hence, we hypothesized that people living in collectivistic cultures would judge

actions that involve personal force and intention as morally less acceptable than people in individualistic cultures. Utilitarian responding in moral dilemma judgements has also been associated with low levels of empathic concern<sup>32</sup> and people living in collectivistic cultures have been suggested to exhibit higher levels of empathic concern<sup>33,34</sup>. Hence, we predicted that individualism-collectivism would also have an effect on utilitarian responding: collectivists would be less utilitarian in general, due to their higher levels of empathic concern.

In addition to testing our confirmatory hypotheses, we also collected a number of additional country-level as well as individual measures for exploratory purposes. These measures have been previously shown to be related to moral judgement such as economic status<sup>35</sup>, individual level individualism-collectivism<sup>35</sup>, and religiosity<sup>36</sup>. We also administered an alternative measure of utilitarian responding<sup>37-40</sup>.

The present investigation is crucial for advancing the field for the following reasons:

- 1) The original article has been very influential (515 citations so far), but replicability has not established yet.
- 2) Our knowledge is scarce on the cultural universality of the effect of personal force and intention in moral judgements.
- 3) The resulting database (with many types of trolley problems and additional measures) could assist and guide future research and applications on moral thinking.

## **Overview**

In the first part of our study, we tested the universality of the role of personal force in moral judgements with a direct replication of Study 1a conducted by Greene et al.. In their study, the authors found evidence that the application of personal force decreases moral acceptability of the utilitarian action (Hypothesis 1a, 1b). In the second part, we tested the universality of the interactional effect of personal force and intention on moral dilemma judgements, by replicating Study 2 of Greene et al. (Hypothesis 2a, 2b) with partially different moral dilemmas. Furthermore, we tested our hypothesis that collectivism moderates the effect of intention and personal force (Hypothesis 3). In addition, we collected various additional measures for exploratory purposes.

## Results

We collected data from 27,502 participants out of 45 countries. Due to our exclusion criteria, we had to exclude 80.6% of the sample for the main analysis (see Table 1 for the various exclusion criteria). Note that, as we registered, we conducted the analysis without excluding the data of the participants who were familiar with the trolley problem (36.2% exclusions), and we also conducted a post-hoc explorative analysis in which we applied no exclusion criteria. All participants were presented with two moral dilemmas that were equivalent in structure but were different in wording: trolley dilemmas and speedboat dilemmas (the former described a situation involving trolley and people on the tracks, the latter described a situation with people on a speedboat and others drowning in the sea). In Study 1, we tested the effect of personal force on moral dilemma judgments (Hypothesis 1a, 1b), while in Study 2, we tested the interaction effect between personal force and intention (Hypothesis 2a, 2b, 3).

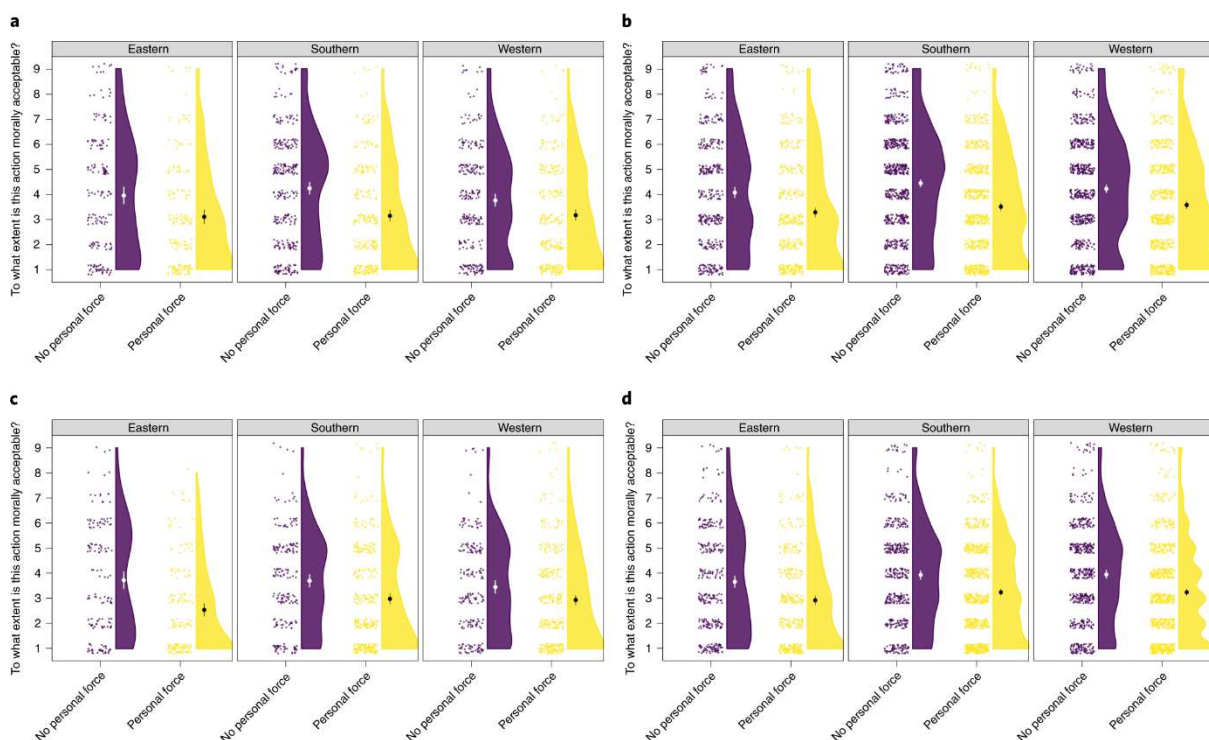
**Table 1** Summary of sample sizes and exclusions in all cultural clusters

	Eastern	Southern	Western	All
<b>Reason to exclude</b>				
N without exclusion	3,877	5,333	18,292	27,502
Careless responding	156 (4.0%)	82 (1.5%)	256 (1.4%)	494 (1.8%)
Confusion	752 (19.4%)	658 (12.3%)	1,718 (9.4%)	3,128 (11.4%)
Familiarity with moral dilemmas	1,669 (43.0%)	2,501 (46.9%)	10,332 (56.5%)	14,502 (52.7%)
Technical problem	531 (13.7%)	413 (7.7%)	1,225 (6.7%)	2,169 (7.9%)
Non-native speaker	347 (9.0%)	177 (3.3%)	1,305 (7.1%)	1,829 (6.7%)
Failed attention check (Study 1a)	720 (18.6%)	943 (17.7%)	1,311 (7.2%)	2,974 (10.8%)
Failed attention check (Study 1b)	849 (21.9%)	1,042 (19.5%)	1,336 (7.3%)	3,227 (11.7%)
Failed attention check (Study 2a)	1,102 (28.4%)	1,071 (20.1%)	4,900 (26.8%)	7,073 (25.7%)
Failed attention check (Study 2b)	1,195 (30.8%)	1,367 (25.6%)	5,528 (30.2%)	8,090 (29.4%)
<b>Final sample</b>				
Study 1a	381	622	566	1,569
Study 1b	327	553	546	1,426
Study 2a	323	690	2,971	3,984
Study 2b	277	576	2,660	3,513

Note. Study 1b and Study 2b refers to the Speedboat dilemmas (recall, all of our subjects answered to one trolley and a speedboat dilemmas)

## The effect of personal force

Findings are represented in Figure 1. To test the effect of personal force on moral judgement, we used one-sided  $t$ -tests. Consistent with our preregistration, we analysed only the continuous acceptability ratings (scale of 1-9), and not the binary choices. In each cultural cluster, we found at least strong evidence ( $BF_{10} > 10$ ) of an effect of personal force on moral judgement, which implies the effect is culturally universal. The results indicate that, when personal force is seen to be necessary to save more lives, people are less likely to favourably judge a consequentialist outcome (i.e., save more people). The results remained robust across dilemma contexts (i.e., trolley or speedboat version) and when including participants who were very familiar with these trolley-problem type scenarios. Therefore, our results replicated the findings of Greene et al. in the original cultural setting (H1a) and in the Southern and Eastern cultural clusters (H1b). The statistical results are summarised in Table 2.



*Figure 1.* Results of study 1 (effect of personal force). a–d, Results for trolley (a,b) and speedboat dilemmas (c,d) with all exclusion criteria applied (a,c) or including familiar participants (b,d). Error bars show 95% CI around the mean. Scale ranges from 1 (completely

unacceptable) to 9 (completely acceptable). Trolley problem: n = 1,569 when all exclusion criteria applied, and n = 3,524 when familiarity exclusion not applied. Speedboat dilemma: n = 1,426 when all exclusion criteria applied, and n = 3,295 when familiarity exclusion not applied.

**Table 2** *The effect of personal force on moral dilemma judgements*

Dilemma	Exclusion	Cluster	BF	RR	t	df	p	Cohen's d	Raw effect	89% CI
Trolley	Exclude	Eastern	1.9*10 <sup>2</sup>	7.00*10 <sup>-3</sup> , 14.00	-3.69	366.23	<.001	0.38	0.85	[0.39, 1.12]
		Southern	2.44*10 <sup>7</sup>	1.00*10 <sup>-5</sup> , 2.80*10 <sup>6</sup>	-6.32	619.93	<.001	0.51	1.10	[0.76, 1.33]
		Western	80.1	1.20*10 <sup>-2</sup> , 4.30	-3.41	553.15	0.001	0.29	0.59	[0.24, 0.79]
	Including familiar	Eastern	9.21*10 <sup>4</sup>	<1.50*10 <sup>-5</sup> , 6.50*10 <sup>3</sup>	-5.19	806.76	<.001	0.36	0.79	[0.51, 1]
		Southern	5.91*10 <sup>12</sup>	<1.00*10 <sup>-5</sup> , 5.50*10 <sup>11</sup>	-8.09	1345.85	<.001	0.44	0.94	[0.73, 1.1]
		Western	4.95*10 <sup>5</sup>	<1.00*10 <sup>-5</sup> , 2.90*10 <sup>4</sup>	-5.51	1338.48	<.001	0.30	0.65	[0.43, 0.8]
Speedboat	Exclude	Eastern	1.16*10 <sup>5</sup>	1.80*10 <sup>-5</sup> , 1.70*10 <sup>4</sup>	-5.26	283.92	<.001	0.59	1.18	[0.77, 1.47]
		Southern	1.01*10 <sup>3</sup>	1.30*10 <sup>-3</sup> , 74.00	-4.19	436.86	<.001	0.37	0.72	[0.37, 0.93]
		Western	25.2	3.30*10 <sup>-2</sup> , 1.20	-3.01	437.36	0.003	0.27	0.51	[0.18, 0.72]
	Including familiar	Eastern	2.4*10 <sup>4</sup>	<6.00*10 <sup>-5</sup> , 1.70*10 <sup>3</sup>	-4.88	680.10	<.001	0.37	0.74	[0.46, 0.95]
		Southern	7.8*10 <sup>6</sup>	<1.00*10 <sup>-5</sup> , 5.50*10 <sup>5</sup>	-5.94	908.97	<.001	0.36	0.69	[0.49, 0.85]
		Western	5.53*10 <sup>7</sup>	<1.00*10 <sup>-5</sup> , 4.0*10 <sup>6</sup>	-6.34	1140.72	<.001	0.35	0.71	[0.51, 0.87]

Note. BF = Bayes Factor, RR = Robustness Region of the prior

**The interaction effect of personal force and intention**

Figure 2 shows when we applied all exclusion criteria, we found strong evidence in the Western cluster (H2a) for the interaction between personal force and intention (BF<sub>10</sub> = 1.5\*10<sup>11</sup>), but moderate inconclusive evidence in the Southern (BF<sub>10</sub> = 9.4) and weak, inconclusive evidence in the Eastern clusters (BF<sub>10</sub> = 0.6) (H2b). More concretely, in the Western cluster, participants judged the acceptability of consequentialist decisions much lower when both personal force and intention had to be applied (i.e., the personal force effect was numerically greater when intention also had to be applied). When we included participants who were familiar with the trolley dilemma, we still found strong evidence in the Western cluster (BF<sub>10</sub> = 1.28\*10<sup>30</sup>) and, interestingly, we also found strong evidence in the Southern cluster (BF<sub>10</sub> = 3.1\*10<sup>6</sup>), but the evidence remained weak and inconclusive in the Eastern cluster (BF<sub>10</sub> = 2.9). Although in the preregistration we expected the effect sizes to be smaller when participants familiar with the trolley problem were included, we observed the direct opposite: when including data of participants familiar with the trolley problem, we found either equivalent or larger effect sizes in all cultural clusters. Notably, the size of the effect almost doubled in the Southern cluster when running the analysis on the sample with familiar and unfamiliar participants included (η<sub>p</sub><sup>2</sup> increased from .014 to .026). All statistical results are presented in Table 3.



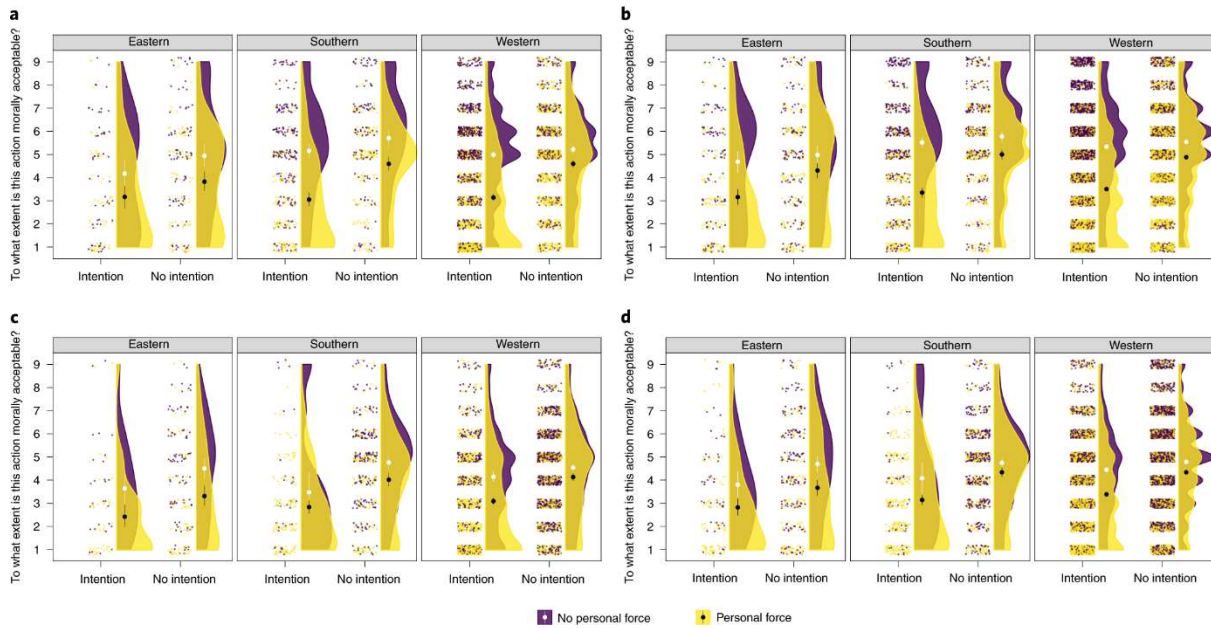


Figure 2. Results of study 2 (personal force and intention interaction). a–d, Results for trolley (a,b) and speedboat dilemmas (c,d) with all exclusion criteria applied (a,c) and including familiar participants (b,d). Error bars represent 95% CI. Scale ranged from 1 (completely unacceptable) to 9 (completely acceptable). Trolley problem: n = 3,984 when all exclusion criteria applied, and n = 9,844 when familiarity exclusion not applied. Speedboat dilemma, n = 3,513 when all exclusion criteria applied, and n = 9,006 when familiarity exclusion not applied.

Table 3

Interaction between personal force and intention on moral judgments

Dilemma	Exclusion	Cluster	BF	RR	b	89% CI	p	Partial $\eta^2$	Raw effect
Trolley	Exclusion	Eastern	0.59	2.20*10 <sup>-2</sup> , 0.64	0.027	[-0.16, 0.19]	0.84	0.000	0.11
		Southern	9.35	2.75*10 <sup>-2</sup> , 0.2	-0.250	[-0.35, -0.09]	0.002	0.014	-1.00
		Western	1.54*10 <sup>11</sup>	5.80*10 <sup>-5</sup> , 1.80*10 <sup>3</sup>	-0.306	[-0.36, -0.24]	<.001	0.019	-1.23
	Include familiar	Eastern	2.85	2.50*10 <sup>-2</sup> , 1.35	-0.213	[-0.33, -0.03]	0.031	0.008	-0.85
		Southern	3.08*10 <sup>6</sup>	2.23*10 <sup>-3</sup> , 60	-0.348	[-0.43, -0.25]	<.001	0.026	-1.39
		Western	1.28*10 <sup>30</sup>	<1.00*10 <sup>-5</sup> , 3.70*10 <sup>9</sup>	-0.292	[-0.33, -0.25]	<.001	0.018	-1.17
Speedboat	Exclusion	Eastern	0.43	4.60*10 <sup>-2</sup> , 0.69	-0.007	[-0.17, 0.2]	0.959	0.000	-0.03
		Southern	0.36	5.10*10 <sup>-2</sup> , 0.65	0.028	[-0.12, 0.16]	0.794	0.000	0.11
		Western	222	3.60*10 <sup>-2</sup> , 1.15	-0.160	[-0.22, -0.08]	<.001	0.005	-0.64
	Include familiar	Eastern	0.42	4.50*10 <sup>-2</sup> , 0.6	0.010	[-0.14, 0.16]	0.926	0.000	0.04
		Southern	1.13	3.20*10 <sup>-2</sup> , 0.94	-0.132	[-0.23, 0.01]	0.097	0.002	-0.53
		Western	4.75*10 <sup>7</sup>	6*10 <sup>-4</sup> , 75	-0.152	[-0.19, -0.11]	<.001	0.005	-0.61

Note. BF = Bayes Factor, RR = Robustness Region of the prior

On the speedboat dilemmas, we found strong evidence for the interaction in the Western cluster, regardless of the familiarity exclusion ( $BF_{\text{all exclusions}} = 222$ ,  $BF_{\text{with familiar}} = 4.8 \times 10^7$ ). However, we found inconclusive evidence in the Eastern and Southern clusters, both before ( $BF_{\text{Eastern}} = 0.4$ ;  $BF_{\text{Southern}} = 0.4$ ) and after ( $BF_{\text{Eastern}} = 0.4$ ;  $BF_{\text{Southern}} = 1.1$ ) familiarity exclusions. Although our results were consistent in the Western and Eastern clusters for both the speedboat and trolley dilemmas, there was a divergence in the Southern cluster. Specifically, we found strong evidence only for the interaction in the Southern cluster when we included familiar participants in the analysis. In general, in all clusters, the observed effect sizes were smaller on the speedboat than on the trolley dilemma.

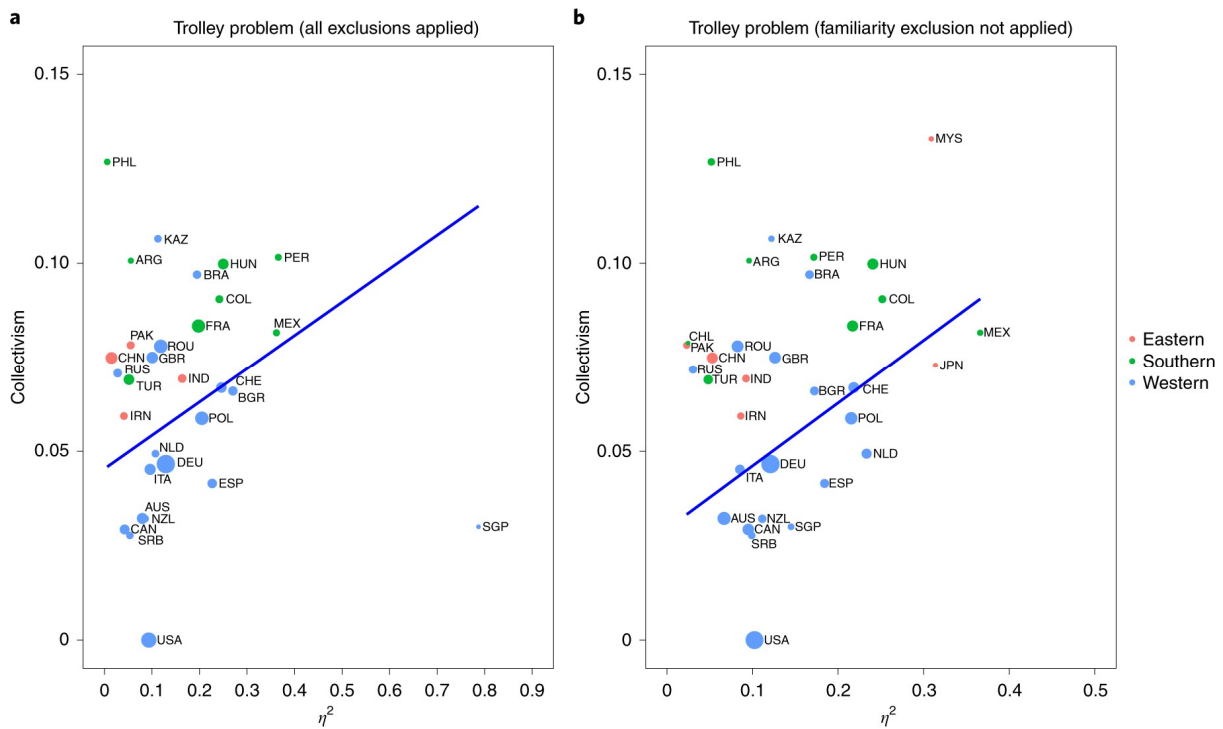
In summary, we conclude that we fully replicated the findings of Greene et al. with respect to the interaction of personal force and intention in the Western cluster (H2a) regardless of dilemma context or exclusion criteria. However, the evidence was inconclusive for all analyses of the Eastern cluster. In the Southern cluster, the conclusion is both context-dependent (i.e., the effect was only detectable in the trolley dilemma) and sensitive to exclusion criteria (i.e., the effect was only detectable when familiar participants were included).

To explore whether our results were sensitive to our choice of priors in the Bayesian analysis, we computed Robustness Regions (“RR”) that indicate the region of priors within which our inference would remain unchanged. The width of this region shows how robust our inferences are to our selection of priors. The RRs were generally wide for all statistical tests (see Tables 2-3), indicating that our results were not sensitive to our choices of prior. Thus, we would arrive at the same conclusions with any possible prior within the realistic range. One exception to this finding where the final conclusion was prior-dependent can be found in the analysis of the Southern cluster in Study 2. Specifically, if the scale of the prior distribution had been  $r = .21$  or higher (instead of  $r = .19$ ), we could have concluded that there was strong evidence for the effect (instead of saying that the test is inconclusive). Here, we would like to stress that we did not reach our registered sample size in this cluster for Study 2 (we registered that for 95% power, we would need 1,800 participants in each cluster of which we only reached 690 - see the Methods for details on sample size estimation). This could explain why our results did not reach our evidence thresholds and remained inconclusive.

### **Cultural correlates**

To test the “effects” of cultural variables, we used linear mixed models predicting moral acceptability ratings from different cultural variables with the random intercept of countries. We tested all five cultural variables one-by-one (i.e., country-level collectivism, and the four individual-level measures of horizontal and vertical collectivism/individualism), in separate linear models on the data with and without familiarity exclusion.

H3 stated that we expected a three-way interaction between country-level collectivism, intention, and personal force. We first tested this hypothesis on the data with familiarity exclusion applied (see Table 4 for statistical results and Figure 3 for the graphical representation of findings). The results of the country-level collectivism scale were inconclusive (trolley:  $BF_{10} = 1.2$ ; speedboat:  $BF_{10} = 0.9$ ). When analysing the individual-level measures of horizontal and vertical collectivism/individualism, all results were inconclusive. We conducted the same analysis on the sample but this time including participants who were familiar with these types of moral dilemmas, but the results were still inconclusive (trolley:  $BF_{10} = 2.2$ ; speedboat:  $BF_{10} = 0.7$ ). Analysing the individual-level individualism/collectivism measures, we found inconclusive evidence in all the scales. In the Introduction (Stage 1), we also hypothesized that country level collectivism would be associated with decreased overall acceptability of the utilitarian option. This hypothesis was not included in the registered analysis plan. Nevertheless, we added this analysis to the Supplementary Analysis section 3. In short, we found no evidence for the association between country-level collectivism and moral acceptability rates. Interestingly, nevertheless, we found strong evidence for a positive correlation between vertical individualism and moral acceptability ratings.



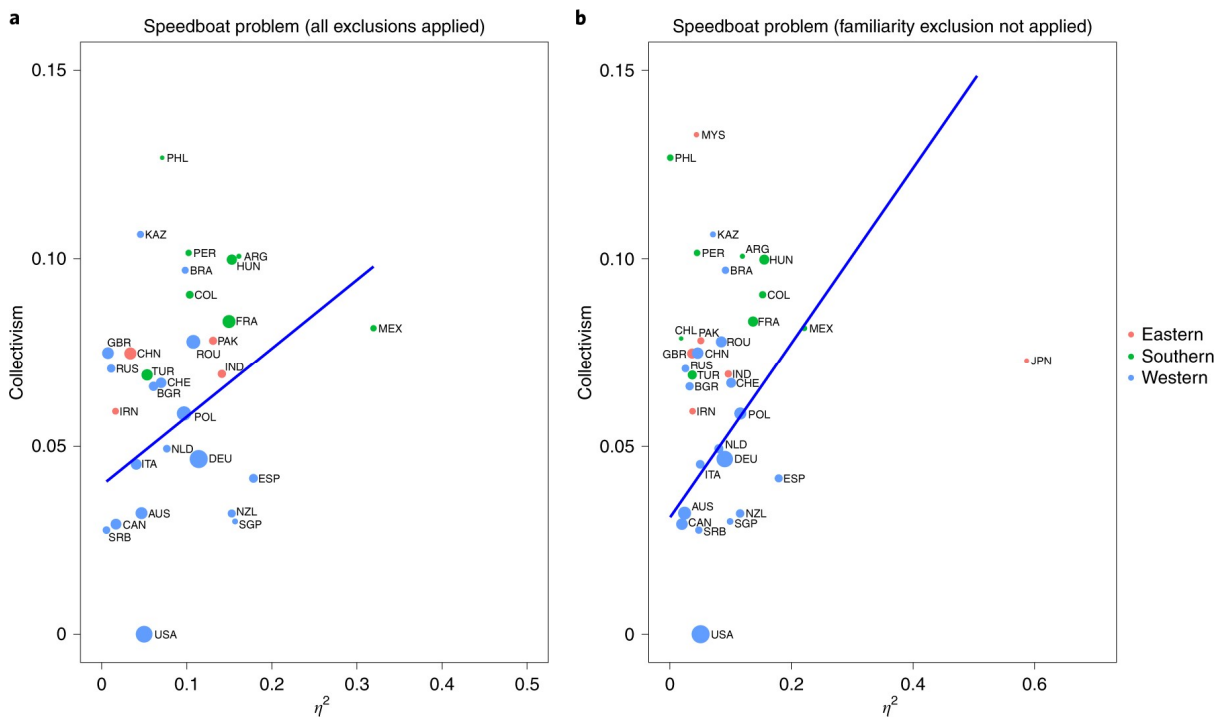
*Figure 3.* Correlation between country-level collectivism and effect size of the interaction between personal force and intention on the trolley problem. a,b, Correlation between country-level collectivism and the  $\eta^2$  effect size of the interaction between personal force and intention with all exclusion criteria applied (a) and including familiar participants (b) on the trolley problem. The size of the circles indicates the size of the sample in a given country. The blue line is the weighted regression. MYS, Malaysia; CHN, China; IND, India; THA, Thailand; MKD, Macedonia; PAK, Pakistan; IRN, Iran; JPN, Japan; GBR, Great Britain; FRA, France; HUN, Hungary; COL, Colombia; ARG, Argentina; TUR, Turkey; ECU, Ecuador; CHL, Chile; PER, Peru; PHL, Philippines; MEX, Mexico; USA, United States; SRB, Serbia; RUS, Russia; DEU, Germany; CAN, Canada; POL, Poland; ITA, Italy; KAZ, Kazakhstan; NZL, New Zealand; NLD, The Netherlands; ROU, Romania; BRA, Brazil; SGP, Singapore; ESP, Spain; AUS, Australia; BGR, Bulgaria; CHE, Switzerland.

We conducted the same analysis on the Speedboat dilemmas. Table 4 and Figure 4 presents the findings. Regardless of the familiarity exclusion criteria, we found inconclusive results in all cases.

**Table 4**

*Individualism/collectivism associations with the interaction between personal force and intention on moral judgments (Trolley dilemmas)*

Dilemma	Variable	With familiarity exclusion				No familiarity exclusion			
		BF	b	89% CI	p	BF	b	89% CI	p
Trolley	Country-level collectivism	1.17	-1.13	[-3.17, 1.12]	0.405	2.17	-1.27	[-2.53, -0.11]	0.096
	H. Collectivism	1.66	-0.03	[-0.06, 0.01]	0.263	2.31	-0.03	[-0.05, 0]	0.096
	H. Individualism	0.70	0.00	[-0.04, 0.04]	0.921	0.94	0.02	[-0.01, 0.04]	0.325
	V. Collectivism	0.88	0.00	[-0.03, 0.04]	0.988	0.71	-0.01	[-0.03, 0.01]	0.538
	V. Individualism	0.72	-0.02	[-0.05, 0.02]	0.451	0.45	-0.01	[-0.03, 0.01]	0.607
Speedboat	Country-level collectivism	0.91	0.66	[-1.43, 2.9]	0.631	0.66	-0.32	[-1.61, 0.83]	0.684
	H. Collectivism	3.11	-0.04	[-0.08, 0]	0.114	0.91	-0.01	[-0.04, 0.01]	0.396
	H. Individualism	1.11	-0.01	[-0.05, 0.03]	0.611	0.70	0.00	[-0.02, 0.03]	0.852
	V. Collectivism	1.53	0.02	[-0.01, 0.06]	0.311	0.96	0.01	[-0.01, 0.04]	0.357
	V. Individualism	0.70	0.00	[-0.04, 0.03]	0.952	0.54	0.01	[-0.01, 0.03]	0.590



*Figure 4. Correlation between country-level collectivism and effect size of the interaction between personal force and intention on the speedboat problem. a,b, Correlation between country-level collectivism and the  $\eta^2$  effect size of the interaction between personal force and intention with all exclusion criteria applied (a) and including familiar participants (b) on the*

speedboat problem. The size of the circles indicates the size of the sample in a given country. The blue line is the weighted regression. MYS, Malaysia; CHN, China; IND, India; THA, Thailand; MKD, Macedonia; PAK, Pakistan; IRN, Iran; JPN, Japan; GBR, Great Britain; FRA, France; HUN, Hungary; COL, Colombia; ARG, Argentina; TUR, Turkey; ECU, Ecuador; CHL, Chile; PER, Peru; PHL, Philippines; MEX, Mexico; USA, United States; SRB, Serbia; RUS, Russia; DEU, Germany; CAN, Canada; POL, Poland; ITA, Italy; KAZ, Kazakhstan; NZL, New Zealand; NLD, The Netherlands; ROU, Romania; BRA, Brazil; SGP, Singapore; ESP, Spain; AUS, Australia; BGR, Bulgaria; CHE, Switzerland.

## Exploratory analysis

### The effect of intention

We registered that we would test the main effect of intention by comparing the standard switch (no intention) and footbridge switch (intention) dilemmas. We found strong evidence in each cultural cluster and in each dilemma type for the effect of intention ( $BF_{10} > 10$ ). Importantly, the effect of intention remained unchanged even when we included participants who were familiar with moral dilemmas in the sample ( $BF_{10} > 10$ ). Tables 5-6 summarize the findings. As registered, we also tested the effect of physical force on moral judgement. In accordance with Greene et al., we found no evidence for this effect. See details in Supplementary Analysis section 2.1.

**Table 5**

*The effect of intention on moral dilemma judgements (Trolley dilemmas)*

Exclusion	Cluster	BF	t	df	p	Cohen's d	Raw effect	89% CI
Exclusion	Eastern	35.5	-3.13	159.97	0.002	0.41	0.99	[0.34, 1.36]
	Southern	4.29*10 <sup>6</sup>	-6.00	214.10	<.001	0.64	1.47	[0.99, 1.78]
	Western	1.95*10 <sup>15</sup>	-8.90	571.04	<.001	0.70	1.46	[1.17, 1.7]
Include familiar	Eastern	6.05*10 <sup>2</sup>	-3.93	234.76	<.001	0.40	0.91	[0.49, 1.2]
	Southern	5.29*10 <sup>13</sup>	-8.63	499.67	<.001	0.61	1.34	[1.04, 1.55]
	Western	3.3*10 <sup>34</sup>	-12.84	1278.97	<.001	0.64	1.33	[1.15, 1.47]
No exclusion	Eastern	30.6	-3.07	1060.61	0.002	0.17	0.39	[0.18, 0.57]
	Southern	1.61*10 <sup>14</sup>	-8.46	1421.86	<.001	0.40	0.89	[0.7, 1.04]
	Western	2.89*10 <sup>26</sup>	-11.01	2999.62	<.001	0.34	0.72	[0.62, 0.82]

**Table 6***The effect of intention on moral dilemma judgements (Speedboat dilemmas)*

<b>Exclusion</b>	<b>Cluster</b>	<b>BF</b>	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>	<b>Raw effect</b>	<b>89% CI</b>
Exclusion	Eastern	10.6	-2.67	192.91	0.008	0.35	0.78	[0.2, 1.12]
	Southern	2.81*10 <sup>5</sup>	-5.51	407.77	<.001	0.54	1.06	[0.68, 1.3]
	Western	3.15*10 <sup>9</sup>	-7.23	327.02	<.001	0.54	1.09	[0.81, 1.31]
Include familiar	Eastern	3.83*10 <sup>4</sup>	-4.99	319.39	<.001	0.48	1.03	[0.64, 1.3]
	Southern	9.55*10 <sup>6</sup>	-6.10	872.90	<.001	0.41	0.81	[0.57, 0.99]
	Western	2.51*10 <sup>16</sup>	-8.77	769.66	<.001	0.43	0.84	[0.68, 0.98]
No exclusion	Eastern	29.6	-3.06	1062.72	0.002	0.17	0.38	[0.18, 0.56]
	Southern	1.83*10 <sup>7</sup>	-6.12	1400.39	<.001	0.29	0.60	[0.43, 0.74]
	Western	2.42*10 <sup>12</sup>	-7.65	3006.15	<.001	0.23	0.47	[0.37, 0.56]

**No exclusion analysis (post-hoc)**

As the exclusion rate was very high in the above analyses (81%), we explored our results while applying no exclusion criteria (including all participants). In Study 1, we found strong evidence for the individual effects of personal force and intention, in each of the three cultural clusters, both in the speedboat and the trolley dilemmas—just as in our main analyses (see Extended Data Figures 1 and 2 for detailed results and data distribution).

For Study 2, Extended Data Figure 3 summarizes the statistical findings. Overall, we can conclude that almost all of our results regarding the effects of personal force and its interaction with intention are not sensitive to our exclusion. Only in the case of the Eastern cluster can we see a difference: without applying exclusions, strong evidence can be found for the effect of personal force and intention in the trolley dilemma, otherwise, we find inconclusive evidence. Here, we can only speculate whether the increased strength of evidence is due to the increased number of participants. The analysis on the speedboat dilemmas yielded the same results with and without exclusions: inconclusive evidence in the Eastern and Southern clusters, and strong evidence in the Western cluster (see Extended Data Figure 4 for the findings on Study 2). Thus, it appears that applying such strong exclusion criteria did not strengthen the replication effort nor substantially alter the inferences we draw about the replicability of the effect of force and intention.

We also conducted the cultural analysis without applying any exclusion criteria and we found that all of the results were inconclusive, with one exception. In the speedboat dilemma, we found moderate evidence that country level collectivism is positively associated with the interaction of personal force and intention (in line with our hypothesis;  $BF_{10} = 5.1$ ; same test for the trolley dilemma:  $BF_{10} = 2.8$ ). We also found moderate evidence ( $BF_{10} = 9.8$ ) that in the trolley dilemma, the interaction between personal force and intention is positively associated with individual-level horizontal collectivism: being higher on horizontal collectivism means a heightened personal force and intention interaction effect size (see Extended Data Figures 5 and 6; same test in the speedboat dilemma was inconclusive:  $BF_{10} = 0.54$ ). Thus, for the moderation of the effect by country-level collectivism, the strict exclusion criteria may have hurt our ability to detect these effects. Although these results appear in line with our prior hypothesis, this analysis was only exploratory, not registered a priori, and hence, should only be interpreted with caution.

As we registered, we added a figure showing the distribution of responses of both subscales of the Oxford Utilitarianism Scale for each country cluster, and also reported means and 95% confidence intervals, as registered. Moreover, we also added a post-hoc analysis correlating each subscales of the OUE with moral acceptability ratings of the moral dilemmas. We found that moral acceptability ratings correlate higher with the “instrumental harm” sub-scale ( $r = 0.40 - 0.45$ ) than with the “impartial beneficence” sub-scale ( $r = 0.05 - 0.20$ ) - with this latter correlation exhibiting somewhat larger cultural variations. Details can be found in the Supplementary Analysis section 2.4.

## Discussion

For centuries, philosophers and psychologists have explored the determinants of moral judgments. Moral dilemmas that force life and death decisions help us explore what norms and psychological processes drive our moral preferences. Initially, researchers thought<sup>41,42</sup> that people are simply susceptible to the doctrine of double effects when making moral judgements; harm is permissible if it occurs as an unintentional side-effect of an overall good outcome. Greene et al.<sup>18</sup>, however, showed that the role of using physical force to kill one (and save more) influenced moral judgments even more than did the intentionality of an action.



In this research, we replicated the design of Greene et al.<sup>18</sup> using a culturally diverse sample across 45 countries to test the universality of their results. Overall, our results support the proposition that the effect of personal force on moral judgments is likely culturally universal. This finding makes it plausible that the personal force effect is influenced by basic cognitive or emotional processes that are universal for humans and independent of culture. Our findings regarding the interaction between personal force and intention were more mixed. We found strong evidence for the interaction of personal force and intention among participants coming from Western countries regardless of familiarity and dilemma contexts (trolley or speedboat), fully replicating the results of Greene et al.<sup>18</sup>. However, the evidence was inconclusive among participants from Eastern countries in all cases. Additionally, this interaction result was mixed for participants from countries in the Southern cluster; we only found strong enough evidence when people familiar with these dilemmas were included in the sample and only for the trolley (not speedboat) dilemma.

Our general observation is that the size of the interaction was smaller on the speedboat dilemmas in every cultural cluster. It is yet unclear whether this effect is caused by some deep-seated (and unknown) differences between the two dilemmas (e.g., participants experiencing smaller emotional engagement in the speedboat dilemmas that changes response patterns), or is caused by some unintended experimental confound (e.g., order effect of the presented dilemmas). Furthermore, in the Eastern and Southern clusters, more participants found the dilemmas confusing than in the Western cluster (see Table 2). The increased confusion rates might have played a role behind the fact that we found no evidence for the personal force and intention interaction in the speedboat dilemmas; participants from the Southern and Eastern clusters might have struggled to follow some versions of the speedboat dilemma, as it was originally written for U.S. participants.

Furthermore, we hypothesised that collectivism would enhance the effect of personal force and intention. This prediction was based on the notion that collectivism increases the sensitivity to certain emotions which mediate these effects. We found no evidence for this hypothesis when we executed our preregistered analysis plan. However, in the exploratory analysis (with no exclusion criteria were applied), we found some moderate evidence for the association of country level collectivism in the speedboat dilemma, and individual level horizontal collectivism in the trolley dilemma with the interactional effect of personal force and intention. Since this analysis was not preregistered, these results should be cautiously interpreted.

The interaction between intention and personal force was sensitive to whether we included participants familiar with moral dilemmas. In the Southern cluster, this led to inconclusive evidence regarding the trolley problem, but contrary to our expectations, the size of all of the interaction effects were larger when we included familiar participants in the analysis. This increase could be due for at least two reasons: (1) familiarity is not the main reason behind the change in response patterns: familiarity correlates with an as yet unknown underlying variable, which induces a selection bias (e.g., educational background); and (2) familiarity is the main reason behind the change in response patterns: for example, being familiar with the trolley problem might have caused people to exhibit a lower emotional response to the problem or caused them to apply different reasoning that ended up affecting their responses. Our results cannot differentiate between the above described explanations (which are not necessarily mutually exclusive).

Although we found no strong evidence for the association between collectivism/individualism and the effects of personal force and intention, future research should test for other cultural variations. There are a number of interesting candidates that we did not examine, including cultural tightness<sup>43</sup> and social mobility<sup>44</sup>. Our database provides opportunities to the field to examine different aspects and cultural moderators of moral judgment.

This research has a number of limitations that future work will need to address. Although we call the personal force effect “universal”, it is only universal to the cultures we tested. This puts a limit to the “universality” of the effects: we did not (nor intended to) reach small scale hunter gatherer societies for example. Moreover, while our sample was more diverse and less WEIRD than that of Greene et al.’s research, it consisted of mostly educated individuals from younger age groups with internet access, raising similar concerns (e.g., still Educated and Industrialized, and possibly Rich, though not strictly Western or Democratic). Secondly, the data collection was conducted before and during the COVID-19 pandemic which could have affected the participants’ responding behaviour in some way (e.g., moral fatigue). Finally, 81% of the sample was not entered into the main confirmatory analyses because of our exclusion criteria, which might have resulted in unintended selection biases. For example, it is possible that more educated participants were more likely to get excluded due to being familiar with moral dilemmas from college. It is also possible that people with less working memory capacity or poor text comprehension abilities were more likely to be excluded due to the stringent attention checks. This is why we included an exploratory analysis in which we analysed data from all of our participants, without applying any exclusions. Our results on the full sample (no exclusion

criteria applied) supported our previous conclusions (that were drawn based on the data with exclusions) except in the cultural analysis, in which we found strong evidence for cultural variations only when no data were excluded. Thus, future work, especially replication work, should take caution when applying stringent exclusion criteria as it may be entirely unnecessary and even hurt the discovery of new effects.

Another limitation of our study might come from the fact that we used a single continuous measure of deontological/utilitarian tendencies. Although common in the field, such an approach has been criticized for being overly simplistic and not being able to pick up on more complex response patterns<sup>45,46</sup>. For example, maximizing outcome and rejecting harm are not necessarily symmetrical (as our continuous measure suggests). Hence, an interesting direction for future research could be to identify whether personal force and intention increase reliance on deontological rules or decrease reliance on consequentialist thinking. Methodological approaches, such as process dissociation, are promising in this regard<sup>40</sup>.

## **Conclusion**

With this replication study, we present empirical results about how people around the world make judgments in moral dilemmas that have long interested moral philosophers and psychologists. Empirical studies in this field have been conducted mostly on WEIRD samples, with little attention paid to cultural universality and variations. Our research allows us to avoid some important selection biases by having participants take the survey in their native language from 45 countries. The shared dataset should allow the assessment of different effects on moral dilemma judgments, such as religion or second language effects.

Overall, we found (1) the negative main effects of personal force and intention on moral dilemma judgments is universal; (2) the interaction between intention and personal force was replicated in the Southern and Western clusters, finding people are less likely to support sacrificing one person's life for the sake of saving the lives of several others, if they have both to intentionally engage in an action to do this and to use personal force; and (3) this interaction is not associated strongly with individual nor country-level collectivism/individualism measures.

## Method

### Participants

A large culturally and demographically diverse sample of participants was recruited from collaborating laboratories through the Psychological Science Accelerator<sup>47</sup>. The data collection team was originally proposed to include 146 labs from 52 countries. All of these participating laboratories obtained IRB approval (verified before the last round of Stage 1 submission). Combined, these labs committed to collect a minimum number of 18,637 participants. More labs were expected to be recruited before data collection commences. Each lab will recruit participants for the study by sending out the survey link along with the consent form to their participant pool, online platforms (such as Mturk), or testing them in the research lab. Due to some dropouts, the data collection team included 140 labs from 45 countries. Eligibility for participation was based on age ( $\geq 18$  years) and being a native speaker of the language of the test (more details on this criterion in the *Controlling for possible confounds* section). Data were collected either from local university participant pools or via data collection platforms (e.g., MTurk). Altogether, 41,090 participants started our survey, and 27,502 finished it whose data were analysed (17961 females, 7956 males, Mean age = 26.0 years, SD = 10.3 years; Study 1: 7,744 participants, 4,329 females, 2,487 males, Mean age = 26.8 years, SD = 11.1 years; Study 2: 19340 participants, 13,632 females, 5,469 males, Mean age = 25.8 years, SD = 9.98 years).

We did not collect any identifiable private data during the project that can be linked to individual survey responses. Each lab ascertained the agreement of the local institutional ethical review board with the proposed data collection. This study was conducted in accordance with the Declaration of Helsinki. The IRB approvals are available on our OSF project page: <https://osf.io/j6kte/>. Participants had to give an informed consent before starting the experiment. Only participants recruited through Mturk or Prolific received monetary compensation.

### Materials

**Moral dilemmas.** We used a total of six trolley dilemmas, namely: *footbridge switch*, *standard footbridge*, *footbridge pole*, *loop*, *obstacle collide* (taken from Greene et al.), and *standard switch*. All the materials are provided in the Supplementary Methods sections 1-3. Each of these

scenarios represents a different condition. For example, in the *standard footbridge* scenario both intention and personal force are required to push the man off the bridge. As in the original experiments, every participant was assigned to only one of these dilemmas. The problems were accompanied by a drawn sketch to aid understanding. Following the original procedure, after presenting each problem, participants were asked whether the described action (e.g., pushing the man to save five people) is morally acceptable or not (*Yes/No* response). After this judgement, participants were asked to indicate on a numbered Likert-type scale ranging from 1 (*completely unacceptable*) to 9 (*completely acceptable*), the extent to which they think that the given action is morally acceptable. Next, participants were asked to type the justification of their decision in an open question format. After participants were presented with the first trolley dilemma, they were presented with a second dilemma from the same condition, without drawn sketches. For the second dilemma, we used the so-called *speedboat dilemmas*. These dilemmas have been taken from Study 1b and 2b of Greene et al., and can be found in the Supplementary Methods section 1, with the exception of the dilemmas in the *obstacle collide* and *standard footbridge* conditions, which were provided by Joshua Greene during the review of the study. The order was fixed for dilemma presentation, so that the trolley version was always presented first. Study 1 was run before Study 2, but within study, participants were randomly assigned to one of the dilemmas within that study.

***Additional measures.*** Although the exploration of individual-level factors associated with moral thinking is not the aim of the present research, to enrich our database for future studies and secondary analyses, we expanded our survey with additional individual-level measures: 1) total yearly household income; 2) place of living (urban or rural area); 3) position on the four-dimensional Individualism-Collectivism scale<sup>34</sup> (16 items) for disentangling cultural differences in participants' responses<sup>48</sup>; 4) religion: Specific religion of the participant will be asked, plus one question to measure their level of religiosity: "On a scale from 1 to 10, how religious are you?". Furthermore, we included the Oxford Utilitarianism Scale<sup>28</sup> (9 items). Following these questions, participants' level of education, age, and sex were also recorded. We also recorded participants' country of origin, and whether the participant came from an immigrant background.

## Procedure

The experiment was administered by using a centralised online survey that participants could answer remotely or in the lab. We used the original instructions of Greene et al., as presented in the Supplementary Methods section 1. After responding to the dilemmas, participants were asked to answer three questions: (1) a measure of careless responding (question about the specifics of the trolley scenario); (2) whether they found the material confusing; and (3) whether they found the description of the problem realistic. After these questions, participants were directed to our series of questionnaires: the Oxford Utilitarianism Scale, followed by the Individualism-Collectivism Scale, and the measures of religion. Next, we administered the demographic questions (income, place of living, country of origin, immigrant background, level of education, age, and sex). Afterwards, we asked three further questions to measure careless responses, participants' familiarity with research questions, and finally, we asked for further comments or any experienced technical problems.

***Controlling for possible confounds.*** To avoid second language effects on moral judgement<sup>49</sup>, only native speakers of the language of the experiment could participate. To ensure this, we asked participants to indicate their native language(s). Bilinguals could choose their preferred language. The data of anyone with a native language different from the language of the survey were removed from data analyses.

Following Greene et al.'s procedure, data from participants who reported that they found the material confusing were excluded from the analyses. Data from participants who reported having experienced technical problems during the experiment were also excluded from all analyses. To avoid careless responses, we added three bogus items at the end of the survey. We asked participants very basic questions (e.g., "I was born on February 30th.") to which incorrect answering indicates careless responding<sup>50</sup>. We excluded data from participants who gave an incorrect response to any of these questions. Moreover, we introduced two additional questions (presented right after the moral dilemmas), asking participants about the specifics of the trolley and speedboat scenarios that they had been presented with, to test whether they had paid attention when reading the scenarios (referred to as attention check in the later test). Specifically, participants were asked to select the option which most accurately described the situation that they had been presented with. Each option described the nature of the physical action that was the key manipulation in the experiment. As attention to the trolley and speedboat

dilemmas was measured by different questions, when analysing the responses, we excluded the data for the correspondingly failed attention check question. For example, people who gave a correct response on the trolley, but not on the speedboat attention check question, were included when analysing the trolley dilemma, and excluded when analysing the speedboat version.

As moral dilemmas are becoming more and more common in psychological research and in summaries of this research in popular media and culture and teaching, it is possible that some participants may have previous knowledge of these dilemmas, which may affect their responses. To address this potential problem, at the end of the experiment participants were asked the following question: “Before this experiment, were you familiar with moral dilemmas of this kind, in which you can save more people by causing the death of one person?” Answers were given on a rating scale from 1 (*absolutely not familiar*) to 5 (*absolutely familiar*). Familiarity with the trolley problem or such moral dilemmas (participants who responded with 4 or 5 on this scale) was used as a further exclusion criterion. Additionally, participating labs were asked to avoid recruiting philosophers or philosophy students because they are likely to have heard about trolley problems, and we wanted to minimise the number of participants to be excluded following data collection.

### **Notable deviations between this study and the design of Greene et al.**

Besides the multinational data collection that forms the crux of our project, the first important methodological difference between this study and the original study is that the original study was conducted by paper and pencil, whereas we administered the experiment online. Of note, recent research found no evidence for a difference between the behaviour of participants who took part in the experiment online versus those who took part in the experiment in the lab. We also added one change in the introduction of the experiment (see Supplementary Methods section 1); participants were not given the opportunity to ask the researcher any questions before the experiment (as the experiment can be administered online, they did not have the opportunity to do so).

The second important change in this experiment is that participants were presented with two moral dilemmas in one condition, instead of one. These additional dilemmas will be analysed separately, as they were in the original experiment. The third difference is that for Study 2, we used different moral dilemmas than those that were used by Greene et al.; the standard switch and footbridge dilemmas were used instead of the loop weight and obstacle push dilemmas,

respectively. These dilemmas are not different from the ones used by Greene et al. in their structural characteristics, only on surface characteristics. That is, in the standard switch the harm is unintended and no personal force is required, while in the standard footbridge dilemma, the harm is intended and requires personal force. By including the standard switch and standard footbridge scenarios instead of the original ones, we gain further insight into the data. Imagine for example, that the personal force effect does not replicate in one of the cultural clusters. One explanation for this is that people are simply not sensitive to the effect of personal force in that cluster. However, it might also be the case that utilitarian response rates to similar dilemmas increase over time<sup>51</sup>. If so, we should see that the replicated difference between the standard footbridge and switch dilemmas is shrinking or

disappeared. Furthermore, by comparing the standard footbridge to the footbridge pole dilemmas, we can test the effect of physical contact, and by comparing the standard switch case to the footbridge switch case to confirm the effect of intention.

Finally, in the original experiment, Greene et al. excluded participants who did not manage to suspend disbelief. Nevertheless, as they noted, this had no effect on their results. Thus, we decided that we would not use this exclusion criterion.

***Cultural classification of countries.*** To test the cultural universality hypothesis, a comprehensive cultural classification is needed that encompasses multiple sources of cultural variability. Hence, to assess our first hypothesis on the universality of the effect of personal force and intention on moral judgements, we used the cultural classification of Awad et al.<sup>35</sup>. Based on surveyed moral preferences, they identified three distinct clusters of countries: Eastern, Southern, and Western. They argued that this cluster structure is broadly consistent with the alternative, but more complex Inglehart-Welzel cultural map<sup>34</sup>. Therefore, we assigned the countries of our participating labs to these cultural clusters, as listed in Supplementary Analysis Section 1, Table S1.

**Language adaptation.** The participating labs translated the survey items into the language of the participant pool, following the translation process of the PSA (<https://psysciacc.org/translation-process/>) detailed below.

1. Translation: Original document is translated from source to target language by A translators resulting in document Version A
2. Back-translation: Version A is translated back from target to source language by B Translators independently resulting in Version B



3. Discussion: Version A and B are discussed among translators and the language coordinator, discrepancies in Version A and B are detected and solutions are discussed. Version C is created.
4. External Readings: Version C is tested on two non-academics fluent in the target language. Members of the fluent group are asked how they perceive and understand the translation. Possible misunderstandings are noted and again discussed as in Step 3.
5. Cultural Adjustments: Data collection labs read materials and identify any needed adjustments for their local participant sample. Adjustments are discussed with the Language Coordinator, who makes any necessary changes, resulting in the final version for each site.

### **Planned analyses**

#### **Preregistered analysis**

#### **Confirmatory Replication Analyses**

As explained in the introduction, we focused our analyses on the question of universality of Greene et al.'s two most important claims. We conducted independent analyses in each cultural cluster and reported them separately. We preregistered the following hypotheses:

*Hypothesis 1a:* There is an effect of personal force on moral judgement in the Western cluster (replication of the original effect).

*Hypothesis 1b:* If the effect of personal force is culturally universal, there is an effect of personal force on the moral acceptability ratings (Greene et al., Study 1) in the Southern and Eastern cultural clusters as well.

*Hypothesis 2a:* There is an interaction between personal force and intention (Greene et al., Study 2) in the Western cluster (replication of original effects). More specifically, the intention factor is larger when personal force is present compared to when personal force is absent.

*Hypothesis 2b:* If this effect is culturally universal, there is an effect in the Southern and Eastern cultural clusters as well.

Unlike in the original study, we employed Bayesian analyses to gain information from our data concerning the strength of evidence for the null and alternative hypotheses. The Bayes factor indicates the relative evidence provided by the data comparing two hypotheses<sup>52</sup>. Regarding the threshold of strong Bayesian evidence, we followed the recommendations of<sup>53</sup> and set the decision threshold of  $BF_{10}$  to  $> 10$  for  $H_1$  and  $< 1/10$  for  $H_0$ . We used informed priors for the alternative model: a one-tailed Cauchy distribution with a mode of zero and a scale  $r = 0.26$  (Hypothesis 1a and 1b) and  $r = 0.19$  (Hypothesis 2a and 2b) on the standardized effect size using the BayesFactor package<sup>54</sup> in R for the analysis. These priors are based on the effect sizes that we expect to find as explained below in the sample size estimation section. We will implement all of our analyses with the R statistical software<sup>55</sup>.

To test Hypothesis 1a and 1b, we compared the moral acceptability ratings given on the footbridge switch problem and footbridge pole dilemma, with the moral acceptability rating of the footbridge switch dilemma expected to be higher. More concretely, we performed three one-sided Bayesian  $t$ -tests with the same comparison in each cultural group. For each cultural cluster, we would conclude that we replicated the original effect if Bayes factor ( $BF_{10}$ )  $> 10$ , we would conclude that we found a null effect if  $BF_{10} < 1/10$ , and we would conclude that the results are inconclusive if we find a  $BF_{10}$  in between these numbers (see below for justification of these thresholds).

To test Hypothesis 2a and 2b, we tested the interaction of personal force and intention in each cultural cluster, separately. We conducted Bayesian linear regression analysis in each cultural cluster. The Bayes factor of interest is defined as the quotient of the model including the interaction and two main effects (numerator) and the model including only the two main effects (denominator). For each cultural group, we would conclude that we replicated the original effect if the Bayes factor of the interaction ( $BF_{10}$ )  $> 10$ , we would conclude that we found a null effect if  $BF_{10} < 1/10$ , and we would conclude that the results are inconclusive if we find a  $BF_{10}$  in between these values (see below for justification of these thresholds). To further understand the direction of the interaction, we will plot out the results in each cultural cluster. To conclude the replication of the original effect, we should find that the intention effect is higher in the personal force condition than in the no personal force condition.

Note that we conducted and reported the frequentist version of the proposed analysis (e.g., *t* tests for each hypothesis, for each cultural class) for the sake of comparability of the original and our results. Nevertheless, we regarded the results of our Bayesian analyses the basis of our statistical inference. Although we registered that the frequentist statistics would only be added as the supplementary material, we added it to the main text for easier comparability. No inference was drawn from the frequentist statistics.

Test assumptions for the statistical tests (t-tests and linear regressions) were assumed to hold true, but they were not formally tested.

### **Robustness analyses**

To probe the robustness of our conclusions to the scaling factor of the Cauchy distribution used as the prior of H1, we reported Robustness Regions for each Bayes factor. Robustness Regions were notated as RR[*min*, *max*], where *min* indicates the smallest and *max* indicates the largest scaling factor that would lead us to the same conclusion as the originally chosen scaling factor<sup>56</sup>.

### **Sampling plan and stopping rule**

As the data were planned to be collected globally, our knowledge was insufficient concerning the noise of the measurement and the rate of exclusion in the various samples, which were needed for an accurate sample size estimation. For this reason, we proposed a sequential data acquisition. That is, first, we launched Study 1 (Hypotheses 1a and 1b), and collected data in sequences from 500 participants per cluster per condition; from 3,000 participants altogether (after all exclusions). We stop data collection after each sequence. At these stops, we conducted our planned Bayesian analyses. Should the BF reach the preset thresholds in a given cluster, we will stop data collection for that cluster. If, in a cluster, the BF thresholds were not reached, we would continue data collection with 200 additional participants per cluster per condition, and then re-analyse the data, repeating this procedure until one of the BF thresholds is reached, or the participant pool is exhausted. Note, however, that we deviated from this sampling plan. See “Deviations from registration” for details.

Should we not have reached this limit with our planned capacity of ~19,000 participants, we would have extended the data collection to a new semester. In the case that we would have not

reached our evidence threshold within 12 months, we would have reported our final results, acknowledging the limited strength of the findings.

We launched Study 2 data collection in a given cluster only when the analysis of Study 1 was conclusive. In Study 2, we conducted the analysis only when we had exhausted our resources.

### **Sample Size estimation**

To calculate our needs for data collection, we conducted a rough sample size estimation. Assuming that the original effect size is found in Study 1 ( $d = 0.4$ ), our sample size estimation indicated that we would require 500 participants per condition per cluster (3,000 altogether), while if the original effect size is to be found in Study 2 ( $d = 0.28$ ), our estimation indicated that we would need 1,800 participants per condition per cluster (21,600 altogether for Study 2) to obtain 95% of power in detecting the effect. A detailed description of the Sample Size estimation can be found in Supplementary Methods section 4.

### **Testing the association between country-level collectivism and the effects of personal force and intention**

Our third hypothesis proposed that collectivism increases the effects of personal force and intention. As a measure of country-level individualism and collectivism, we added the Collectivism measure from the Cultural Distance WEIRD scale (countries' differences in terms of individualism from the United States)<sup>57</sup> as a continuous variable in our model. We tested whether collectivism interacted with personal force and intention (Hypothesis 3), as explained in the introduction. Hypothesis 3 expected to find a three-way interaction between collectivism, intention, and personal force, for which we used the dilemmas we used to test Hypotheses 2a and 2b. In this analysis, we used a Cauchy distribution with a scale of  $r = 0.37$  (same we used to test Hypothesis 2a and 2b, i.e., the test of the interaction) as prior. Should we find evidence for null effect ( $BF < 1/10$ ) of the interaction of individualism/collectivism, personal force, and intention, we would conclude that individualism/collectivism does not moderate the effect of personal force and intention.

## **Analysis of the additional moral dilemmas**

### **Study 1.**

As we explained above, each participant had to give a response on two moral dilemmas. For Study 1 (effect of personal force), we conducted the same analysis on the rest of the moral dilemmas, without the trolley versions, as in the original study (Study 1b; Greene et al.).

### **Study 2.**

We conducted the same analysis (interaction of personal force and intention) on the rest of Speedboat dilemmas, without the trolley versions.

## **Further tests**

**Effect of physical contact and intention.** With this set of items, we were able to assess the effect of physical contact, by comparing the standard footbridge and footbridge pole dilemmas. We also assessed the effect of intention by comparing the standard switch case with the footbridge switch case. These analyses were done in every cluster, and we used Bayesian t-tests for these comparisons. We used the same prior we use for the assessment of the effect of physical force ( $r = 0.26$ ). This analysis was done separately on the trolley and speedboat dilemmas.

**Comparing the standard switch and standard footbridge dilemmas.** For the reasons explained earlier, we compared the standard footbridge and standard switch dilemmas, in each cultural cluster. For this, we conducted a Bayesian t-test, with the same prior previously used for the assessment of the effect of physical force ( $d = 0.26$ ). This analysis was done separately for the trolley and speedboat dilemmas.

**Oxford Utilitarianism Scale.** We computed a figure showing the response distribution of each subscales of the Oxford Utilitarianism Scale<sup>39</sup> for each cultural cluster to explore potential cultural differences (along with means and 95% CI). The results of this can be found in the Supplementary Analysis section 2.4.

**Individual-level horizontal and vertical individualism-collectivism.** Triandis and Gelfand<sup>45</sup> defined individualistic and collectivistic cultural tendencies among 4 dimensions: *vertical individualism*, *vertical collectivism*, *horizontal individualism*, and *horizontal collectivism*. We added these continuous measures to our Bayesian linear regression analysis. The predictive power of all four measures were assessed separately.

**Including familiar participants.** A potentially large number of participants were excluded due to familiarity with the trolley dilemma, and there was a possibility that this exclusion criterion will affect the data from some countries or cultural clusters more than others. To avoid this potential sampling bias, we computed all above-listed analyses on moral dilemmas (confirmatory and exploratory) on the full sample in which we did not exclude the participants who were familiar with moral dilemmas. Second, we computed all analyses specifically on data coming from people who were familiar with moral dilemmas in order to compare the results of “familiar” and “unfamiliar” participants. This latter analysis can be found in the Supplementary Analysis section 2.3 and was limited to the confirmatory hypothesis tests.

**Pilot testing.** To ascertain that the survey software operates without any technical problems, we planned to conduct a pilot test in which each participating lab would have been expected to collect data from 10 participants. We would have only assessed the expected functioning of the survey software without analysing the collected data.

**Timeline.** We planned to finish data collection within six months from Stage 1 in principle acceptance and we planned to submit our report within one month from then.

### **Deviations from registration**

We preregistered that we would collect data from 3,000 participants for Study 1 (test of personal force; H1a, H1b), after exclusions. Unexpectedly, the exclusion criteria led to 80.6% exclusion of our collected data. At the point when this was realized, it seemed likely that Study 1 would exhaust the available sample pool, not leaving capacity for Study 2. Therefore, with the agreement of the journal editor, we decided to collect participants for Study 1 only until our Bayes Factor evidence thresholds were reached after all exclusion criteria were applied. This modification allowed us to collect data for Study 2 as well.

At the time of this decision, the distribution of responses has been taken into account: we had collected data from 3,473 participants: 1319 from the “Western cluster”, 1762 from the “Southern” cluster, and 392 from the “Eastern” cluster. After exclusions, 789 participants remained (78% excluded): 296 from the “Western” cluster (78% excluded), 429 from the “Southern” cluster (76% excluded), and 64 from the “Eastern” cluster (84% excluded).

Instead of conducting a pilot study as preregistered, in order to avoid wasting any (much needed) participants, participating researchers from all labs tested the experiment before it was sent out to assure that there are no grammatical mistakes or functionality problems.

Due to COVID-19 crisis, data collection took 6 months longer than expected (with the agreement of the editor).

### **Exploratory analysis**

During the data pre-processing, we excluded 229 participants from three US-based labs as they received a wrong survey link. Furthermore, 13,359 participants started, but did not finish the experiment, therefore their data were also dropped from further analyses. These participants did not count towards our final sample and are not part of the data in any way. The final sample used for data analyses consisted of 27,502 participants. Further information on the demographics of our participants can be found in the Supplementary Analysis section 1.

Note that we limited the use of Robustness Regions for the confirmatory hypothesis tests.

### **Data availability statement**

Collected anonymised raw and processed data are publicly shared on the Github page of the project: <https://github.com/marton-balazs-kovacs/trolleyMultilabReplication/tree/master/data>.

### **Code availability statement**

Code for data management and statistical analyses have been written in R and are available at: <https://github.com/marton-balazs-kovacs/trolleyMultilabReplication>.

## Protocol Registration Information

The Stage 1 protocol for this Registered Report was accepted in principle on 30th January 2020.

The protocol, as accepted by the journal, can be found at <https://doi.org/10.6084/m9.figshare.11871324.v1>

## Supplementary Information

Supplementary Methods 1–4, Tables 1–10 and Figs. 1–5.

[https://static-content.springer.com/esm/art%3A10.1038%2Fs41562-022-01319-5/MediaObjects/41562\\_2022\\_1319\\_MOESM1\\_ESM.pdf](https://static-content.springer.com/esm/art%3A10.1038%2Fs41562-022-01319-5/MediaObjects/41562_2022_1319_MOESM1_ESM.pdf)

## Author Contribution

**Conceptualization:** B. Bago and B.A.

**Data curation:** B. Bago, M. Kovacs, and T.N.

**Formal analysis:** B. Bago, M. Kovacs, T.N., Z.K., and B.P.

**Funding acquisition:** P.A., P.M., K.Q., I. Zettler, and R. Hoekstra

**Investigation:** B. Bago, M. Kovacs, M.A., S. Adamus, S. Albaloooshi, N.A.-A., S. Alper, S.A.-S., S.G.A., S. Amaya, P.K.A., G.A., D.A., P.A., J.J.B.R.A., A.A., P.B., K.B., B. Bashour, E. Baskin, L. Batalha, C.B., J. Bavolar, F.B., M. Becker, B. Becker, A.B., M. Białek, E. Bilancini, D.B., L. Boncinelli, J. Boudesseul, B.T.B., E.M.B., M.M.B., D.P.C., N.C.C., J.B.C., C.R.C., W.J.C., P.C., H.C.-P., R.F.C., O.Ç., R.C.C., V.C.A., C.P.C., S.C., Y.D., J.A.M.d.G., W.C.d.V., E.G.D.B., C.D., B.J.W.D., X.D., F.D., A.D., N.B.D., J.E., C.E.-S., L.E., T.R.E., G.F., F.M.F., S.F., A. Findor, A. Fleischmann, F.F., R.F., D.-A.F., C.H.Y.F., S.G., O.G., A.-R.G.-N., M.E.G., I.G., T.G., B.G., M. Gollwitzer, A.G., M. Grinberg, A.G.-B., E.A.H., A.H., W.A.N.M.H., J.H., K.R.H., J.J.J.H., E.H., M.H., C.A.H., R. Huskey, A. Ikeda, Y.I., G.P.D.I., O.I., C.I., A. Iyer, B. Jaeger, S.M.J.J., W.J.-L., B. Jokić, P.K., V.K., G.K., F.K.-M., A.T.A.K., K.M.K., B.J.K., H.E.K., R.I.K., M. Kowal, E.K., L.K., A.K., A.O.K., F.L., C.L., J.L., E.B.L., A.L., I.Y.-M.L., L.B.L., M.C.L., J.N.L., C.A.L., S.C.L., M.L., Y.L., H.L., T.J.S.L., S.L., M.T.L., P.L., J.G.L., T.L., M. Máčel, S.P.M., M. Maganti, Z.M.-M., L.F.M., H.M., G.M.M., D.M.S., C.-J.M., A.D.A.M., M. Mazidi, J.P.M., N.M., M.C.M., L.M., T.L.M., A. Mirisola, M. Misiak, P.M., M.M.-J., A. Monajem, D.M., E.D.M., E.N., I.N., D.P.O., J.O., N.C.O., A.A.Ö., M. Panning, M.P.-P., N.P., P.P., M.P.-C., M. Parzuchowski, J.V.P., J.M.P., M. Peker, K.P., L.P., I.P., M.R.P., N.P.-J., A.J.P., M.A.P., E. Pronizius, D.P., E. Puvia, V.Q., K.Q., A.Q., B.R., D.A.R., U.-D.R., C.R., K. Reynolds, M.F.F.R., J.P.R., R.M.R., P.R., F.R.-D., S.R.-F., B.T.R., K. Rybus, A. Samekin, A.C.S., N.S., C.S., K.S., K.A.Š., M. Sharifian, J. Shi, Y.S., E.S., M. Sirota, M. Slipenkyj, Ç.S., A. Sorokowska, P.S., S. Söylemez, N.K.S., I.D.S., A. Sternisko, L.S.-W., S.L.K.S., S. Stieger, D.S., J. Strube, K.J.S., R.D.S.-C., N.M.S., B. Takwin, S.T., A.G.T., K.E.T.,



L.E.T., M. Tonković, B. Trémolière, L.V.T., B.N.T., M. Twardawski, M.A. Vadillo, Z.V., L.A.V., B.V., D.V., M.V., M.A. Vranka, S. Wang, S.-L.W., S. Whyte, L.S.W., A.W., X.W., F.X., S. Yadanar, H.Y., Y.Y., O.Y., S. Yoon, D.M.Y., I. Zakharov, R.A.Z., I. Zettler, I.L.Ž., D.C.Z., J.Z., X.Z., and B.A.

**Methodology:** B. Bago, Z.K., B.P., R. Hoekstra, and B.A.

**Project administration:** B. Bago, M. Kovacs, J.P., and B.A.

**Resources:** B. Bago, M. Kovacs, S.G.A., G.A., P.A., H.C.-P., R.C.C., Y.D., X.D., W.J.-L., F.K.-M., C.L., H.M., A.A.Ö., V.Q., A.C.S., Y.S., J. Strube, N.M.S., M.V., I. Zakharov, and B.A.

**Supervision:** B. Bago, J.P., P.A., A.A.Ö., P.S., M.V., and B.A.

**Validation:** B. Bago, M. Kovacs, T.N., Z.K., B.P., and B.A.

**Visualization:** B. Bago, M. Kovacs, T.N., and B.P.

**Writing - original draft:** B. Bago and B.A.

**Writing - review & editing:** B. Bago, M. Kovacs, J.P., B.P., M.A., S. Adamus, S. Albaloooshi, N.A.-A., S. Alper, S.A.-S., S.G.A., S. Amaya, P.K.A., G.A., D.A., P.A., J.J.B.R.A., A.A., P.B., K.B., B. Bashour, E. Baskin, L. Batalha, C.B., J. Bavolar, F.B., M. Becker, B. Becker, A.B., M. Białek, E. Bilancini, D.B., L. Boncinelli, J. Boudesseul, B.T.B., E.M.B., M.M.B., D.P.C., N.C.C., J.B.C., C.R.C., W.J.C., P.C., H.C.-P., R.F.C., O.Ç., R.C.C., V.C.A., C.P.C., S.C., Y.D., J.A.M.d.G., W.C.d.V., E.G.D.B., C.D., B.J.W.D., X.D., F.D., A.D., N.B.D., J.E., C.E.-S., L.E., T.R.E., G.F., F.M.F., S.F., A. Findor, A. Fleischmann, F.F., R.F., D.-A.F., C.H.Y.F., S.G., O.G., A.-R.G.-N., M.E.G., I.G., T.G., B.G., M. Gollwitzer, A.G., M. Grinberg, A.G.-B., E.A.H., A.H., W.A.N.M.H., J.H., K.R.H., J.J.J.H., E.H., M.H., C.A.H., R. Huskey, A. Ikeda, Y.I., G.P.D.I., O.I., C.I., A. Iyer, B. Jaeger, S.M.J.J., W.J.-L., B. Jokić, P.K., V.K., G.K., F.K.-M., A.T.A.K., K.M.K., B.J.K., H.E.K., R.I.K., M. Kowal, E.K., L.K., A.K., A.O.K., F.L., C.L., J.L., E.B.L., A.L., I.Y.-M.L., L.B.L., M.C.L., J.N.L., C.A.L., S.C.L., M.L., Y.L., H.L., T.J.S.L., S.L., M.T.L., P.L., J.G.L., T.L., M. Máčel, S.P.M., M. Maganti, Z.M.-M., L.F.M., H.M., G.M.M., D.M.S., C.-J.M., A.D.A.M., M. Mazidi, J.P.M., N.M., M.C.M., L.M., T.L.M., A. Mirisola, M. Misiak, P.M., M.M.-J., A. Monajem, D.M., E.D.M., E.N., I.N., D.P.O., J.O., N.C.O., A.A.Ö., M. Panning, M.P.-P., N.P., P.P., M.P.-C., M. Parzuchowski, J.V.P., J.M.P., M. Peker, K.P., L.P., I.P., M.R.P., N.P.-J., A.J.P., M.A.P., E. Pronizius, D.P., E. Puvia, V.Q., K.Q., A.Q., B.R., D.A.R., U.-D.R., C.R., K. Reynolds, M.F.F.R., J.P.R., R.M.R., P.R., F.R.-D., S.R.-F., B.T.R., K. Rybus, A. Samekin, A.C.S., N.S., C.S., K.S., K.A.Š., M. Sharifian, J. Shi, Y.S., E.S., M. Sirota, M. Slipenkyj, Ç.S., A. Sorokowska, P.S., S. Söylemez, N.K.S., I.D.S., A. Sternisko, L.S.-W., S.L.K.S., S. Stieger, D.S., J. Strube, K.J.S., R.D.S.-C., N.M.S., B. Takwin, S.T., A.G.T., K.E.T., L.E.T., M. Tonković, B. Trémolière, L.V.T., B.N.T., M. Twardawski, M.A. Vadillo, Z.V., L.A.V., B.V., D.V., M.V., M.A. Vranka, S. Wang, S.-L.W., S. Whyte, L.S.W., A.W., X.W., F.X., S. Yadanar, H.Y., Y.Y., O.Y., S. Yoon, D.M.Y., I. Zakharov, R.A.Z., I. Zettler, I.L.Ž., D.C.Z., J.Z., X.Z., and B.A.

### Competing interests statement

The authors declare no competing interests.

**References**

1. Mill, J. S. & Bentham, J. *Utilitarianism and other essays*. (Penguin, 1987).
2. Kant, I. *Groundwork for the metaphysics of morals*. (Yale University Press, 1785).
3. Greene, J. D. *Moral tribes: emotion, reason and the gap between us and them*. (Penguin Press, 2013).
4. London, J. A. How should we model rare disease allocation decisions? *Hastings Cent. Rep.* 42, 3 (2014).
5. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* 352, 1573–1576 (2016).
6. Foot, P. The problem of abortion and the doctrine of double effect. *Oxf. Rev.* 5, 5–15 (1967).
7. Baron, J. Utilitarian vs. deontological reasoning: method, results, and theory. in *Moral inferences* (eds. Bonnefon, J.-F. & Trémolière, B.) 137–151 (Psychology Press, 2017).
8. Baron, J. & Gürçay, B. A meta-analysis of response-time tests of the sequential two-systems model of moral judgment. *Mem. Cognit.* 45, 566–575 (2017).
9. Cushman, F., Young, L. & Hauser, M. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychol. Sci.* 17, 1082–1089 (2006).
10. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400 (2004).
11. Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M. & Cohen, J. D. An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108 (2001).
12. Gürçay, B. & Baron, J. Challenges for the sequential two-system model of moral judgement. *Think. Reason.* 23, 49–80 (2017).
13. Mikhail, J. Universal moral grammar: Theory, evidence and the future. *Trends Cogn. Sci.* 11, 143–152 (2007).
14. Boyle, J. Medical ethics and double effect: the case of terminal sedation. *Theor. Med. Bioeth.* 25, 51–60 (2004).
15. Gross, M. L. Bioethics and armed conflict: Mapping the moral dimensions of medicine and war. *Hastings Cent. Rep.* 34, 22–30 (2004).
16. Gross, M. L. Killing civilians intentionally: double effect, reprisal, and necessity in the Middle East. *Polit. Sci. Q.* 120, 555–579 (2005).
17. Tully, P. A. The doctrine of double effect and the question of constraints on business decisions. *J. Bus. Ethics* 58, 51–63 (2005).
18. Greene, J. D. et al. Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition* 111, 364–371 (2009).
19. Cushman, F. Action, outcome, and value: A dual-system framework for morality. *Personal. Soc. Psychol. Rev.* 17, 273–292 (2013).
20. Cushman, F., Gray, K., Gaffey, A. & Mendes, W. B. Simulating murder: The aversion to harmful action. *Emotion* 12, 2 (2012).
21. Ellsworth, R. M. & Walker, R. S. Sociobiology of lethal violence in small-scale societies. in *The Routledge International Handbook of Biosocial Criminology* 85–102 (Routledge, 2014).
22. Abarbanell, L. & Hauser, M. D. Mayan morality: An exploration of permissible harms. *Cognition* 115, 207–224 (2010).
23. Barrett, H. C. et al. Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proc. Natl. Acad. Sci.* 113, 4688–4693 (2016).
24. Cushman, F., Young, L. & Greene, J. D. Our multi-system moral psychology: Towards a consensus view. in *The Oxford handbook of moral psychology* 47–71 (2010).

25. Koenigs, M. et al. Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446, 908–911 (2007).
26. Perkins, A. M. et al. A dose of ruthlessness: Interpersonal moral judgment is hardened by the anti-anxiety drug lorazepam. *J. Exp. Psychol. Gen.* 142, 612–620 (2013).
27. Szekely, R. D. & Miu, A. C. Incidental emotions in moral dilemmas: The influence of emotion regulation. *Cogn. Emot.* 29, 64–75 (2015).
28. Johnson, R. C. et al. Guilt, shame, and adjustment in three cultures. *Personal. Individ. Differ.* 8, 357–364 (1987).
29. Tracy, J. L. & Matsumoto, D. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proc. Natl. Acad. Sci.* 105, 11655–11660 (2008).
30. Scollon, C. N., Diener, E., Oishi, S. & Biswas-Diener, R. Emotions across cultures and methods. *J. Cross-Cult. Psychol.* 35, 304–326 (2004).
31. Heinrichs, N. et al. Cultural differences in perceived social norms and social anxiety. *Behav. Res. Ther.* 44, 1187–1197 (2006).
32. Gleichgerrcht, E. & Young, L. Low levels of empathic concern predict utilitarian moral judgment. *PloS One* 8, e60418 (2013).
33. Luo, S. et al. Interaction between oxytocin receptor polymorphism and interdependent culture values on human empathy. *Soc. Cogn. Affect. Neurosci.* 10, 1273–1281 (2015).
34. Cheon, B. K. et al. Cultural influences on neural basis of intergroup empathy. *NeuroImage* 57, 642–650 (2011).
35. Awad, E. et al. The Moral Machine experiment. *Nature* 563, 59–64 (2018).
36. Koenig, L. B., McGue, M., Krueger, R. F. & Bouchard Jr, T. J. Genetic and environmental influences on religiousness: Findings for retrospective and current religiousness ratings. *J. Pers.* 73, 471–488 (2005).
37. Kahane, G. On the wrong track: Process and content in moral psychology. *Mind Lang.* 27, 519–545 (2012).
38. Kahane, G., Everett, J. A., Earp, B. D., Farias, M. & Savulescu, J. ‘Utilitarian’ judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition* 134, 193–209 (2015).
39. Kahane, G. et al. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychol. Rev.* 125, 131–164 (2017).
40. Conway, P., Goldstein-Greenwood, J., Polacek, D. & Greene, J. D. Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition* 179, 241–265 (2018).
41. Hauser, M. *Moral minds: How nature designed our universal sense of right and wrong.* (Ecco/HarperCollins Publishers, 2006).
42. Hauser, M. D., Young, L. & Cushman, F. Reviving Rawls’ linguistic analogy. *Moral Psychol.* 2, 107–143 (2008).
43. Gelfand, M. J., Nishii, L. H. & Raver, J. L. On the nature and importance of cultural tightness-looseness. *J. Appl. Psychol.* 91, 1225 (2006).
44. Awad, E., Dsouza, S., Shariff, A., Rahwan, I. & Bonnefon, J.-F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc. Natl. Acad. Sci.* 117, 2332–2337 (2020).
45. Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R. & Hütter, M. Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *J. Pers. Soc. Psychol.* 113, 343 (2017).
46. Conway, P. & Gawronski, B. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *J. Pers. Soc. Psychol.* 104, 216 (2013).

47. Moshontz, H. et al. The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Adv. Methods Pract. Psychol. Sci.* 1, 501–515 (2018).
48. Bond, M. H. & van de Vijver, F. J. R. Making scientific sense of cultural differences in psychological outcomes: Unpackaging the Magnum Mysterium. in *Culture and psychology. Cross-cultural research methods in psychology* (eds. Matsumoto, D. & van de Vijver, F. J. R.) 75–100 (Cambridge University Press, 2011).
49. Costa, A. et al. Your morals depend on language. *PloS One* 9, e94842 (2014).
50. Meade, A. W. & Craig, S. B. Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455 (2012).
51. Hannikainen, I. R., Machery, E. & Cushman, F. A. Is utilitarian sacrifice becoming more morally permissible? *Cognition* 170, 95–101 (2018).
52. Dienes, Z. Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5, 781 (2014).
53. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev.* 25, 128–142 (2018).
54. Morey, R. D., Rouder, J. N. & Jamil, T. BayesFactor: Computation of Bayes factors for common designs. (2015). [Computer Software]
55. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2018). [Computer Software]
56. Dienes, Z. How do I know what my theory predicts? Accessed via: <https://psyarxiv.com/yqaj4>. *Adv. Methods Pract. Psychol. Sci.* 2, 364–377 (2019).
57. Muthukrishna, M. et al. Beyond WEIRD psychology: measuring and mapping scales of cultural and psychological distance. *Psychol. Sci.* 31. 678-701 (2018).

## 4. STATISTICAL PRACTICE

### 4.1. The role of human fallibility in psychological research: A survey of mistakes in data management<sup>43</sup>

Marton Kovacs<sup>1,2</sup>, & Rink Hoekstra<sup>3</sup>, Balazs Aczel<sup>1</sup>

<sup>1</sup>Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>2</sup>Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>3</sup>University of Groningen, Groningen, The Netherlands

---

<sup>43</sup> Published as:

Kovacs, M., Hoekstra, R., & Aczel, B. (2021). The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management. *Advances in Methods and Practices in Psychological Science*, 4(4), 25152459211045930.

**Abstract**

Errors are an inevitable consequence of human fallibility and researchers are no exception. Most researchers can recall major frustrations or serious time delays due to human errors while collecting, analyzing, or reporting data. The present study is an exploration of mistakes made during the data management process in psychological research. We surveyed 488 researchers regarding the type, frequency, seriousness, and outcome of mistakes that have occurred in their research team during the last 5 years. The majority of respondents suggested that mistakes occurred with very low or low frequency. Most respondents reported that the most frequent mistakes led to insignificant or minor consequences, such as time loss or frustration. The most serious mistakes caused insignificant or minor consequences for about a third of respondents, moderate consequences for almost half of respondents, and major or extreme consequences for about one-fifth of respondents. The most frequently reported types of mistakes were 'ambiguous naming/defining of data', 'version control error', and 'wrong data processing/analysis'. Most mistakes were reportedly due to 'poor project preparation or management' and/or 'personal difficulties' (physical or cognitive constraints). These initial exploratory findings do not aim to provide a description, representative for psychological scientists, but to lay the groundwork for a systematic investigation of human fallibility in research data management and the development of solutions to reduce errors and mitigate their impact.

**Keywords:** human error, data management mistakes, research workflow, life-cycle of the data

## Introduction

Everybody makes mistakes, and scientists are no exception. The research process is a highly complex affair involving a variety of self-taught, unsupervised, and ad-hoc manual procedures that are vulnerable to human error. Such errors include accidentally overwriting data, analysing the wrong dataset, misapplying a randomization procedure, mislabelling experimental conditions, or copying and pasting the wrong test statistics. When errors are discovered, it is common to blame the researcher, but some errors should be expected as an inevitable consequence of human fallibility (Hardwicke et al., 2014).

The field of psychology is currently immersed in a self-reflective era during which the credibility of the literature has come under serious scrutiny (Nelson et al., 2018; Vazire, 2018). Much attention in this discussion has been paid to the impact of existing methodological and statistical practices which have been identified as threats to the validity of scientific claims and the efficiency of knowledge accumulation (John et al., 2012; Simmons et al., 2011). The impact of basic human error, however, has received relatively sparse attention and existing evidence is limited to specific circumstances. For example, reviewing published studies, Rosenthal (1978) found that researcher observations of participants were occasionally miscoded. In a more recent study, Nuijten et al. (2016) performed an automated assessment of thousands of psychology articles and observed at least one statistical reporting inconsistency in half of them. Finally, Hardwicke et al. (2018) attempted to directly reproduce target values reported in 35 psychology articles by repeating the original analyses. 24 of these articles contained at least one value that could not be reproduced within a 10% margin of error. While these studies highlight the role of human error in specific circumstances, what is missing is a systematic assessment of the nature, frequency, and severity of data management mistakes in psychology. A detailed characterisation of data management mistakes may help with the identification and dissemination of solutions that are most needed to improve this aspect of psychological research.

The goal of the present survey is to start the exploration of the role of human error in the management of psychological data. Research data management is an umbrella term concerning all stages of a research project that have an effect on the data. This is the definition we use throughout the paper. These stages typically consist of many manual procedures, making them especially vulnerable to human error. We aimed to survey researchers from the field of psychology and ask them to describe and rate mistakes that they encountered in their own

research. Given the sparsity of research in this topic and the non-representativeness of our sample, our goal was explicitly exploratory and descriptive.

## Disclosures

### Preregistration

This was an exploratory study and it was not our intention to test any hypotheses. Nevertheless, we preregistered a study protocol (<https://osf.io/myu3v>) outlining our rationale, methods, and analysis plan to make clear which aspects of the study were pre-planned and which were developed during or after data collection.

In the preregistration, we proposed to group the collected mistakes into traditional data management stages, but we have used an updated version of the data management stages which we think is more nuanced (see *Figure S1*). The data preprocessing procedures and the validation of the grouping process (see Method section) were not preregistered. We are not aware of any other deviations from the preregistered protocol.

### Data, materials, and online resources

All data and materials, as well as the R code for the analyses and figures, can be accessed at the project's OSF page: <https://osf.io/fg7yb/>. A list of links to specific external materials can be found in Table 1.

Table 1

Links to All External Materials Related to the Study

Name	Link
The main survey exported from Qualtrics	<a href="https://osf.io/67dfz/">https://osf.io/67dfz/</a>
Preregistration of the primary study	<a href="https://osf.io/myu3v">https://osf.io/myu3v</a>
OSF repository of the project	<a href="https://osf.io/fg7yb/">https://osf.io/fg7yb/</a>



Definitions of the groups	<a href="https://github.com/marton-balazs-kovacs/researchers_mistake_script/tree/master/Data/Processed/grouping/definition">https://github.com/marton-balazs-kovacs/researchers_mistake_script/tree/master/Data/Processed/grouping/definition</a>
Examples of data management mistakes showed during the validation of the grouping process	<a href="https://osf.io/3sf9j/">https://osf.io/3sf9j/</a>
Instructions for the raters during validation	<a href="https://osf.io/awr6s/">https://osf.io/awr6s/</a>
Link to the preprint	<a href="https://psyarxiv.com/xcy kz/">https://psyarxiv.com/xcy kz/</a>

---

## Reporting

We report the rationale for our sample size, all data exclusions, all manipulations, and all measures conducted during the study.

## Ethical Approval

Ethical permission was provided by Eotvos Lorand University Faculty of Education and Psychology Ethical board in Hungary. We collected no identifying information from the respondents. This study was conducted in accordance with the Declaration of Helsinki.

## Method

### Sample

We contacted 16,412 corresponding authors of articles published between 2010 and 2018 in a journal having ‘psychology’ among its labels in the ScienceDirect database. Participation was voluntary and anonymous. To encourage participation, we offered to support the Center for Open Science with 0.20 USD for each completed survey. The detailed description of the email address collection method and the recruitment can be found in the Supplementary materials.

### Materials

We developed a questionnaire (summarised in Table 2) and corresponding scales (see Table 3 and 4) for the exploration of the mistakes made during the data management process in psychological research (available at: <https://osf.io/67dfz/>).

Table 2

List of Questions from the Survey about the Data Management Mistakes

Property of the Mistake	Question	Variable Type
Mistakes in general	The frequency of mistakes in general	Likert-type scale
Most frequent mistake	Description of the mistake	Free-text
	Cause of the mistake	Free-text
	Outcome of the mistake	Multiple choice with free-text option
	Frequency of the mistake	Likert-type scale
	Seriousness of the mistake	Likert-type scale
Most serious mistake	Description of the mistake	Free-text
	Cause of the mistake	Free-text
	Outcome of the mistake	Multiple choice with free-text option
	Seriousness of the mistake	Likert-type scale

In this questionnaire, we first aimed to measure how often researchers commit data management mistakes in general. Therefore, we asked them how frequently they believe any kind of data management mistake happens in their research team, responding on a 5-point Likert-type scale from ‘very low’ to ‘very high’ frequency (Table 3). Next, we asked the respondents to specify *the most frequent* mistake that has happened in their research team during the last 5 years and how frequently that mistake occurs (on the same frequency scale as above); how serious they think the outcome of that mistake was (on a 5-point Likert-type scale ranging from ‘Insignificant’ to ‘Extreme’ severity, see Table 4); the cause of that mistake (free text response); and what negative outcome occurred (select one from financial loss; erroneous conclusion; time loss; inefficiency; frustration; other, please specify).

We also asked researchers to write down *the most serious* mistake that has happened in their research team during the last 5 years, how serious they think the outcome of that mistake was (on the same seriousness scale as above), the cause of that mistake (free text response) and what negative outcome occurred (select one from financial loss; erroneous conclusion; time loss; inefficiency; frustration; other, please specify).

Finally, as background information questions, we asked respondents to specify their research field (they could choose one from the following options: social psychology; applied psychology; personality psychology; clinical psychology; developmental and educational psychology; experimental and cognitive psychology; neurophysiology and physiological psychology; methodology and statistics; or other), and the number of years they have worked in that field.

Table 3

Frequency Scale for Research Data Management Mistakes

Frequency Level	Description
Very Low	Occurs never or rarely.
Low	Occurs in some of the projects.
Moderate	Occurs in half of the projects.
High	Occurs in most of the projects.
Very High	Occurs in (almost) all of the projects.

Table 4

Seriousness Scale for Research Data Management Mistakes

Seriousness Level	Possible Consequences	Example
Insignificant	Minutes of time loss. Insignificant financial loss. No effect on conclusions.	Occasional typos in the variable names.

Minor	Some project delay and/ or money loss. Short-timed frustration. No effect on conclusions.	Have to rerun the analysis.
Moderate	Definite time and/or money loss. Mild frustration. Potential effect on some conclusions.	Have to record part of the whole sample again.
Major	Great project delay and/or money loss. Affecting some conclusions of the article. A considerable level of frustration.	Have to redo the whole data collection.
Extreme	Project failure. Serious time and/or money loss. Strongly affecting the central conclusion of the article. Damaged professional reputation.	Article withdrawal.

---

### Procedure

The participants received the Qualtrics survey link in an email (available at: <https://osf.io/67dfz/>). All questions were optional aside from the background information questions. The topic of the survey was introduced by eight brief examples of research data management mistakes (partially sourced from a pilot study, see Supplementary materials). The completion of the survey took a median of 6 minutes.

### Number of responses

Out of the 16,412 sent emails 14,033 were delivered and the remaining 2,379 bounced. All in all, 779 researchers (response rate: 5%) started our survey, out of which we excluded 19 respondents who did not accept the informed consent form, and 271 respondents who did not answer any of the questions listed in Table 2. We also excluded one respondent who did not answer any of the compulsory questions regarding their background. The survey software and personal correspondence indicated that some respondents redistributed the survey link among their colleagues. As the respondents who answered the forwarded survey also indicated their field of research and the years, they have spent on the field we decided to keep their responses (24 respondents after exclusions). Ultimately, the data of 488 respondents remained for further analysis.

## Data preprocessing

The data preprocessing pipeline was considerably different for the investigation of the frequency and seriousness ratings of the mistakes, and for the free-text responses (description of the mistakes, their causes, and their outcomes). Thus, we describe below the preprocessing of the ratings and the free-text responses separately.

### *Preprocessing of the frequency and seriousness ratings*

For the preprocessing of the frequency and seriousness ratings, we used the data from the remaining 488 respondents after the initial exclusions. The respondents had to provide a frequency rating and seriousness rating for the most frequent mistake, and a seriousness rating for the most serious mistake that they described (for the description of the rating scales see Tables 3 and 4). In some cases, the respondents described more than one mistake in their free-text response, but we included only one rating per question per respondent. Only those frequency and seriousness ratings were included in our analyses where the corresponding description of the mistake passed the following exclusion procedures. First, as describing a mistake was not compulsory, we worked with the description of 449 most frequent mistakes, and the description of 404 most serious mistakes after excluding the missing responses. Second, we excluded responses where the description of the mistake provided by the participant was ambiguous (e.g., respondent wrote *see above* and it was not clear which answer they were referring to), irrelevant to the given question, or the researcher stated that the mistake occurred before the 5-year time-frame we were interested in. After this exclusion, we were left with the descriptions of 419 most frequent mistakes and 297 of the most serious mistakes. Table 5 contains the number of mistake descriptions that we excluded in this step for each exclusion criteria. Finally, as providing a rating for the described mistakes was also not compulsory, one seriousness rating was not reported for a description of a most serious mistake. Therefore, it is missing from the analyses. At the end of the data preprocessing, we have been left with 419 frequency and seriousness ratings of the most frequent mistakes, and 296 seriousness ratings of the most serious mistakes. These ratings were provided by 426 respondents. For the overall frequency of mistakes in the team question, we had 486 responses left after excluding two missing responses.

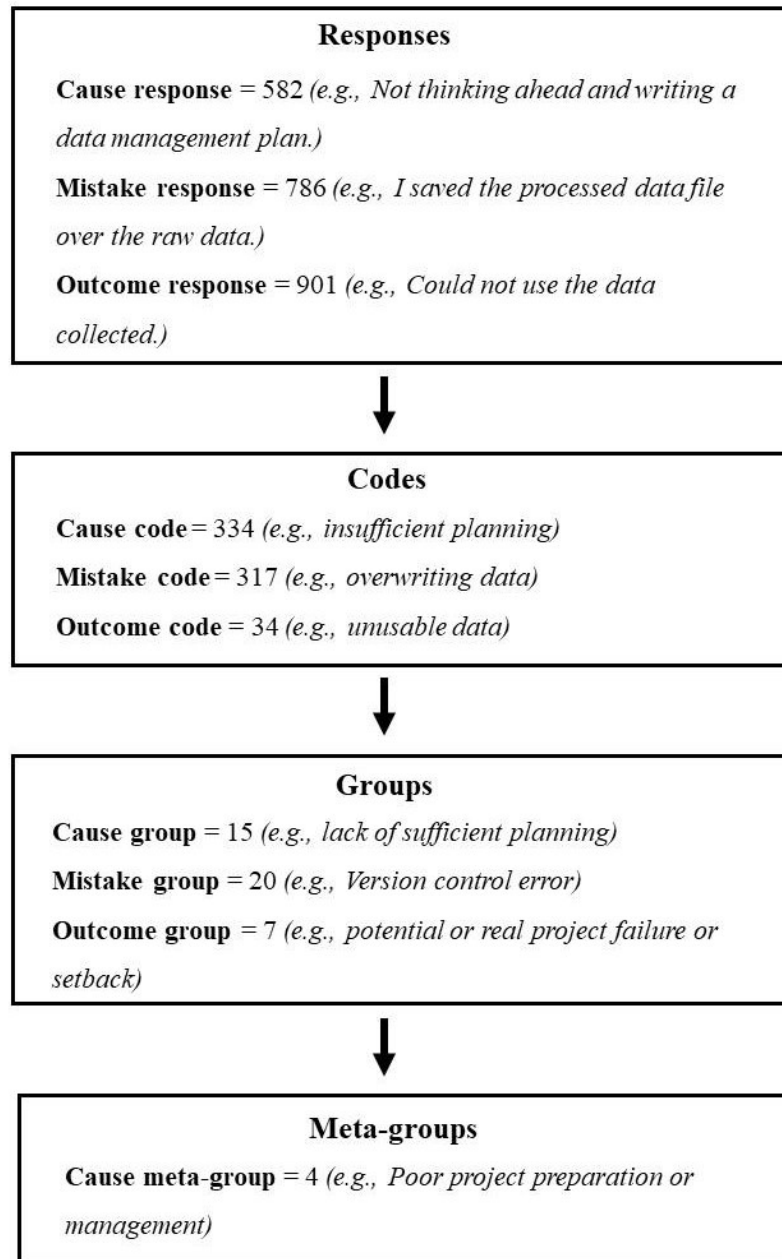
Table 5

Number of Mistakes Descriptions Excluded for Each Exclusion Criteria

Exclusion Criteria	Property of the Mistake	Number of Mistakes Excluded for the Property
Ambiguous	Most frequent	0
	Most serious	40
Irrelevant content	Most frequent	22
	Most serious	62
Out of timeframe	Most frequent	8
	Most serious	5

### *Preprocessing of the free-text responses*

To analyse the free-text responses describing the research data management mistakes, their causes, and their outcomes we categorised them into groups based on similarity by using thematic analysis (Braun & Clarke, 2006), a qualitative method that helps identify and highlight central features in texts (see Figure 1 for a summary). The grouping process was carried out by two team members (BA and MK) and all disagreements were resolved by discussion. Below, we describe creation of the groups in detail.



*Figure 1.* This flowchart illustrates the categorisation of free-text responses into groups. The number of responses indicate their counts after both the separation of the responses and the exclusions. Here, we only report the final number of items for each level of grouping. Illustrative examples are shown as italicized text in parentheses.

**Preparing data for the grouping process.** For the preprocessing of the free-text responses, we started the process with responses from 488 respondents. Respondents were asked to describe their most frequent and most serious mistakes, and their causes and outcomes in a free-text

response (see Table 2). For the outcomes of the mistakes, we provided a list of options with the possibility of writing a free-text response if none of the provided options were applicable. However, we applied the same data preprocessing method to the outcomes of the mistakes as to the descriptions of the mistakes and their causes for the sake of simplicity. The preprocessing methodology was applied separately to the descriptions of mistakes, causes, and outcomes. Answering these questions was not compulsory, therefore, there were missing responses. Moreover, as mentioned in the *Preprocessing of the frequency and seriousness ratings* section, we excluded the responses where the description of the mistake was ambiguous, irrelevant to the given question, or the researcher stated that the mistake occurred before the prescribed time-frame (i.e., past 5 years). We applied the same exclusion criteria to the descriptions of the causes and the outcomes as well. When the respondents provided more than one description of a cause, a mistake, or an outcome in their free-text response, we treated each response separately in the grouping process. Thus, after the initial exclusions and the separation of the responses, we had 931 descriptions of causes, 835 descriptions of mistakes, and 920 descriptions of outcomes. Figure 2 shows the number of responses left after each stage of the grouping process broken down by the aspects of a research data management mistake (cause of the mistake, the mistake itself, outcome of the mistake) and property of the mistake (most frequent mistake, most serious mistake). Further, we excluded additional responses as explained in the *Coding process* and the *Grouping process* sections.



	<b>Cause</b>	<b>Mistake description</b>	<b>Outcome</b>
<b>Excluding missing responses</b>	430	449	462
	<i>365</i>	<i>404</i>	<i>380</i>
<b>Excluding irrelevant, ambiguous, and out of timeframe responses</b>	418	419	457
	<i>302</i>	<i>297</i>	<i>368</i>
<b>Separation</b>	549	529	507
	<i>382</i>	<i>306</i>	<i>413</i>
<b>Exclusion in creating codes</b>	517	514	504
	<i>355</i>	<i>291</i>	<i>411</i>
<b>Exclusion in creating groups</b>	356	506	496
	<i>226</i>	<i>280</i>	<i>405</i>

*Figure 2.* The flowchart illustrates the number of responses broken down by aspects of the mistake and property (most frequent and most serious) after each preprocessing stage of the free-text responses. The number of responses for the most frequent mistakes are written in the upper row, while the number of responses for the most serious mistakes are written in italics in the lower row.

**Creating codes.** As the first step of the grouping process, we summarized each response by a short plain-text code in a systematic way. Each code highlighted a central feature of the given answer. We excluded all responses from further steps of the thematic grouping where we did not find the text to contain sufficient information regarding the given survey question. At the end of the coding process, we had 317 different codes for the descriptions of the mistakes, 334 for the causes of the mistakes, and 34 for the outcomes of the mistakes.

**Creating groups.** As the second step, we categorised the codes into higher-level groups. A group describes the essence of a collection of codes. Each time a code did not fit any of the existing groups, we created a new group based on the given code. At this stage, we excluded those responses of which the codes could not be categorised into any of the groups, as the code did not contain sufficient or relevant information. Based on the codes, we identified 20 different groups of mistake types, 15 groups of causes of mistakes, and 7 groups of outcomes of mistakes. Following this, we created a definition for each group by listing the codes that have been assigned to that group. Finally, each free-text response inherited the group label assigned to its code. At the end of the thematic grouping process, there were 786 descriptions of mistakes, 582 causes of mistakes, and 901 outcomes of mistakes assigned to groups.

**Creating meta-groups.** As the third step, we created four *meta-groups* to decrease the number of groups for the causes of mistakes to ease comprehension and aid visualization. The creation of the meta-groups was carried out through a discussion in a non-systematic way. The four meta-groups were created based on overlapping themes between the groups. Table 6 shows which cause groups were assigned to which meta-groups.

Table 6

Meta-Groups for Mistake Causes

Meta-Group	Cause Group
Poor project preparation or management	bad or lack of planning, bad or lack of standards, bad skill management, miscommunication, failure to automate an error prone task, time management issue
External difficulties	high task complexity, technical issues
Lack of knowledge	lack of knowledge/experience
Personal difficulties	carelessness, inattention, lack of control, overconfidence, physical or cognitive constraints

## Results

### Background information

Among the 488 respondents the three most commonly identified psychology fields were experimental and cognitive psychology ( $N = 88$ ), social psychology ( $N = 62$ ), clinical psychology, and developmental and educational psychology ( $N = 45$  for both), although the largest group ( $N = 116$ ) of the respondents could not associate themselves with any of the listed research fields. The median time spent in their field was 15 years ( $IQR = 15$ ). The summary of the respondents' research fields and the distribution of their years spent in their field can be found in the *Supplementary materials*.

### General overview of data management mistakes

To obtain a general overview of data management mistakes, we investigated the overall frequency of mistakes, the frequency and seriousness of the most frequent mistakes, and the seriousness of the most serious mistakes (see Table 2 for the questions). All the results for this section are shown in *Figure 3* and the text below provides a summary of the results.

**The overall frequency of mistakes.** Responses suggested that the overall occurrence of mistakes was infrequent; 79% (384 out of 486) of respondents reported that mistakes occurred with very low or low frequency whereas for 21% (102) of the remaining respondents, mistakes had moderate, high, or very high frequency.

**The most frequent mistakes.** When researchers were asked how frequently the most frequent mistake happened in their research team, 75% (314 out of 419) of them indicated that it had low or very low frequency, while for the remaining 25% (105) of the teams the most frequent mistake had moderate, high, or very high frequency.

The most frequent mistakes reportedly led to insignificant or minor consequences (e.g., minutes of time loss; insignificant financial loss; no effect on conclusions) for 69% (289 out of 419) of respondents, moderate consequences for 25% (104) of respondents, and major or extreme consequences for the remaining 6% (26) of respondents.

**The most serious mistakes.** When asked about the most serious data management mistake that occurred in their team during the last five years, 31% (93 out of 296) of respondents reported that the mistake led to insignificant or minor consequences (e.g., minutes of time loss; insignificant financial loss; no effect on conclusions), 46% (137) reported that the mistake led

to moderate consequences, and the remaining 22% (66) reported that the mistake led to major or extreme consequences.

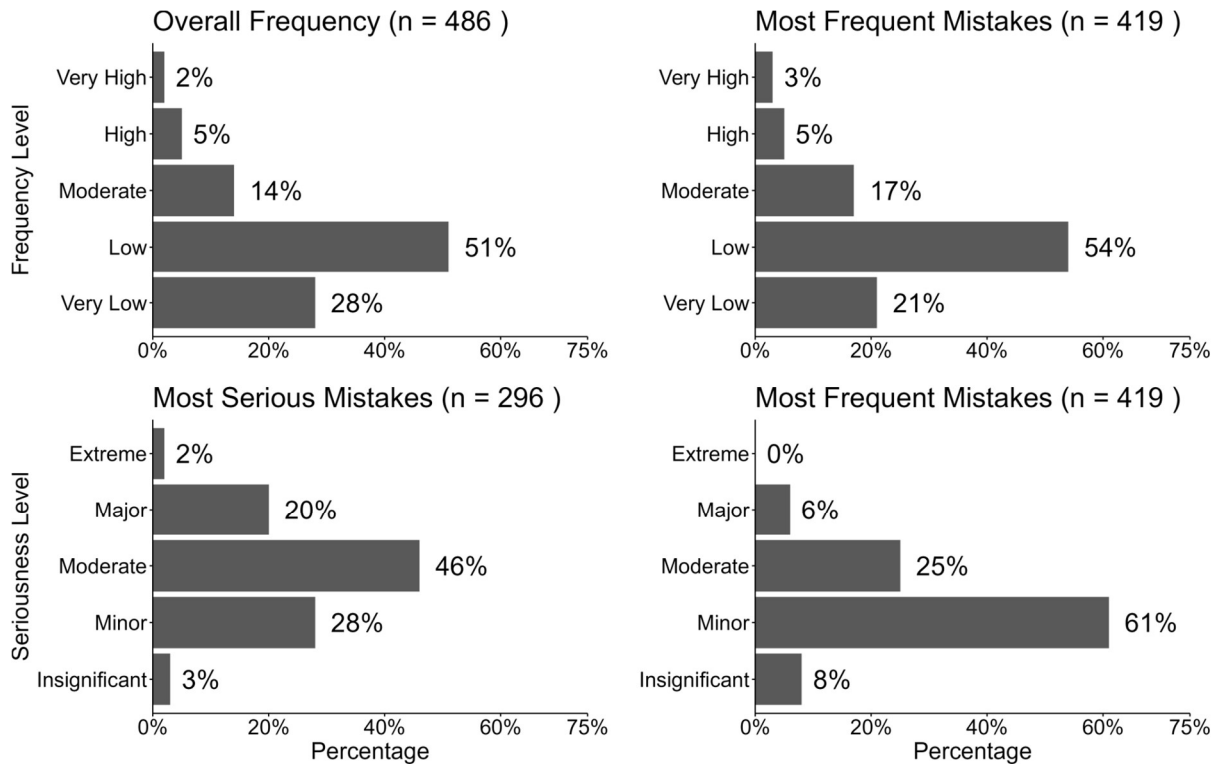


Figure 3. Distribution of all responses presented in the General overview of data management mistakes section. Each plot shows the percentages on the X axis, while the levels of either the frequency scale (see Table 2) or the seriousness scale (see Table 3) are shown on the Y axis. Percentages may not sum to 100 due to rounding. The counts behind these percentages are reported in Figure S4.

**Data management mistake types, causes, and outcomes**

**Frequency of data management mistake types.** Through the grouping process, we sorted the 786 descriptions of the most frequent (N = 506) and most serious (N = 280) data management mistakes into 20 different mistake types. To determine which type of mistakes are the most frequent, in our sample we counted how many times a mistake type was reported by respondents. For this analysis we kept multiple responses provided by single respondents. Table 7 shows how many times a mistake type reportedly occurred for the most frequent and most serious mistakes. The three most frequently reported mistake types for the most frequent mistakes were ‘ambiguous naming/defining of data’ (86 out of 506), ‘version control error’ (62), and ‘wrong data processing/analysis’ (47). The three most frequently reported mistake

types for the most serious mistakes were ‘wrong data processing/analysis’ (32 out of 280), ‘data coding error’ (26), ‘loss of materials/documentation/data’ (26).

Table 7

Data Management Mistake Type Groups and the Number of Their Occurrences

Mistake Type Group	Most Frequent Mistakes (n = 506)	Most Serious Mistakes (n = 280)
Ambiguous naming/defining of data	86	16
Version control error	62	20
Wrong data processing/analysis	47	32
Data coding error	45	26
Data input error	37	21
Loss of materials/documentation/data	30	26
Programming error	35	19
Data transfer error	41	8
Data selection/merging error	21	17
Technical/infrastructural problem	17	21
Oversight in study design or measurement	14	18
Bad or poor documentation	21	7
Participant allocation error	15	8
Bad connection of data points	8	8
Wrong reporting	7	8
Wrong software or hardware settings	5	10

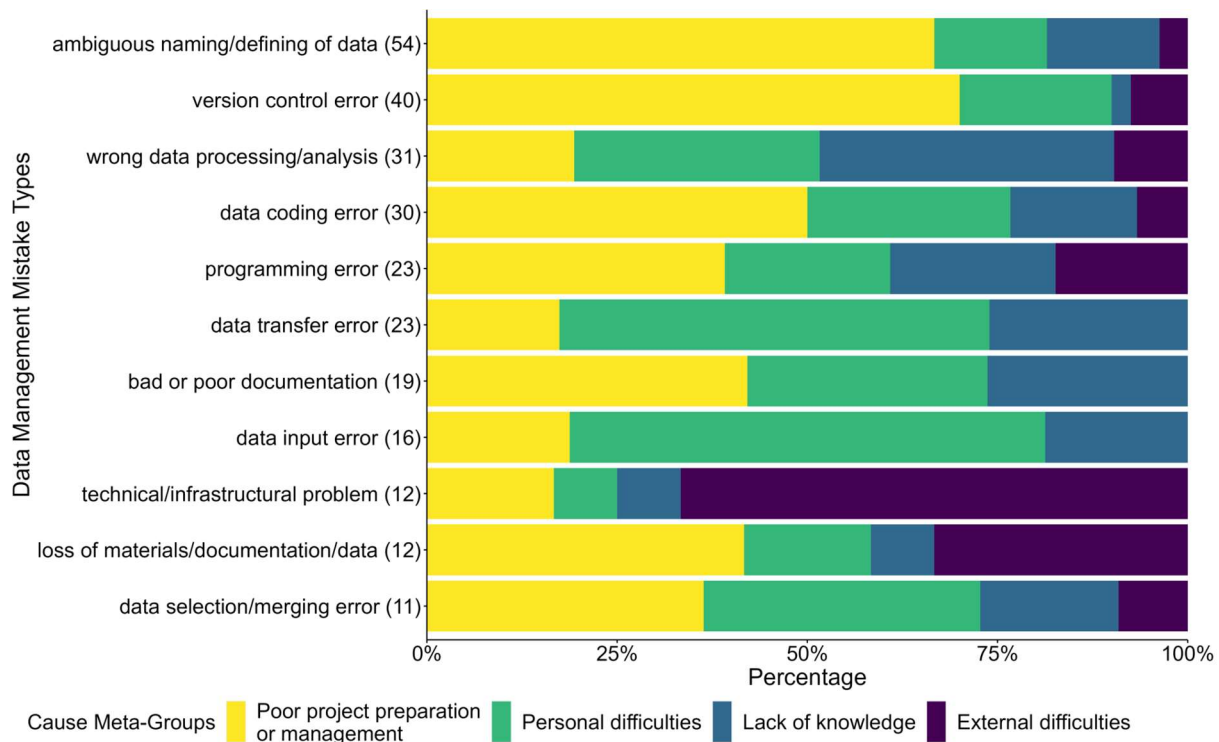
Data or file organisation error	6	4
Deviation from the protocol	5	4
Violation in ethics	2	4
Project management error	2	3

---

Note: The mistake types are arranged in a descending order according to the number of times they were reported for the most frequent and most serious mistakes together.

**Mistake types and their reported causes.** *Figure 4* shows the data management mistake types for the most frequent mistakes, and the proportions of the meta-level grouping for their reported causes. The relationship between the mistake types and causes can also be viewed separately for the most serious mistakes (see *Figure S5*) in the Supplementary materials. Cases were omitted from the analyses where the respondent described more than one mistake and more than one cause was associated with them, since here the mistake and its cause could not be unambiguously connected. In case of a one to many mapping, we assumed that the respondent wished to report several causes that led to a mistake or one cause that led to several mistakes.

The most common causes assumed by the researchers to be responsible for these most frequent mistake types were ‘poor project preparation or management’ (43%) and ‘personal difficulties’ (29%). For the most serious mistake types the most common causes were the same with 39% for the ‘poor project preparation or management’ and 37% for the ‘personal difficulties’.



*Figure 4.* The frequency of the data management mistake types for the most frequent mistakes and the proportions for the meta-level grouping (see Table 6) of their reported causes. The mistake types are presented in decreasing order from the top to the bottom by the number of research teams who reported the specific mistake type. Mistake types with fewer than 10 occurrences are not displayed. The numbers in parentheses represent the number of times a given mistake type was reported after cases with multiple mistakes/causes were omitted.

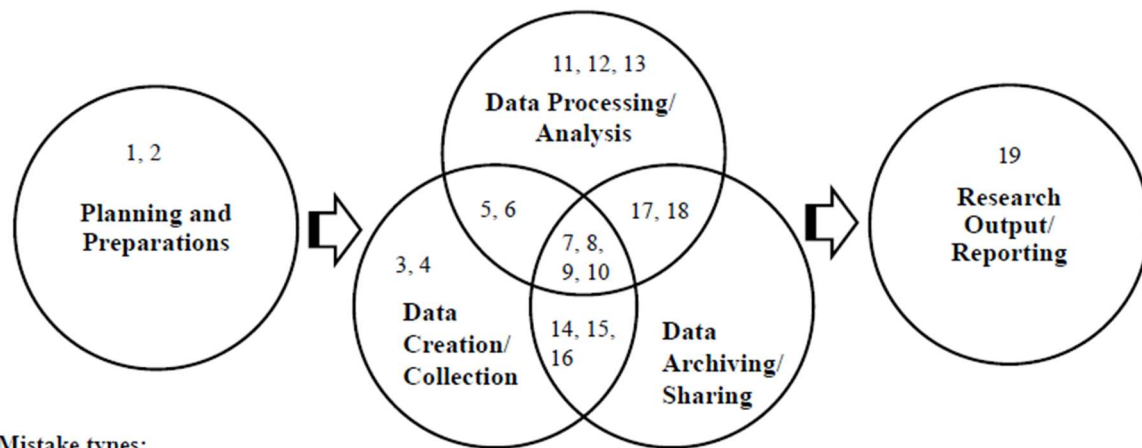
**Mistake types and their reported outcomes.** *Figure 5* shows the frequency of the data management mistake types for the most frequent mistakes and the proportions of their reported outcomes. The relationship between the mistake types and reported outcomes can also be viewed separately for the most serious (see *Figure S6*) mistakes in the Supplementary materials. Cases where the respondent described more than one mistake and reported more than one negative outcome associated with those were omitted from this analysis. The most commonly reported outcomes that we could clearly associate with the mistake types for the most frequent mistakes were ‘time loss’ (67%) and ‘frustration’ (21%) respectively. The most common outcomes associated with the most serious mistakes were the same with 46% for ‘time loss’ and 26% for ‘frustration’.



*Figure 5.* The frequency of data management mistake types for the most frequent mistakes and the proportions of the reported outcomes. The mistake types are presented in decreasing order from the top to the bottom by the number of research teams who reported the specific mistake type. Mistake types with fewer than 10 occurrences are not displayed. Numbers in parentheses represent the number of times a respondent reported the given mistake type.

**Mistakes Types and Data Management Stages.** We categorized each mistake type according to the data management stage (or overlap of multiple stages) during which it was likely to have occurred (*Figure 6*). Most types of mistakes belong to the overlap of the Data Processing/Analysis, Data Creation/Collection, and the Data Archiving/Sharing sections. The data management model used for the present categorization was developed by the authors. See the Supplementary materials for a more detailed description of development.





#### Mistake types:

1 oversight in study design or measurement	7 ambiguous naming/defining of data	14 bad connection of data points
2 project management error	8 data or file organisation error	15 bad or poor documentation
3 data input error	9 deviation from the protocol	16 violation in ethics
4 participant allocation error	10 programming error	17 loss of materials/documentation/data
5 data coding error	11 data selection/merging error	18 version control error
6 wrong software or hardware settings	12 data transfer error	19 wrong reporting
	13 wrong data processing/analysis	20 technical/infrastructural problem

*Figure 6.* Mistake types categorized by research data management stage. The numbers indicate the mistake types (see <https://osf.io/76d24/>). Mistake 20 (technical or infrastructure problems) is not part of any stage as it is an external factor but can have an effect on the efficiency of the data management pipeline.

## Discussion

The results of this survey showed that data management mistakes are ubiquitous in many labs conducting psychological research. Most respondents believed that data management mistakes occur infrequently in their own research, one-fifth of them observed those in moderate, high, or very high frequency. The most serious mistakes only had minor consequences for the third of the research teams in our sample, whereas for one-fifth of them they came with major or extreme repercussions (such as project failure or erroneous conclusions). Naturally, this survey was not capable of detecting undiscovered or unreported mistakes and, therefore, it is plausible that our numbers underestimate the actual frequency of data management mistakes. These exploratory findings do not aim to provide exact estimates but to help identify some common data management mistakes and potential causes and outcomes, which may facilitate the education about existing solutions, and the development of novel mistake mitigation strategies.

Respondents reported a variety of mistakes occurring across the research data management pipeline. Deciding which mistakes are of highest priority to address will require consideration of their frequency and seriousness, as well as the potential resources needed in order to address them. The majority of respondents reported that the most frequent mistakes, involving

‘ambiguous naming or defining of data’, ‘version control error’, and ‘wrong data processing/analysis’, can be associated with the ‘data processing and analysis’ stage. These mistakes were mostly assumed to be the result of ‘poor project planning or management’. Most frequently, the cost of these mistakes is ‘time loss’ and ‘frustration’. We assume that ‘erroneous conclusions’ are less frequent outcomes of these mistakes only because the reporters have discovered the mistakes before publicizing their results. Hence, the proportion of conclusions that remain defective in the literature due to data management mistakes is dependent on the efficiency of the existing checking procedures.

Most mistake types were categorized to more than a single stage as they can happen at several points of the data management pipeline. The mistakes that were typical of most stages were found to be ‘ambiguous naming/defining of data’, ‘data or file organization error’, ‘deviation from the protocol’, and ‘programming error’.

A number of generic solutions and guidelines have been proposed to assist researchers within their data management. Based on personal experience, Rouder et al. (2019) described five principles to minimize and mitigate research mistakes: (1) a lab culture focused on learning from mistakes; (2) implementing computer automation; (3) standardization; (4) coded analysis; and (5) elaborate manuscripts. Others have pointed towards the need for formal training in data management (Barone et al., 2017; Tenopir et al., 2016). Importantly, an increasing number of university library services provide dedicated support for data management plans (Michener, 2015). Data librarians are specialized in providing support in managing research data (Semeler et al., 2019). Various guidelines and checklists have been developed to help researchers adopt transparent research workflows (Aczel, et al., 2020; Klein et al., 2018), comprehensive reporting (e.g., see <https://www.equator-network.org/>), reusability of data holdings (Wilkinson et al., 2016), as well as ethical and efficient research management (Bareille et al., 2017; Giesen, 2015). Dedicated software tools (e.g., R Markdown; Baumer & Udwin, 2015) are available to make data management more efficient, transparent, and less error-prone. In Table 8 we have presented a non-comprehensive collection of existing error-mitigation tools or strategies corresponding to a number of our mistake types. It is to be mentioned that the cause of the mistake can play an important role in the efficiency of the error-mitigation strategies. For example, if a person makes mistakes in data management not because of the lack of knowledge but because of some personal difficulties then the potential solution will require more than mistake-specialised strategies.

Table 8

Existing Error Mitigation Strategies for the Most Frequent and/or Serious Data Management Mistakes.

Mistake Type	Existing Error-Mitigation Strategy
Ambiguous naming or defining of data	Using naming standards (e.g., Gorgolewski et al., 2016) Using codebooks (Arslan, 2019; Johnson et al., n.d.) Creating data management plans (Michener, 2015)
Version control error	Using a version control system such as Git (Blischak et al., 2016)
Wrong processing/analysis data	Co-piloting (Veldkamp et al., 2014) Creating data management plans (Michener, 2015) Using statistical code language (Python, R)
Loss of materials/documentation/data	Using a version control system such as Git (Blischak et al., 2016) Sharing information in online repositories (Klein et al., 2018)
Programming error	Co-piloting (Veldkamp et al., 2014) Use software tests and code commenting (Michener, 2015)
Oversight in study design and measurement	Clear project structuring (Rybicki, 2019) Registered report format (Chambers, 2013)
Poor documentation	Transparent research workflow (Klein et al., 2018) Pre-registration (Nosek et al., 2019)
Data or file organisation error	Using data specification standards (e.g., Gorgolewski et al., 2016) or file organisation standards (e.g., <i>The DRESS Protocol</i> , n.d.)

This survey was intended to be exploratory and descriptive and several caveats and limitations should be considered when interpreting the results. Firstly, because the survey relied on researchers' self-report, the study will not have detected mistakes that were undiscovered,

forgotten, or otherwise unreported. The findings may, therefore, highlight the existence of some pertinent data management mistakes, and perhaps their relative frequency, but should not be interpreted as reliable estimates of mistake prevalence in psychological science. Secondly, although the number of researchers responding to the survey and passing our exclusions is adequate (488) for exploratory purposes, the overall response rate (before exclusions) was very low (5%), suggesting that the findings are potentially strongly affected by self-selection bias. The overall direction of influence of such bias is difficult to predict as potential differences between respondents and non-respondents are non-trivial (e.g., those who have made more mistakes may have been more likely to take part in the survey as it was more relevant to them or less likely to take part because reporting mistakes may have felt more embarrassing for them). Thirdly, we gained only limited knowledge about the background of the respondents as many could not assign themselves to any of the psychological subfields offered in the survey and chose instead the 'other' category. Finally, the survey yielded a large quantity of partly qualitative data and it was necessary to rely on our own subjective assessment in order to generate a meaningful summary. We attempted to improve objectivity by having at least two team members dual code all responses; but some subjectivity was required, nonetheless.

Psychological science is currently undergoing a period of heightened concern about the credibility and validity of its research practices and results (Vazire, 2018). Meta-research efforts have focused on documenting major threats to credibility, such as fraud, questionable research practices, and low transparency (Hardwicke et al., 2019), but have paid relatively sparse attention to the role of basic human error. The present study has highlighted some pertinent mistakes that can percolate into the research pipeline, reducing efficiency and potentially undermining the validity of scientific claims. Future work may look to build on these findings and develop a systematic exploration of human fallibility in research data management. Repeating our methodology on a representative sample could provide valuable information in this regard and identify the weaknesses of research efficiency. We suggest three major research questions for the continuation of this endeavor: (1) what practices do researchers use to improve efficiency and quality control in data management; (2) what prevents researchers from using existing solutions; and (3) what is needed to increase adoption of these solutions.

### **Conflicts of Interest**

The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

**Author Contribution**

**Conceptualization:** Marton Kovacs, Rink Hoekstra, and Balazs Aczel.

**Data curation:** Marton Kovacs and Balazs Aczel.

Formal analysis: Marton Kovacs.

**Investigation:** Marton Kovacs and Balazs Aczel.

**Methodology:** Marton Kovacs, Rink Hoekstra, and Balazs Aczel.

Project administration: Balazs Aczel.

**Resources:** Marton Kovacs and Balazs Aczel.

**Supervision:** Balazs Aczel.

**Validation:** Marton Kovacs, Rink Hoekstra, and Balazs Aczel.

**Visualization:** Marton Kovacs and Balazs Aczel.

**Writing - original draft:** Marton Kovacs and Balazs Aczel.

**Writing - review & editing:** Marton Kovacs, Rink Hoekstra, and Balazs Aczel.

**Acknowledgments**

We are grateful to Tom Harwicke for his contribution to the conceptualization and revision of this study. We say thanks to Andrei Tamas Foldes for providing the database of email addresses for data collection, just as to Marjan Bakker, Patrick Forscher, Michele Nuijten, and Simine Vazire for their thoughts and comments on the present research. We are also thankful to Beata Bothe, Zoltan Kekecs, Tamas Nagy, Bence Palfi, Istvan Toth-Kiraly, Janos Salamon, Barnabas Szaszi, and Aba Szollosi for giving us feedback on an earlier version of the survey. We are thankful for their help with the validation of the grouping process to Bence Bakos, Patricia David, Nandor Hajdu, Emma Kis, Gabor Makovics, Eszter Molnar, Peter Szecsi, Orsi Szoke, Attila Szuts, Boglarka Zach, and Dorina Zelena.

**Prior Versions**

A preprint of the article was posted prior to publication: <https://psyarxiv.com/xcykz/>.

## References

- Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... & Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4(1), 4-6.
- Arslan, R. C. (2019). How to Automatically Document Data With the codebook Package to Facilitate Data Reuse. *Advances in Methods and Practices in Psychological Science*, 2(2), 169–187. <https://doi.org/10.1177/2515245919838783>
- Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLOS Computational Biology*, 13(10), e1005755. <https://doi.org/10.1371/journal.pcbi.1005755>
- Baumer, B., & Udwin, D. (2015). R Markdown. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3), 167–177. <https://doi.org/10.1002/wics.1348>
- Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A quick introduction to version control with Git and GitHub. *PLoS Computational Biology*, 12(1), e1004668.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, 49(3), 609–610. <http://dx.doi.org/10.1016/j.cortex.2012.12.016>
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., ... Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1), 1–9. <https://doi.org/10.1038/sdata.2016.44>
- Hardwicke, T. E., Jameel, L., Jones, M., Walczak, E. J., & Weinberg, L. M. (2014). Only human: Scientists, systems, and suspect statistics. *Opticon* 1826, 16(25), 1–12.
- Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2019). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*. <https://doi.org/10.1146/annurev-statistics-031219-041104>

- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Johnson, H. R., Stash, H., Papadatou-Pastou, M., Isager, P. M., Carlsson, R., & Aczel, B. (n.d.). *Getting Started Creating Data Dictionaries: How to Create a Shareable Dataset* Erin M. Buchanan 12abc Sarah E. Crain 1abc Arielle Cunningham 1abc.
- Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., Ijzerman, H., Nilsson, G., Vanpaemel, W., & Frank, M. C. (2018). A Practical Guide for Transparency in Psychological Science. *Collabra: Psychology, 4*(1), 20. <https://doi.org/10.1525/collabra.158>
- Michener, W. K. (2015). Ten simple rules for creating a good data management plan. *PLoS Computational Biology, 11*(10), e1004525.
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., Veer, A. E. van 't, & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences, 23*(10), 815–818. <https://doi.org/10.1016/j.tics.2019.07.009>
- Rosenthal, R. (1978). How often are our numbers wrong? *American Psychologist, 33*(11), 1005–1008.
- Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing Mistakes in Psychological Science. *Advances in Methods and Practices in Psychological Science, 2*(1), 3–11. <https://doi.org/10.1177/2515245918801915>
- Rybicki, J. (2019). Best Practices in Structuring Data Science Projects. In Z. Wilimowska, L. Borzemski, & J. Świątek (Eds.), *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018* (pp. 348–357). Springer International Publishing.
- Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science, 51*(3), 771–780. <https://doi.org/10.1177/0961000617742465>

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*, 0956797611417632.

Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., Frame, M., & Baird, L. (2016). Data management education from the perspective of science educators. *International Journal of Digital Curation*. <https://doi.org/10.2218/ijdc.v11i1.389>

*The DRESS Protocol*. (n.d.). Retrieved February 24, 2020, from <https://www.projecttier.org/tier-protocol/dress-protocol/>

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4), 411–417.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS One*, 9(12), e114876.



## 4.2. Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation<sup>44</sup>

Aczel, B.<sup>1a</sup>, Palfi, B.<sup>2,3</sup>, Szollosi, A.<sup>4</sup>, Kovacs, M.<sup>1</sup>, Szaszi, B.<sup>1,5</sup>, Szecsi, P.<sup>1</sup>, Zrubka, M.<sup>1</sup>, Gronau, Q. F.<sup>6</sup>, van den Bergh, D.<sup>6</sup>, & Wagenmakers, E.-J.<sup>6</sup>

<sup>1</sup>Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

<sup>2</sup>School of Psychology, University of Sussex, Brighton, UK

<sup>3</sup>Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK

<sup>4</sup>School of Psychology, University of New South Wales, Sydney, Australia

<sup>5</sup>Doctoral School of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary

<sup>6</sup>Department of Psychology, University of Amsterdam, PO Box 15906, 1001 NK Amsterdam, The Netherlands

### Abstract

In the traditional statistical framework, nonsignificant results leave researchers in a state of suspended disbelief. This study examines, empirically, the treatment and evidential impact of nonsignificant results. Our specific goals were twofold: to explore how psychologists interpret and communicate nonsignificant results, and to assess how much these results constitute evidence in favor of the null hypothesis. Firstly, we examined all nonsignificant findings mentioned in the abstracts of the 2015 volume of *Psychonomic Bulletin & Review*, *Journal of Experimental Psychology: General*, and *Psychological Science* ( $N = 137$ ). In 72% of cases, nonsignificant results were misinterpreted, in the sense that authors inferred that the effect was absent. Secondly, a Bayes factor reanalysis revealed that fewer than 5% of the nonsignificant findings provided strong evidence (i.e.,  $BF_{01} > 10$ ) in favor of the null hypothesis compared to the alternative hypothesis. We recommend that researchers expand their statistical toolkit in order to correctly interpret nonsignificant results and to be able to evaluate the evidence for and against the null hypothesis.

Keywords: nonsignificant results, NHST, Bayes factor analysis

---

<sup>44</sup> Published as: Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ... & Wagenmakers, E. J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, 1(3), 357-366.

“Never use the unfortunate expression ‘accept the null hypothesis’.”

Wilkinson & the Task Force on Statistical Inference, 1999, p. 599

The interpretation of statistically nonsignificant findings is a vexing point of traditional psychological research.<sup>45</sup> Within the framework of null hypothesis significance testing (NHST; Fisher, 1925; Neyman & Pearson, 1933), decisions about the null hypothesis are based on the  $p$ -value. Under NHST logic, we are entitled to reject the null hypothesis whenever our  $p$ -value is smaller than or equal to a predefined  $\alpha$  threshold (mostly set at .05; but see Benjamin et al. 2017). In contrast, the  $p$ -value does not entitle us to claim support in favor of the null hypothesis. According to the common interpretation, any  $p$ -value higher than  $\alpha$  indicates that we have to withhold judgment about the null hypothesis (Cohen, 1994). This asymmetric characteristic of the NHST framework frustrates the interpretation and communication of nonsignificant results (Edwards, Lindman, & Savage, 1963; Nickerson, 2000). It is known that  $p > .05$  results are subject to misinterpretation among researchers (Goodman, 2008), but the magnitude of this bias on the communication of psychological findings remains unexplored. Here we examine the degree to which nonsignificant findings are miscommunicated in current psychological publications; in addition, we use Bayes factors to assess how much these findings support the null hypothesis relative to a composite alternative hypothesis (e.g., Etz & Vandekerckhove, 2017).

Nonsignificant findings in psychological research are both disliked and misinterpreted, and this brings dire consequences. Firstly, the common aversion to nonsignificant findings (e.g., Ferguson, 2012; Greenwald, 1975) not only causes publication bias (e.g., Franco, Malhotra, & Simonovits, 2014) but also harms the validity of the reported outcomes. For example, most questionable research practices aim to transform otherwise nonsignificant  $p$ -values into significant  $p$ -values (e.g., Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016; Lilienfeld & Waldman, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015; Pritchet, Powell, & Horne, 2016). Secondly, nonsignificant findings are commonly misinterpreted, usually because researchers regard nonsignificant  $p$ -values as support in favor of the null hypothesis (i.e., misconception #2 in Goodman, 2008). However,  $p$ -values larger than our threshold indicate only that the test was incapable of rejecting the null hypothesis; this

---

<sup>45</sup> Throughout the paper, whenever we use the expression ‘(non)significant results’, we refer to statistically and not theoretically (non)significant results.

could have occurred because the effect does not exist, but also because the power of the test was insufficient to detect a true effect (Dienes, 2014, 2016). Indeed, an examination of the psychology literature suggests that a high proportion of statistically nonsignificant results are false negatives (Hartgerink, Wicherts, & Assen, 2017).

Finally, when confronted with nonsignificant findings, researchers may seek refuge in a description of the sample rather than inference concerning the population; such a tendency is revealed by expressions such as “no difference was observed between the groups”. Such statements about the sample are problematic, as for continuous data the observed difference is never exactly zero, even when the null hypothesis holds exactly. The question that bears scientific interest always concerns the extent to which observed effects generalize to the population. One could argue that sometimes the authors do not mean literally what they write in these cases and that expert readers can reach the proper interpretation. Nevertheless, these expressions represent a type of miscommunication that can create ambiguity for experts and misunderstanding for lay readers. Despite much recent discourse of methodological challenges in the empirical sciences (e.g., Munafò et al., 2017), the ways in which nonsignificant findings are discussed and interpreted have remained relatively unexplored. One previous study (Hoekstra et al., 2006) explored whether the recommendations of the fifth edition of the APA Publication Manual (APA, 2001) improved the way authors report and interpret the results of significance testing. They found that that both before and after the publication of the new guideline nonsignificant effects were interpreted as claims of no effect in 60% of cases.

In this observational study, we investigated the prevalence of the different interpretations of nonsignificant findings. We also explored the evidential value of these results with a Bayes factor analysis (e.g., Jeffreys, 1961; Kass & Raftery, 1995). Unlike NHST, Bayes factors indicate how much the data favor one hypothesis over another (Dienes, 2008). Therefore, where the necessary information was available, we computed Bayes factors for all reported nonsignificant *t*-test results. This allows us to explore the degree to which reported nonsignificant results actually provide support for the null hypothesis.

## Method

The data analysis methodology was preregistered online by uploading a plan to the Open Science Framework (OSF) prior to conducting the analyses. The preregistration and the collected data are available on the OSF at <https://osf.io/f2n7c>. The statistical analyses of the link between Bayes factors and *p*-values, and Bayes factors and sample sizes were not specified

in the preregistration. Further minor deviations from this plan are described in the Supplementary Materials. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

### Sample

We selected the abstracts of every empirical research article with human participants published in 2015 in the journals *Psychological Science* ( $n = 150$ ), *Psychonomic Bulletin & Review* ( $n = 167$ ), and *Journal of Experimental Psychology: General* ( $n = 95$ ; overall  $N = 412$ ). All three are prominent journals that cover broad areas of psychological research. From this collection, we selected the articles that contained at least one negative empirical statement in their abstracts. By negative statement, we mean that the authors explicitly stated the absence of an effect (e.g., “had no effect”, “were the same”) or they referred to a nonsignificant finding (e.g., “was not significant”).

For those articles that contained negative statements in their abstracts, we screened the main texts and supplements to collect: (1) the corresponding  $p$ -values; (2) the type of the statistical analysis; (3) the corresponding sentence from the abstract; and (4) the sentence describing the results of the analysis. Additionally, when the claim was based on a  $t$ -statistic (one-sample, paired-sample, or independent-sample  $t$ -tests), we collected the  $t$ -value and the number of participants in each experimental group.<sup>46</sup>

### Screening procedure

The data-collection procedure was the following: one author screened the selected abstracts and judged whether they contained a negative statement. If an abstract contained such a statement, the author extracted the additional data from the article. Each collected article was then re-examined by another author to ensure that the statement was based on the selected statistical test. Next, two authors independently categorized each of the extracted claims from the abstracts using three categories and two sub-categories (see Table 1 for hypothetical examples):

---

<sup>46</sup> When the data were not available, we requested the data from the corresponding author via e-mail. When group sample sizes were not provided for independent-sample  $t$ -tests, we took the half of the total number of participants for the sample size of each group. When the exact  $p$ -value of a  $t$ -test was not reported, we calculated it from the  $t$  and  $df$ , if these values were available.

(1) *Correct-frequentist*: the statement refers only to the fact that the analysis did not yield a significant result and it does not imply that the effect is absent in the population; (2) *Incorrect-frequentist*: the statement indicates that the authors inferred to the absence of an effect from a non-significant result. We differentiated two subcategories here: (a) the statement is generalized to the whole population; or (b) the statement is restricted to the current sample; (3) *Bayesian analysis*: for their statement the authors used Bayes factors to quantify evidence in favor of the null hypothesis. In case of a disagreement at any point during the categorization, the agreement of at least three authors was needed to reach conclusion in that particular case. These disagreements were resolved by discussion.

Table 1

*Hypothetical Examples for the Categories of Claims Concerning Nonsignificant Findings*

Category	Example
(1) Correct-frequentist	“The analysis did not show a significant effect of the intervention.”
(2a) Incorrect-frequentist – whole population	“The results establish that the intervention has no effect on the dependent variable.”
(2b) Incorrect-frequentist – current sample	“There was no difference between the participants in the intervention group and the control group.”
(3) Bayesian-analysis	“The Bayes factor favored the null hypothesis over the alternative hypothesis.”

### Bayes factor calculation

To gauge the strength of evidence for the null hypothesis we calculated Bayes factors, that is, the likelihood of the data under the null hypothesis (i.e., equal population means) divided by the likelihood of the data under the alternative hypothesis. Bayes factors greater than 1 indicate relative evidence for the null, whereas Bayes factors smaller than 1 indicate relative evidence for the alternative hypothesis. In order to aid the interpretation of the Bayes factors, we employed the Jeffreys (1961) classification scheme (see also Lee & Wagenmakers, 2013): Bayes factors between 1/3 and 3 are labeled *anecdotal evidence*, Bayes factors between 3 and 10 (or between 1/3 and 1/10) indicate *moderate evidence*, and Bayes factors greater than 10 or smaller than 1/10 indicate *strong evidence*.

We restricted the Bayes factor calculation to the case of the  $t$ -test. To obtain the Bayes factors that correspond to the reported  $t$ -statistics and degrees of freedom, we applied the default settings of the *ttest.tstat* function of the BayesFactor R package (Morey, Rouder, & Jamil, 2015). The default settings specify the alternative hypotheses by assigning effect size a two-tailed Cauchy distribution with medium scale (i.e.,  $r = \sqrt{2}/2$ ). This so-called default JZS prior

### **Bayesian parameter estimation**

In order to explore how Bayes factors and  $p$ -values, and Bayes factors and sample sizes related to one another, we decided to conduct exploratory analyses (neither these hypotheses nor the details of the analyses were preregistered). For these two correlation analyses, we decided to conduct Bayesian parameter estimation instead of hypothesis testing. Therefore, we reported the correlation coefficients (Kendall's  $\tau$ ) with the 95% credible intervals (CI). The investigated associations were non-linear, thus, we opted to compute Kendall's  $\tau$  to estimate the population effect sizes (e.g., Kendall and Gibbons, 1990). To calculate Kendall's  $\tau$ , we used the *KendallTauB* function from the DescTools R package (Signorell, 2017). We passed on the  $\tau$ -value and the sample size to compute the 95% CIs with the *credibleIntervalKendallTau* function created by van Doorn, Ly, Marsman and Wagenmakers (2016). We employed the two-sided default prior distribution of  $\tau$ , which is a non-uniform distribution on  $\tau$  constructed from a uniform distribution on the Pearson's  $\rho$  (parametric yoking; van Doorn et al., 2016).

(Rouder, Speckman, Sun, Morey, & Iverson, 2009) constitutes one of several proposed methods to specify the predictions of the alternative hypothesis. As detailed below, we repeated our Bayes factor reanalysis using two alternative prior distributions in order to explore the robustness of the results.

## **Results**

### **Preregistered analyses**

#### ***Screening***

From the 412 screened abstracts, we found at least one negative statement in 132 abstracts (*Psychological Science*,  $n = 39$ ; *Psychonomic Bulletin & Review*,  $n = 58$ ; *Journal of Experimental Psychology: General*,  $n = 35$ ). The 132 abstracts contained 137 negative

statements (*Psychological Science*,  $n = 39$ ; *Psychonomic Bulletin & Review*,  $n = 61$ ; *Journal of Experimental Psychology: General*,  $n = 37$ ). We linked these statements to 175 statistical tests from the articles and we collected 122 reported  $p$ -values from these tests (*Psychological Science*,  $n = 26$ ; *Psychonomic Bulletin & Review*,  $n = 46$ ; *Journal of Experimental Psychology: General*,  $n = 50$ ). The number of reported  $p$ -values is substantially less than the number of tests since the tests contained some non-frequentist statistics (e.g., Bayes factors) and in several cases, the  $p$ -value was not reported (e.g., non-significant regression slopes or ANOVAs) and could not be retrieved from the authors.<sup>47</sup>

### ***Categories of statements***

We found that 72% ( $n = 98$ ) of the negative statements misinterpreted the nonsignificant result, as 23% ( $n = 32$ ) fell in the “Incorrect-frequentist – whole population” (Category 2a) and 48% ( $n = 66$ ) in the “Incorrect-frequentist – current sample” category (Category 2b). Only 18% ( $n = 25$ ) were categorized as correct frequentist reporting (Category 1). With 10% ( $n = 14$ ), the application of Bayes factors (Category 3) was the least common category. Table 2 indicates the frequencies of the different categories of negative claims broken down by journal.

Table 2

#### *Frequencies of the Negative Statements Broken Down by Category and Journal*

Category	Psych Sci	PB&R	JEP:Gen	Total
(1) Correct-frequentist	4	9	12	25
(2a) Incorrect-frequentist – whole population	7	15	10	32
(2b) Incorrect-frequentist – current sample	25	29	12	66
(3) Bayesian-analysis	3	8	3	14
Total	39	61	37	137

*Note.* Journal names are abbreviated as follows: Psychological Science (Psych Sci); Psychonomic Bulletin & Review (PB&R); Journal of Experimental Psychology: General

<sup>47</sup> We contacted 19 authors in total; 4 did not reply, 10 provided the required information, and 5 did not provide the required information.

(JEP:G). “Whole population” refers to statements that are generalized to the whole population; “Current sample” refers to statements that are restricted to the current sample.

### ***Bayesian analyses***

From the 175 statistics that we collected from the articles, we identified 67 *t*-tests and we were able to acquire the necessary information for the analyses of 63 tests.<sup>48</sup> Calculating the Bayes factors ( $BF_{01}$ ) with medium scale ( $r = \sqrt{2}/2$ ) Cauchy prior under the alternative hypothesis, the 63 *t*-tests yielded 16 *anecdotal* (25%), 45 *moderate* (71%) and 2 *strong* (3%)  $BF_{01}$ s, all of them in favor of the null hypothesis. Both of the *strong*  $BF_{01}$ s were obtained in studies with sample sizes over 300 participants (see the exploratory analyses and Figure 3 for a more thorough description of the link between sample size and  $BF_{01}$ s).

### ***Robustness test***

The above results were obtained for a specific prior distribution (i.e., a two-tailed medium-scale Cauchy distribution on the standardized effect size). To probe the robustness of the results against other prior distributions, we calculated the corresponding  $BF_{01}$ s of the 63 *t*-tests using normal (Dienes, 2014) and informed priors (Gronau, Ly, & Wagenmakers, 2017; see the Supplementary Materials for a detailed description). Figure 1 shows  $BF_{01}$ s ordered by their size as calculated with the default prior. The figure also indicates the proportions of the  $BF_{01}$ s in the different evidence categories. The default prior resulted in 74.6% ( $n = 47$ )  $BF_{01}$ s greater than 3 (providing at least a *moderate* evidence for the null), whereas with the informed prior this was the case for only 44.5% ( $n = 28$ ) of the  $BF_{01}$ s.  $BF_{01}$ s computed with the normal prior showed even weaker evidential support for the null, as only 25.4% ( $n = 16$ ) of them exceeded  $BF_{01} = 3$ . In sum, applying the informed rather than the default prior changes the evidential category of the  $BF_{01}$ s in 20 cases (31.7%); and application of the normal prior results in 33 (52.4%) changes compared to the default prior. However, as it is apparent from Figure 1, the differences between the values of the  $BF_{01}$ s calculated with the distinct models are in most cases not substantial. The high number of different evidence categorizations is due to the fact that the majority of the  $BF_{01}$ s are scattered around the category thresholds.

---

<sup>48</sup> None of the tests referred to randomization failures.



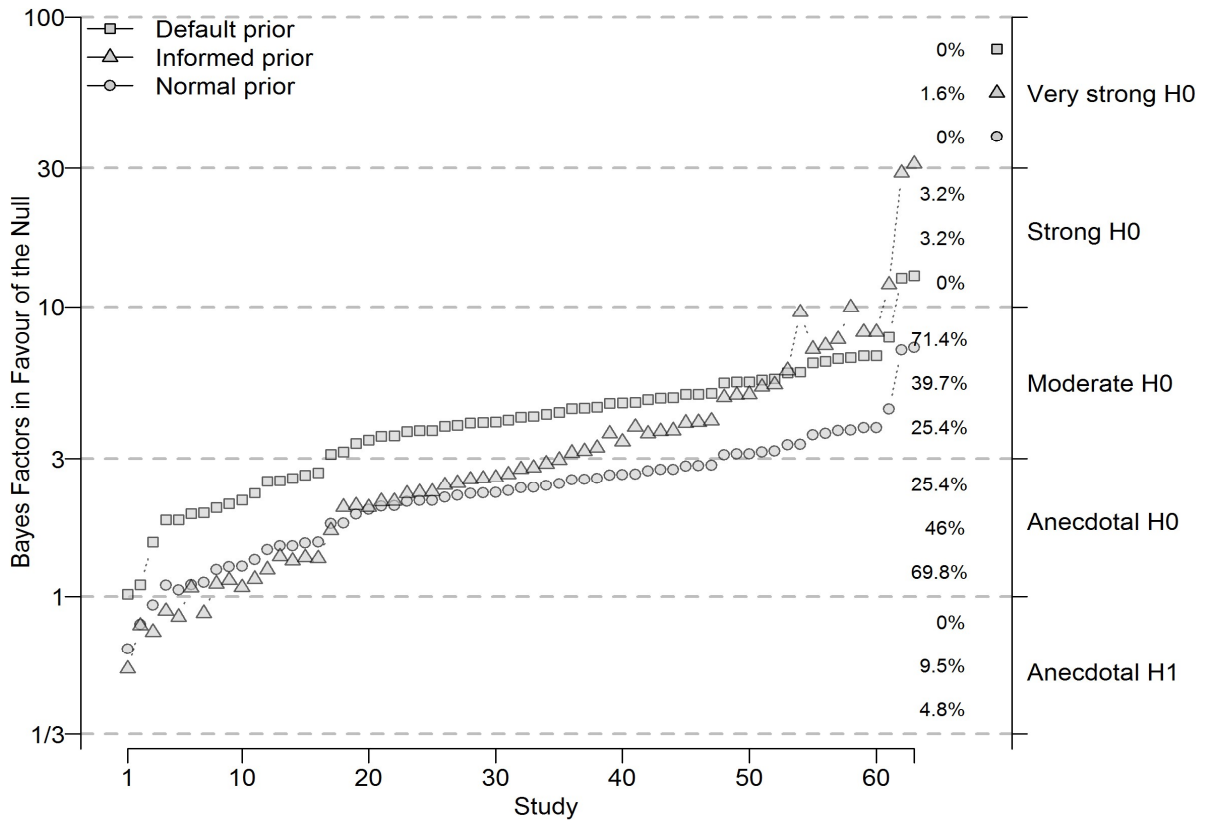
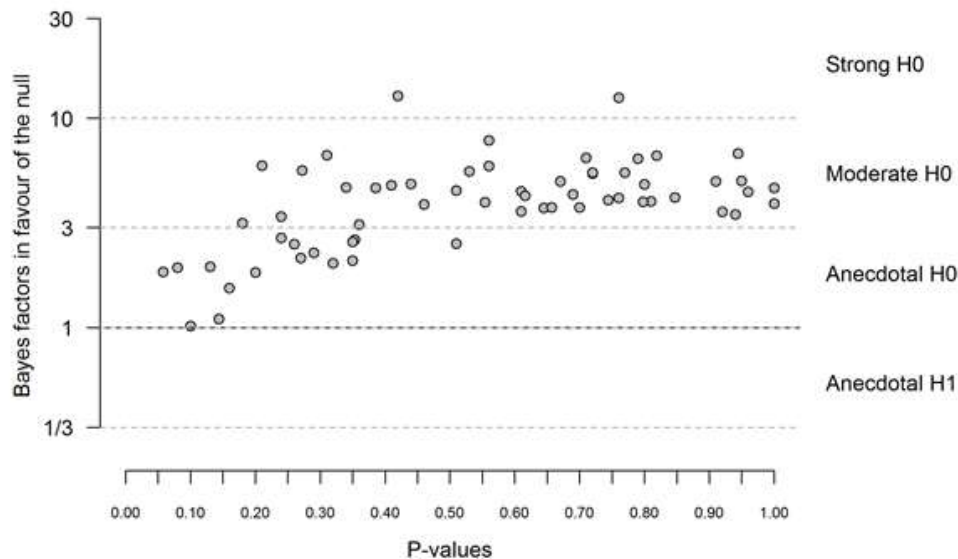


Figure 1. Three different Bayes factors in favor of the null hypothesis for each of the 63 nonsignificant *t*-tests reported in the selected literature. The Bayes factors were calculated with default, informed, and normal prior specifications of the alternative hypothesis. Note that the scaling of the y-axis has been  $\log_e$  transformed to help visualize the relationship. The right-hand side of the y-axis represents Jeffreys' (1961) classification scheme. Within each region, the numbers describe the proportion of all results falling in that category when using default, informed, and normal prior specifications, respectively. The figure is reproduced from: <https://doi.org/10.6084/m9.figshare.5721076.v2>

## Non-preregistered analyses

### *Bayesian analyses*

To explore the extent to which  $p$ -values and the corresponding  $BF_{01s}$  are associated, we plotted the reported  $p$ -values<sup>49</sup> against  $BF_{01s}$  (see Figure 2) and conducted Bayesian parameter estimation by computing Kendall's tau and its 95% CI. The correlation analysis revealed that the relationship between the  $p$ -values and the  $BF_{01s}$  is moderate and it is likely that the true value of the correlation falls between .20 and .50 ( $\tau = .38$ , 95% CI [.20, .52]). Figure 2 shows that this moderate relation is driven primarily by the correlation between the low  $p$ -values (smaller than .3) and  $BF_{01s}$ , and that the values of the  $BF_{01s}$  level off for  $p$ -values higher than .3. The figure also shows that high  $p$ -values do not guarantee *strong* evidence for the null.

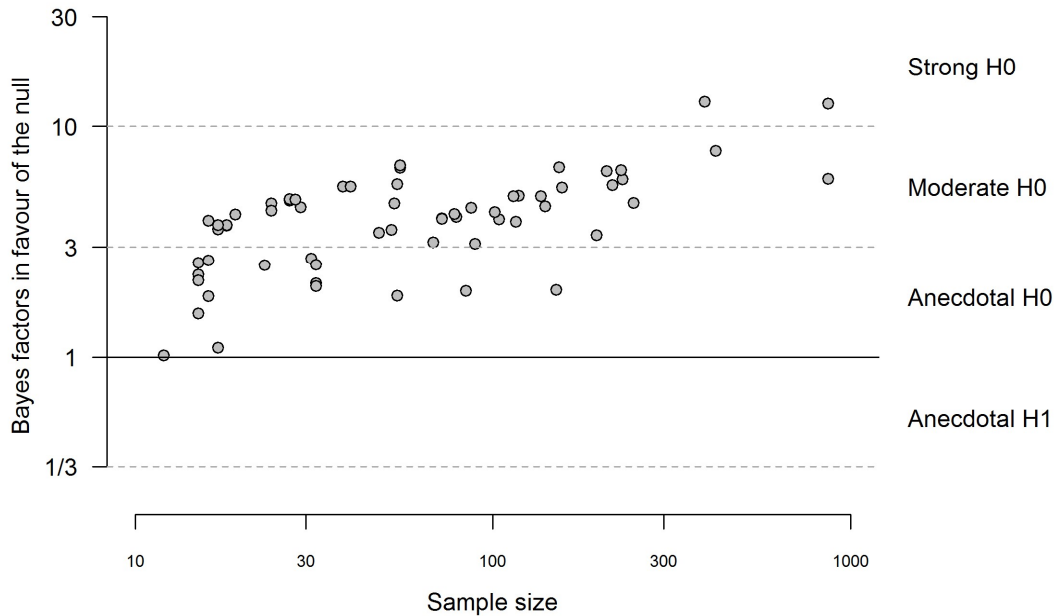


*Figure 2.* Relationship between the  $p$ -values and the corresponding default  $BF_{01s}$  ( $\tau = .38$ , 95% CI [.20, .52]). Note that the scaling of the y-axis has been  $\log_e$  transformed to help visualize the relationship. Dots situated above the bold vertical line indicate evidence for the null hypothesis. The figure is reproduced from: <https://doi.org/10.6084/m9.figshare.5721076.v2>

Next, we investigated the relationship between sample sizes and  $BF_{01s}$ . It is apparent from Figure 3 that the majority of the anecdotal  $BF_{01s}$  (13 cases, 81.25% of all anecdotal BFs) were obtained in studies with small sample sizes ( $n < 35$ ). In contrast, 48% (12 cases) of the small samples produced moderate evidence towards the null hypothesis. Strong evidence was only

<sup>49</sup> Note that 4 of the 63  $p$ -values were obtained from one-sided tests. As the focus of our interest is how researchers interpret the nonsignificant  $p$ -values, we did not modify these values to the results of two-sided tests.

reached in studies with large samples ( $n > 300$ ) and all of these studies provided at least moderate evidence in favor of the null. To estimate the strength of the association between sample size and  $BF_{01}$ , we calculated the correlation coefficient and its 95% credible interval (CI). We found a positive correlation between sample size and  $BF_{01}$  ( $\tau = .45$ , 95% CI [.26, .59]).



*Figure 3.* The relationship between the sample sizes of the investigated studies and the corresponding  $BF_{01}$ s ( $\tau = .45$ , 95% CI [.26, .59]). Both of the scaling factors of the axes are  $\log_e$  transformed. The figure is reproduced from: <https://doi.org/10.6084/m9.figshare.5721076.v2>

## Discussion

The goal of this study was twofold: to explore how psychology researchers interpret and communicate nonsignificant results, and to assess how much these results truly constitute evidence in favor of the null hypothesis. To this aim, we collected all the negative statements from the abstracts of three domain-general psychology journals and we extracted and reanalyzed their corresponding statistics.

The analysis of the null-statements of the abstracts demonstrates that there are several ways in which researchers interpret nonsignificant results. Importantly, only 28% ( $n = 39$ ) of these statements are in agreement with the logic of the employed statistical methods (frequentist: 18%,  $n = 25$ ; Bayesian: 10%,  $n = 14$ ). Regarding the incorrect inferences, in the smaller fraction of the statements (23%,  $n = 32$ ), researchers concluded that there is no effect in the population.

The most prevalent strategy when interpreting nonsignificant results, however, was to limit the relevance of the results to the observed sample (48%,  $n = 66$ ). While it is possible that the words that researchers use to describe their results did not reflect what they meant to say, awareness must be raised to this habit since interpreting the results of an inferential test with respect to the observed sample is not meaningful.

In our exploratory analysis, the comparison of the extracted statistical results to all the reported statistical results from the same year of the journals (see Supplementary Results) give the impression that researchers are less likely to build an argument on a nonsignificant result if the corresponding  $p$ -value is small compared to when it is large. These observations underscore the apparent confusion and uncertainty in the interpretation of nonsignificant results and they also reflect that the field has no generally applied strategy to discuss nonsignificant findings.

The apparent confusion and uncertainty in research practice possibly originate from the fact that while researchers are motivated to discuss all of their findings, the NHST framework is not designed to be informative about negative results (Fisher, 1935). As Fisher (1935) wrote: "*In relation to any experiment we may speak of this hypothesis as the 'null hypothesis,' and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.*" (p. 18). This limitation resulted not just in a great number of uncommunicated negative results (Franco et al., 2014), but also, as this study shows, in unwarranted interpretation of negative findings.

To assess the extent to which the reported nonsignificant findings constitute the absence of evidence (i.e., nondiagnostic results produced by low power) or evidence of absence (i.e., support for the null hypothesis) we conducted a Bayes factor reanalysis. The interpretation of Bayes factor results is always conditional on the level of support we expect from our data regarding our hypotheses (Aczel, Palfi, & Szaszi, 2017). As apparent in Figure 2, almost all results remained under  $BF_{01} = 10$  and a great proportion of them under  $BF_{01} = 3$ . While there are different Bayes factor labeling traditions (Schönbrodt, 2015), values lower than 3 are most often interpreted as anecdotal, and values lower than 10 are generally not considered strong evidence (Lee & Wagenmakers, 2013). Note that, when the null hypothesis and the alternative hypothesis are deemed equally likely a priori, a Bayes factor of 3 raises the model probability for the null hypothesis from 50% to 75% (leaving a full 25% for the alternative hypothesis), and a Bayes factor of 10 raises it from 50% to 91% (leaving 9% for the alternative hypothesis).

This result, and its robustness to alternative prior specification of the alternative hypothesis, suggests that the nonsignificant findings that were elevated to the abstracts of the investigated studies provide at best only moderate evidence for the authors' negative claim. In a considerable number of cases, the nonsignificant findings presented in the abstracts carry evidence that is not worth more than a bare mention (Etz & Vandekerckhove, 2017; Jeffreys, 1961). This weakness can be partly due to the typically low sample sizes in psychology (see e.g., Aczel, Palfi, Szaszi, Szollosi, & Dienes, 2015; Kekecs et al., 2016). Hoekstra et al. (2017), reanalyzing nonsignificant results in medicine, found much stronger evidence for the null hypothesis with samples two or three magnitudes greater than ours. Our result showing a moderate link between sample size and Bayes factor further corroborates this explanation.

Taken together, our results extend the list of reasons why the current practice of research in psychological science needs to be reconsidered. It is a long-known problem that positive results are more attractive (Giner-Sorolla, 2012) and that they are more likely to be published than negative results (Franco et al., 2014; Rosenthal, 1979). This publication bias is often blamed for creating misleading and non-replicable findings and for resulting in immense loss of resources (Lilienfeld & Waldman, 2017). Here, we show that even when these negative findings are reported they are often miscommunicated or lack sufficient evidential support. In fact, the situation did not improve since Hoekstra et al. (2006) observed that 61% of the psychology articles published between 2002 and 2004 claimed no effect or of a negligible effect based only on statistically nonsignificant results. We suggest that this 'curse' on negative results is not just due to a lack of attraction, but also to its problematic status within the NHST tradition as well as to the chronic underestimation of required sample sizes in psychological experiments.

It is to be noted that our sample was constrained to *t*-tests of papers published in 3 journals from 2015. Nevertheless, we would not expect substantially different pattern of results since in a recent Bayesian reanalysis of over 300,000 published significant *t*-, *F*- and *r*-test results indicated that the strength of evidence is comparable among the different statistical tests in psychological studies (Aczel et al., 2017). The generalizability of any Bayesian analysis is subject to the predictions of the tested hypotheses, determined by their prior distributions. Here, we examine the robustness of our conclusions with a range of different prior distributions and we obtained the same pattern of results.

Transparency of research conduct and communication is of primary importance for improving the field. However, the field may also benefit from adopting a more inclusive statistical approach. For instance, the proponents of Bayes factors argue that it could help

alleviate several the current challenges. Bayes factors can be interpreted not just against, but also for the null hypothesis; Bayes factors are insensitive to stopping rules, allowing the experimenter to stop data-collection whenever the evidence for one of the hypotheses is sufficiently compelling (Dienes, 2016; Rouder, 2014; but see de Heide & Grünwald, 2017). The Bayes factor is not the only tool to test the absence of an effect or to demonstrate that an effect is too small to be practically relevant. For instance, parameter estimation with confidence intervals (e.g., Cumming, 2014) can inform us about the size of an effect, and equivalence testing (Lakens, 2017), a frequentist procedure that is conceptually similar to the Bayesian Region Of Practical Equivalence (ROPE; e.g., Kruschke, 2014, chapter 12), provides a way to accept the null hypothesis if a region of negligible effect sizes can be determined. Nonetheless, these alternative methods cannot be applied to test a point-null hypothesis, which was the primary focus of the current study.

It has long been highlighted that psychological experiments are often underpowered (Cohen, 1990). The statistical power of a typical two-group between-subjects design is estimated to be less than .35 (Bakker et al. 2012) and only 3% of the psychological studies report using power analysis. While these issues might be traced back to some inappropriate rules of thumb existing among research psychologists (Bakker et al., 2016), our results provide further evidence that without a substantial increase in statistical power, psychologists' data can provide only weak evidence in favor of the null hypothesis.

## **Conclusion**

Our findings reveal that in the published literature nonsignificant findings are often misinterpreted. Moreover, Bayesian reanalyzes reveal that most studies that report a nonsignificant finding provide only limited evidence for the null hypothesis. These observations suggest that nonsignificant findings, as traditionally reported, can easily mislead the reader. Specific statistical training, a more skeptical mindset, and an extension of the standard statistical toolbox are possible remedies to promote a more adequate communication and a more appropriate assessment of negative results.

## **Author Contributions**

BA, BP, AS, BS and EJW conceptualized and wrote the manuscript. MK, PS and MZ contributed to the data collection and methodology. BP, MK, GQF and DvdB contributed to

the analysis and visualization of the results. All authors reviewed and approved the final version of the manuscript.

### **Acknowledgments**

We thank Maarten Marsman for his help with the code and Melissa Wood for her comments on earlier versions of the manuscript. Bence Palfi is grateful to the Dr Mortimer and Theresa Sackler Foundation which supports the Sackler Centre for Consciousness Science. Aba Szollosi was supported by the “Nemzet Fiatal Tehetségeiért” Scholarship (NTP-NFTÖ-16-1184) and E-J Wagenmakers was supported by a Vici grant from the Netherlands Organisation for Scientific Research (NWO, 016.Vici.170.083).

## References

- Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE*, *12*(8), e0182651. <https://doi.org/10.1371/journal.pone.0182651>
- Aczel, B., Palfi, B., Szaszi, B., Szollosi, A., & Dienes, Z. (2015). Commentary: Unlearning implicit social biases during sleep. *Frontiers in psychology*, *6*:1428. <https://doi.org/10.3389/fpsyg.2015.01428>
- American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5<sup>th</sup> ed.). American Psychological Association, Washington, DC, USA
- Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological science*, *27*(8), 1069-1077.
- Bakker, M., van Dijk, A., Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour*, *1*.
- Cohen, J. (1990). Things I have learned (so far). *American psychologist*, *45*(12), 1304-1312.
- Cohen, J. (1994). The earth is round ( $p < 0.05$ ). *American Psychologist*, *49*, 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29.
- de Heide, R., & Grünwald, P. D. (2017). Why optional stopping is a problem for Bayesians. *arXiv preprint arXiv:1708.08278*.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Basingstoke, UK: Palgrave Macmillan.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. <https://doi.org/10.1016/j.jmp.2015.10.003>
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193-242.
- Etz, A., & Vandekerckhove, J. (2017). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*. <https://doi.org/10.3758/s13423-017-1262-3>
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. <https://doi.org/10.1177/1745691612459059>



- Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.
- Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, *7*(6), 562–571. <https://doi.org/10.1177/1745691612457576>
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. <https://doi.org/10.1053/j.seminhematol.2008.04.003>
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. <https://doi.org/10.1037/h0076157>
- Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian t-tests. Retrieved from <http://arxiv.org/abs/1704.02479>
- Hartgerink, C. H. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data*, *1*(3), 14. <https://doi.org/10.3390/data1030014>
- Hartgerink, C. H., van Aert, R. C., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. <https://doi.org/doi.org/10.7717/peerj.1935>
- Hartgerink, C. H., Wicherts, J. M., & van Assen, M. A. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology*, *3*(1).
- Hoekstra, R., Finch, S., Kiers, H. A., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review*, *13*(6), 1033-1037.
- Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2017). Bayesian reanalysis of null results reported in the New England Journal of Medicine: Strong yet variable evidence for the absence of treatment effects. *Manuscript*
- Jeffreys, H. (1961). *The theory of probability*. Oxford, UK: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795. <https://doi.org/10.1080/01621459.1995.10476572>
- Kekecs, Z., Szollosi, A., Palfi, B., Szaszi, B., Kovacs, K. J., Dienes, Z., & Aczel, B. (2016). Commentary: Oxytocin-gaze positive loop and the coevolution of human–dog bonds. *Frontiers in neuroscience*, *10*:155. <https://doi.org/10.3389/fnins.2016.00155>
- Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). London: Edward Arnold.
- Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362.

- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.
- Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. Chichester, UK: John Wiley & Sons.
- Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12-2). Retrieved from <https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf>
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. <https://doi.org/10.1038/s41562-016-0021>
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231*, 289–337.
- Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.
- Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. <https://doi.org/10.3758/s13428-015-0664-2>
- Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036–1042. <https://doi.org/10.1177/0956797616645672>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. <https://doi.org/10.3758/s13423-014-0595-4>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Schönbrodt, F. (2015, April 17). Grades of evidence - A cheat sheet. Retrieved from <http://www.nicebread.de/grades-of-evidence-a-cheat-sheet/>
- Signorell, A. (2017). DescTools: Tools for descriptive statistics (Version 0.99.22). Retrieved from <https://cran.r-project.org/web/packages/DescTools/index.html>
- van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2016). Bayesian inference for Kendall's rank correlation coefficient. *The American Statistician*.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>

### 4.3. Discussion points for Bayesian inference<sup>50</sup>

**Authors:**

Balazs Aczel<sup>1\*</sup>, Rink Hoekstra<sup>2</sup>, Andrew Gelman<sup>3</sup>, Eric-Jan Wagenmakers<sup>4</sup>, Irene G. Klugkist<sup>5</sup>, Jeffrey N. Rouder<sup>6</sup>, Joachim Vandekerckhove<sup>6</sup>, Michael D. Lee<sup>6</sup>, Richard D. Morey<sup>7</sup>, Wolf Vanpaemel<sup>8</sup>, Zoltan Dienes<sup>9</sup>, and Don van Ravenzwaaij<sup>2</sup>

**Affiliations:**

<sup>1</sup>ELTE, Eötvös Loránd University, Budapest, Hungary

<sup>2</sup>University of Groningen, Groningen, The Netherlands

<sup>3</sup>Columbia University, New York, USA

<sup>4</sup>University of Amsterdam, Amsterdam, The Netherlands

<sup>5</sup>Utrecht University, Utrecht, Utrecht, The Netherlands

<sup>6</sup>University of California, Irvine, USA

<sup>7</sup>University of Cardiff, Cardiff, UK

<sup>8</sup>University of Leuven, Leuven, Belgium

<sup>9</sup>University of Sussex, Brighton, UK

\*Correspondence should be sent to [aczel.balazs@ppk.elte.hu](mailto:aczel.balazs@ppk.elte.hu)

---

<sup>50</sup> published as:

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E. J., Klugkist, I. G.,... & van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, 4(6), 561-563.

## **Standfirst**

Why is there no consensual way of conducting Bayesian analyses? We present a summary of agreements and disagreements of the authors on several discussion points regarding Bayesian inference. We also provide a thinking guideline to assist researchers on conducting Bayesian inference in the social and behavioural sciences.

## **Debates among Bayesians**

Despite its many advocates, Bayesian inference is currently employed by only a minority of social and behavioural scientists. One possible barrier is a lack of consensus on how best to conduct and report such analyses. Employing Bayesian methods involves making choices about prior distributions, likelihood functions, and robustness checks, as well as on how to present, visualize, and interpret the results (for a glossary of the main Bayesian statistical concepts see Box 1). Some researchers may find this wide range of choices too daunting to use Bayesian inference in their own study. This paper highlights the areas of agreement and the arguments behind disagreements, established on the back of a self-questionnaire explained in detail in the Supplement.

The overall message is that instead of following rituals<sup>1,2</sup>, researchers should understand the reasoning behind the different positions and make their choices on a case by case basis. To assist the reader in this task, we provide a summary of our views on seven discussion points in Bayesian inference, serving as an inspiration for a ‘thinking guideline’ as a guide towards conducting Bayesian inference in the social and behavioural sciences.

Our paper attempts to highlight the degree of debate that persists around the topic and explains why there are no easy-to-implement heuristics on how to use Bayesian analyses. Information about the genesis of this project can be found in the Supplementary Information and on OSF (<https://osf.io/6eqx5/>).

***Box 1 Glossary for the main statistical concepts discussed in this Comment*****Bayes factor**

The relative support provided by the data for one model over another model in the form of an odds ratio.

**Bayesian estimation**

Branch of Bayesian statistical inference in which (an) unknown population parameter(s) is/are estimated.

**Bayesian testing**

Branch of (Bayesian) statistical inference in which competing hypotheses are tested.

**Credible intervals**

A probabilistic interval that is believed to contain a given parameter.

**Likelihood**

The probability (density) of the data given a model for a particular (set of) parameter(s).

**Likelihood function**

A function of the parameters of a statistical model, given specific observed data. Consider, for instance, a coin with an unknown rate probability  $r$  of coming up heads on a single flip. For the specific data of two flips, each coming up heads  $\{H, H\}$ , the likelihood function of  $r$  is  $L(r|H,H) = \Pr(\{(H,H)\}|r) = r^2$ . For instance, given these observed data, the likelihood of the specific value  $r = 0.6$  is  $0.6^2 = 0.36$ .

**Posterior (distribution)**

Used in Bayesian inference to quantify an updated state of belief about some hypotheses (such as parameter values) after observing data.

**Prior (distribution)**

Used in Bayesian inference to quantify a state of belief about some parameter values *given a model* before having observed any data. Typically represented as a probability distribution over different states of belief.

**Posterior model probability**

Used in Bayesian inference to quantify an updated state of belief about the plausibility of a given model after observing data. The ratio of prior model probabilities times the Bayes factor for these same models gives the ratio of posterior model probabilities.

**Prior model probability**

Used in Bayesian inference to quantify a state of belief about the plausibility of a given model without taking observed data into account.

**Robustness check**

Used in Bayesian inference to verify the extent to which the obtained results are affected by (typically modest) variations of prior distribution and/or likelihood function.

**Discussion Points**

1. *When would you recommend using Bayesian parameter estimation and when Bayesian testing (i.e., Bayes factors)? Do you think there is a fundamental difference between the two?*

There are (mathematical) similarities between testing and estimation, although the two approaches often have different goals in practice. Bayesian testing is generally used to test whether an effect is present; in contrast estimation is used to assess the size/strength of the effect. A big difference between the two approaches lies in the nature of the (joint) prior distribution, which tends to be discontinuous for testing, but continuous for estimation. An argument to consider estimation more informative, especially when credible intervals are calculated, is that it provides information about the uncertainty of the estimated parameter(s). Bayes factors are generally considered suitable to assess evidence for or against competing hypotheses (or models). Researchers tend to use estimation when they want to examine a single model or several models very similar to each other but testing when they examine (at least two) models that differ from each other.

*2. A. How should the prior distribution and likelihood function for Bayesian analyses be chosen?*

Typically, there is a lot more emphasis on the choice of prior than on the choice of likelihood in Bayesian inference, but it is just as important to use the right model -- instantiated by the likelihood function -- for the data. Some Bayesian statisticians favour subjective priors over objective/default/uninformative ones, because uninformative priors are unrealistic, or because every scientific endeavour begins with an (informed) choice of both prior and likelihood. Uninformative priors should be chosen when assessing evidence for certain parameter values, but informative priors should be chosen when assessing evidence for one model over another. When using informative priors, uninformative priors can serve a role in fitting baseline models for comparison. A slightly less wide-spread strategy is choosing priors and likelihoods iteratively, obtaining prior predictive distributions of the model, and checking whether they lead to plausible data patterns. For example, it can be valuable to choose a sceptic's prior, a believer's prior, and a personal prior, and compare the possibly diverging results to determine how much the obtained results are influenced by prior beliefs.

*2.B. When and how do you think robustness checks should be performed in Bayesian analyses?*

Robustness checks are performed to verify whether the obtained results are affected by for modest variations of the prior distribution but should also be used to verify the influence of the choice of the likelihood function on the obtained results. The main argument for the importance of performing robustness checks over reasonable variations in modelling choices is to increase confidence in the obtained results: ideally results should be reasonably unaffected by a researcher's idiosyncratic choice of prior or likelihood function when reasonable alternatives exist. When performing robustness checks, it is crucial to determine first which modelling choices may impact the results and perform your checks accordingly. They are primarily important when working with non-informative, and therefore more arbitrary priors.

*3. What do you think about using point null hypotheses versus (small) interval hypotheses when testing within the Bayesian framework?*

First of all, it is important to consider if the research question is best served by testing rather than estimating. A researcher should consider what a practically relevant effect is before having

seen the data and set up an interval test accordingly. There is some agreement regarding the practical usefulness of the point null as a model to reflect invariance, but the viewpoint is open to critique: In the end, it may not matter that much, it would be rare for a point null and a small interval around null to lead to practically different conclusions, since the point null is a useful model as an approximation of a near-zero interval. In some cases, the parsimonious point null helps flag the need for more data in case a (much) more complex model is believed to be true. Ultimately, researchers should use whichever they are most interested in (or both, to test robustness).

*4. How would you recommend reporting Bayesian analysis results?*

Although there is no agreement on a necessary reporting format, there are some important markers that are considered helpful in assessing the evidence. These include the model and its assumptions, prior distributions, choice of likelihood and posterior, potential hypotheses to be evaluated, details about samples from the posterior<sup>3</sup> when applicable, and robustness tests. It is helpful to report results in terms of competing and completely specified models. Providing figures that show estimates with uncertainty, accompanied by Bayes factors when applicable is important.

*5. How would you recommend visualizing the results of a Bayesian analysis on diagrams?*

For Bayesian estimation, it is good practice to plot posteriors of parameters as a measure of uncertainty in case of estimation. Unless it creates an information overload, marginal predictions of a model and observed data should be plotted together, so that readers can see how authors came to their conclusions.

For Bayesian testing, plots can include information on whether the Bayes factor reaches a meaningful threshold to facilitate the reader in drawing conclusions. It may be unwise to standardize data visualization as no solution fits all purposes.

*6. How would you recommend interpreting Bayesian analysis results (with a robustness test)?*

There are good arguments why it may be better to focus on the scientific rather than on the statistical interpretation because it helps the reader understand what the results mean and what



the uncertainties of the presented conclusions are. One helpful chain of interpretation would go from (modelling) assumptions to observed data to conclusions, possibly with a similar chain for an alternative (but plausible) set of assumptions. When interpreting Bayes factors, presenting them through the lens of betting, especially when accompanied by real-world examples of odds (i.e., Team A is deemed three times more likely to win than Team B) may be a helpful way of providing an intuition of the meaning of a Bayes factor. The same holds for providing illustrative visualizations and ranges for your qualitative conclusions when interpreting results.

*7. A. Should we use Bayesian analysis for making decisions about the evidence?*

One option for making decisions involves using Bayes factors. As an example, consider a researcher who obtains a Bayes factor of 10 for the hypothesis that a new medicine against migraine reduces symptoms over the hypothesis that the new medicine does not reduce symptoms. Should this Bayes factor be used to make a decision (i.e., endorse the new medication, so that it can be sold by pharmacies)?

Some Bayesian statisticians think we should, offering that Bayes factors are suitable to do so. This, however, requires reliance on related utilities as well as probabilities (see supplementary materials for a concrete example). A second option involves doing Bayesian utility analysis based on the posterior from a single fitted model. Other Bayesian statisticians state that making decisions about the evidence is optional and perhaps better left to policy makers rather than researchers. This echoes similar debates among frequentists<sup>4</sup>.

*7. B. Would you recommend a decision threshold, an a priori sample size, or anything else?*

There are arguments speaking against decision thresholds, e.g., (1) the behaviour of Bayes factors for different kinds of hypotheses is insufficiently understood such that it may lead to arbitrary decision making, both about the fate of the manuscript that reports them and about the true state of the world; (2) the strength of evidence (and the number of data points) needs to be understood within the research context; (3) even the smallest study can contribute useful information; (4) basing a decision on decision thresholds alone does not incorporate utilities. One of us believes that standard decision thresholds are useful as a convention because it facilitates making a decision about the evidence (see previous question) and has been active in

having journals implement them. Perhaps a compromise is to consider standard decision thresholds a useful heuristic for evaluating the statistical evidence, without using them as a basis for publishing papers.

### **Questions to consider**

This list of discussion points shows some of the disagreement that exists on major discussion points, but also that differing opinions are supported by arguments. The bottom line, endorsed by all authors, is: Use common sense. To assist the reader in this task, we compiled a ‘thinking guideline’ (Box 2) which aims to orient the attention to the questions that should be considered when conducting Bayesian statistics.

#### **Box 2 Thinking Guideline for Bayesian Inference**

##### **Questions to consider when conducting Bayesian statistics**

###### **1. Why use Bayesian statistics?**

Possible reasons include: (1) given a model, the strength of evidence only depends on data that were actually observed; (2) the results do not depend on the intention of the researcher; (3) the evidence is quantified as relative for one model or hypothesis over another model or hypothesis; and (4) the possibility to include prior information or beliefs.

For general introductions to Bayesian inference, see ref<sup>5-8</sup>.

###### **2. Are you interested in estimation or testing?**

Conduct a test when a binary question of some kind needs to be answered (e.g., “Can people see into the future?”). In such cases, a particular parameter value, such as zero, often has a special status when testing. Estimate parameters, possibly after having conducted a test, when your main interest is about the extent of the effect (e.g., “Assuming that they can, what is their predictive accuracy?”)<sup>9,10 p 274,11 p 385</sup>.

### **3. How will you choose the prior distribution and likelihood function for Bayesian analyses?**

If you have relevant prior information available, for example based on prior study results, incorporate this in your prior distribution<sup>12-15</sup>. If not, consider using a ‘default’ (testing), or uninformative (estimation) prior. When you have several plausible candidates for your likelihood function, perform model comparisons.

### **4. How do you plan to demonstrate the robustness of your analysis?**

Examine whether similar results would be obtained for different, but plausible, choices for the prior distribution. Perform model comparison when one has different, but plausible, choices for the likelihood function. One can couple robustness checks to decision thresholds, to verify for what range of prior assumptions a certain decision would be taken.

### **5. How do you plan to communicate your results?**

Think about whether your results are best communicated through descriptive (summary) statistics (when the results are easily presented in the main text), graphics (when a visualisation conveys the information better), or tables (when there is too much information to present in a figure)<sup>14</sup>. The choice should also be guided by the research topic, the intended audience, and the type of analysis.

**6. Whatever you do, at each choice and decision in your analysis, be prepared to answer the ‘why’ question!**

Statistical analyses are sequences of choices. Understanding the implications of these choices and carefully thinking about them on a case by case basis are the responsibility of the author. Step-by-step guidelines and rituals can never substitute statistical thinking.

To conduct statistical inference is to make choices, for Bayesian inference, this dilemma remains. We hope that the thinking guideline that we present here is able to guide some of the choices necessary for analysing work in the behavioural and social sciences and informs researchers of some of the opinions of those in the field.

## **Method**

### *Participants*

This project employed an iterative survey method to explore the agreements and disagreements among experts on conceptual and practical questions in Bayesian analysis. The first two and last authors facilitated the project (henceforth facilitators). The facilitators approached seven Bayesian statisticians (henceforth experts). The criteria for the selection of these experts were that Bayesian inference constituted a central topic of their scientific work in recent years and that they were active in the social sciences. One expert declined, and three new experts were suggested and subsequently approached by the facilitators.

Ultimately, the following nine experts agreed to participate in this study (henceforth they will be addressed by their initials): Andrew Gelman (AG), Eric-Jan Wagenmakers (EJW), Irene Klugkist (IK), Jeffrey N. Rouder (JR), Joachim Vandekerckhove (JV), Michael D. Lee (MDL), Richard D. Morey (RM), Wolf Vanpaemel (WV), and Zoltan Dienes (ZD).

### *Materials*

The first version of the survey consisted of eight questions regarding topics the facilitators deemed relevant. The first wave of experts was asked whether these questions were clear, and whether any important issues were missing. The experts were given the opportunity to suggest modifications in the phrasing of the questions or recommend new questions, which they did in a few instances. Eventually, the survey consisted of nine questions about the following topics: testing vs. estimation, choosing priors and robustness, point null vs. small intervals, reporting of results, visualization, interpretation, and decision making. The survey questions and the summary of the responses are presented in the Manuscript.

### *Procedure*

Once the survey was finalized, the participating experts were asked to provide their answers to the questions. After all of the responses were collected, the facilitators summarized the answers to each question. If required, the experts were asked to clarify their positions. In the second round, a summary for each question and the detailed responses of all of the experts were shared with the panel members. Thus, experts were given the opportunity to modify their original answers. Following this, the facilitators used the experts' comments to amend and extend the summary text. The facilitators implemented modifications in the summary until all of the experts were satisfied with the text. The first round of the study took about two months and the second round took about one month.

The nine experts continued their participation in the study until the end of the project. All of the experts accepted the final version of the opinion summary. Their full responses to the questions are available from <https://osf.io/6eqx5/>.

### *Preregistration*

The preregistration of our procedure is available here: <https://osf.io/q37as/>.

### *Notable Deviation from Preregistration*

Although the preregistration protocol stated we would include 7-8 experts, we ended up with 9.

### **Note**

We limited our panel to Bayesian statisticians who are predominantly active in the social sciences, to present a paper applicable to Bayesian inference in this field. Thus, the presented opinions apply to problems in the social and behavioural sciences. Two anonymous reviewers outside the social sciences expressed views different from those of the experts on several topics. For example, one reviewer argued that robustness checks should never be performed when estimating, because they do not make any sense, given there is only one prior. This highlights that different fields may have different schools when it comes to Bayesian analysis, likely because of the different nature of data under consideration.

We found many benefits of surveying expert opinions as a way to provide the readers with a review of arguments behind diverging views. Probably, it resulted in a more comprehensive and balanced picture than any review a single author could write. We recommend the use of this approach in the future, especially for reviews on contested topics.

### **Author Contributions**

B.A., R.H., and D.v.R. conceptualized the project, conducted the study survey and wrote the manuscript. A.G., E-J.W., I.G.K., J. N.R., J.V., M.D.L., R.D.M., W.V., and Z.D. contributed to the summary of this review and added suggestions to the manuscript. The authorship order follows the alphabetical order of their first names. All authors reviewed and approved the final version of the manuscript.

**References**

1. Gigerenzer, G. *J. Socio-Econ.* **33**, 587–606 (2004).
2. Gigerenzer, G. *Adv. Methods Pract. Psychol. Sci.* **1**, 198–218 (2018).
3. van Ravenzwaaij, D., Cassey, P. & Brown, S. D. *Psychon. Bull. Rev.* **25**, 143–154 (2018).
4. Fisher, R. *Journal of the Royal Statistical Society: Series B (Methodological)* **17**, 69–78 (1955).
5. Dienes, Z. *Understanding psychology as a science: An introduction to scientific and statistical inference.* (Palgrave Macmillan, 2008).
6. Etz, A. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 5–34 (2018).
7. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan.* (Academic Press, 2014).
8. Wagenmakers, E.-J. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
9. Haaf, J. M., Ly, A. & Wagenmakers, E.-J. *Nature* **567**, 461 (2019).
10. Fisher, R.A. *Statistical Methods for Research Workers*, 2nd Edit. *Oliver Boyd Edinb.* (1928).
11. Jeffreys, H. *Theory of Probability, section 3.23.* (Oxford: Clarendon Press, 1948).
12. Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. *The Am. Stat.* 1–14 (2019).
13. Wagenmakers, E.-J. *et al. Psychon. Bull. Rev.* **25**, 58–76 (2018).
14. Matzke, D., Boehm, U. & Vandekerckhove, J. *Psychon. Bull. Rev.* **25**, 77–101 (2018).
15. van Doorn, J. *et al.* The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. (2019).

**Competing interests**

The authors declare no competing interests.

## 4.4. One statistical analysis must not rule them all

### Comment

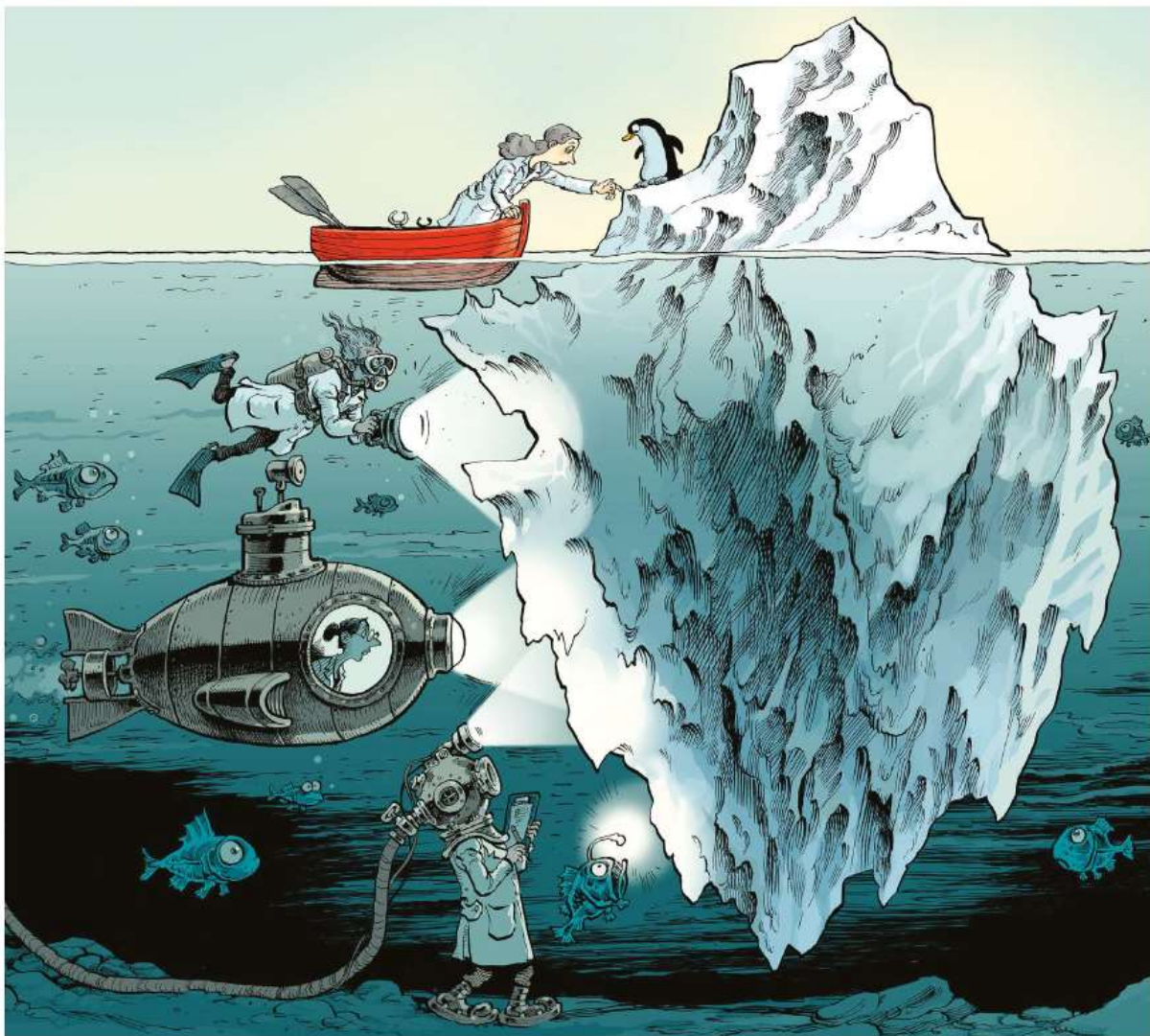


ILLUSTRATION BY DAVID PARKINS

## One statistical analysis must not rule them all

Eric-Jan Wagenmakers, Alexandra Sarafoglou & Balazs Aczel

Any single analysis hides an iceberg of uncertainty. Multi-team analysis can reveal it.

**A** typical journal article contains the results of only one analysis pipeline, by one set of analysts. Even in the best of circumstances, there is reason to think that judicious alternative analyses would yield different outcomes.

For example, in 2020, the UK Scientific Pandemic Influenza Group on Modelling asked nine teams to calculate the reproduction number  $R$  for COVID-19 infections<sup>1</sup>. The

teams chose from an abundance of data (deaths, hospital admissions, testing rates) and modelling approaches. Despite the clarity of the question, the variability of the estimates across teams was considerable (see 'Nine teams, nine estimates').

On 8 October 2020, the most optimistic estimate suggested that every 100 people with COVID-19 would infect 115 others, but perhaps as few as 96, the latter figure implying that



**Comment**

the pandemic might actually be retreating. By contrast, the most pessimistic estimate had 100 people with COVID-19 infecting 166 others, with an upper bound of 182, indicating a rapid spread. Although the consensus was that the trajectory of disease spread was cause for concern, the uncertainty across the nine teams was considerably larger than the uncertainty within any one team. It informed future work as the pandemic continued.

**Flattering conclusion**

This and other ‘multi-analyst’ projects show that independent statisticians hardly ever use the same procedure<sup>2-6</sup>. Yet, in fields from ecology to psychology and from medicine to materials science, a single analysis is considered sufficient evidence to publish a finding and make a strong claim.

Over the past ten years, the concept of *P*-hacking has made researchers aware of how the ability to use many valid statistical procedures can tempt scientists to select the one that leads to the most flattering conclusion. Less understood is how restricting analyses to a single technique effectively blinds researchers to an important aspect of uncertainty, making results seem more precise than they really are.

To a statistician, uncertainty refers to the range of values that might reasonably be taken by, say, the reproduction number of COVID-19 or the correlation between religiosity and well-being<sup>6</sup>, or between cerebral cortical thickness and cognitive ability<sup>7</sup>, or any number of statistical estimates. We argue that the current mode of scientific publication – which settles for a single analysis – entrenches ‘model myopia’, a limited consideration of statistical assumptions. That leads to overconfidence and poor predictions.

To gauge the robustness of their conclusions, researchers should subject the data to multiple analyses; ideally, these would be carried out by one or more independent teams. We understand that this is a big shift in how science is done, that appropriate infrastructure and incentives are not yet in place, and that many researchers will recoil at the idea as being burdensome and impractical. Nonetheless, we argue that the benefits of broader, more-diverse approaches to statistical inference could be so consequential that it is imperative to consider how they might be made routine.

**Charting uncertainty**

Some 100 years ago, scholars such as Ronald Fisher advanced formal methods for hypothesis testing that are now considered indispensable for drawing conclusions from numerical data. (The *P* value, often used to determine ‘statistical significance’, is the best known.) Since then, a plethora of tests and methods have been developed to quantify inferential

uncertainty. But any single analysis draws on a very limited range of these. We posit that, as currently applied, uncertainty analyses reveal only the tip of the iceberg.

The dozen or so formal multi-analyst projects completed so far (see Supplementary Information) show that levels of uncertainty are much higher than that suggested by any single team. In the 2020 Neuroimaging Analysis Replication and Prediction Study<sup>2</sup>, 70 teams used the same functional magnetic resonance imaging (MRI) data to test 9 hypotheses about brain activity in a risky-decision task. For example, one hypothesis probed how a brain region is activated when people consider the prospect of a large gain. On average across the hypotheses, about 20% of the analyses constituted a ‘minority report’ with a qualitative conclusion opposite to that of the majority. For the three hypotheses that yielded the

**“Formal methods cannot cure model myopia, because they are firmly rooted in the single-analysis framework.”**

most ambiguous outcomes, around one-third of teams reported a statistically significant result, and therefore publishing work from any of one these teams would have hidden considerable uncertainty and the spread of possible conclusions. The study’s coordinators now advocate that multiple analyses of the same data be done routinely.

Another multi-analyst project was in finance<sup>3</sup> and involved 164 teams that tested 6 hypotheses, such as whether market efficiency changes over time. Here again, the coordinators concluded that differences in findings were due not to errors, but to the wide range of alternative plausible analysis decisions and statistical models.

All of these projects have dispelled two myths about applied statistics. The first myth is that, for any data set, there exists a single, uniquely appropriate analysis procedure. In reality, even when there are scores of teams and the data are relatively simple, analysts almost never follow the same analytic procedure.

The second myth is that multiple plausible analyses would reliably yield similar conclusions. We argue that whenever researchers report a single result from a single statistical analysis, a vast amount of uncertainty is hidden from view. And although we endorse recent science-reform efforts, such as large-scale replication studies, preregistration and registered reports, these initiatives are not designed to reveal statistical fragility by exploring the degree to which plausible alternative analyses can alter conclusions. In summary, formal methods, old and new,

cannot cure model myopia, because they are firmly rooted in the single-analysis framework.

We need something else. The obvious treatment for model myopia is to apply more than one statistical model to the data. High-energy physics and astronomy have a strong tradition of teams carrying out their own analyses of other teams’ research once the data are made public. Climate modellers routinely perform ‘sensitivity analyses’ by systematically removing and including variables to see how robust their conclusions are.

For other fields to make such a shift, journals, reviewers and researchers will have to change how they approach statistical inference. Instead of identifying and reporting the result of a single ‘correct’ analysis, statistical inference should be seen as a complex interplay of different plausible procedures and processing pipelines<sup>8</sup>. Journals could encourage this practice in at least two ways. First, they could adjust their submission guidelines to recommend the inclusion of multiple analyses (possibly reported in an online supplement)<sup>9</sup>. This would motivate researchers to either conduct extra analyses themselves or to recruit more analysts as co-authors. Second, journals could invite teams to contribute their own analyses in the form of comments on a recently accepted article.

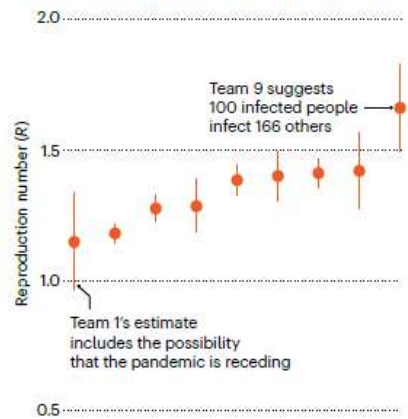
**False alarm?**

Certainly, large-scale changes in how science is done are possible: expectations surrounding the sharing of data are growing. Medical journals now require that clinical trials be registered at launch for the results to be published. But proposals for change inevitably prompt critical reactions. Here are five that we’ve encountered.

**Won’t readers get confused?** Currently, there are no comprehensive standards for, or conventions on, how to present and interpret the

**NINE TEAMS, NINE ESTIMATES**

Comparing models of the rate of COVID-19’s spread in the United Kingdom in early October 2020 revealed a degree of uncertainty masked by any one model.



SOURCE: REF.1

results of multiple analyses, and this situation could complicate how results are reported and make conclusions more ambiguous. But we argue that potential ambiguity is a key feature of multi-team analysis, not a bug. When conclusions are supported only by a subset of plausible models and analyses, readers should be made aware. Facing uncertainty is always better than sweeping it under the rug.

**Aren't other problems more pressing?** Problems in empirical science include selective reporting, a lack of transparency around analyses, hypotheses that are divorced from the theories they are meant to support, and poor data sharing. It is important to make improvements in these areas – indeed, how data are collected and processed, and how variables are defined, will greatly influence all subsequent analyses. But multi-analyst approaches can still bring insight. In fact, multi-analyst projects usually excel in data sharing, transparent reporting and theory-driven research. We view the solutions to these problems as mutually reinforcing rather than as a zero-sum game.

**Is it really worth the time and effort?** Even those who see benefit in multiple analyses might not see a need for them to happen at the time of publication. Instead, they would argue that the original team be encouraged to pursue multiple analyses or that shared data can be reanalysed by other interested researchers after publication. We agree that both would be an improvement over the status quo (sensitivity analysis is a severely underused practice). However, they will not yield the same benefits as multi-team analyses done at the time of publication.

Post-publication analyses are usually published only if they drastically undercut the original conclusion. They can give rise to squabbles more than constructive discussion, and would come out after the authors and readers have already drawn conclusions based on a single analysis. Information about uncertainty is most useful at the time of analysis. However, we doubt whether a single team can muster the mental fortitude needed to reveal the fragility of their findings; there might be a strong temptation to select those analyses that, together, present a coherent story. In addition, a single research team usually has a somewhat narrow expertise in data analysis. For instance, each of the nine teams that produced different estimates for  $R$  would probably feel uncomfortable if they had to code and produce estimates using the other teams' models. Even for simple statistical scenarios (that is, a comparison of two outcomes – such as the proportions of people who improve after receiving a drug or placebo – and a test of a linear correlation), several teams can apply widely divergent statistical models and procedures<sup>10</sup>.

Some sceptics doubt that multi-team

analyses will consistently find broad enough ranges of results to make the effort worthwhile. We think that the outcomes of existing multi-analyst projects counter that argument, but it would be useful to gather evidence from yet more projects. The more multi-analyst approaches are undertaken, the clearer it will be as to how and when they are valuable.

**Won't journals balk?** One sceptical response to our proposal is that multi-analyst projects will take longer, be more complicated to present and assess, and will even require new article formats – complications that will make journals reluctant to embrace the idea. We counter that the review and publication of a multi-analyst paper do not require a fundamentally different process. Multi-team projects have been published in a variety of journals, and most journals already publish comments attached

**“Journals, governments and philanthropists should actively recruit or support multi-analysis teams.”**

to accepted manuscripts. We challenge journal editors to give multi-analyst projects a chance. For instance, editors might test the waters by organizing a special issue consisting of case studies. This should make it readily apparent whether the added value of the multi-analyst approach is worth the extra effort.

**Won't it be a struggle to find analysts?** One response to our proposal is that the bulk of multi-team analyses published so far are the product of demonstration projects wrapped into a single paper. These papers encompass several analyses with long author lists comprised mainly of enthusiasts for reform; most other researchers would see little benefit in being a minor contributor to a multi-analyst paper, especially one at the periphery of their core research interest. But we think enthusiasm has a broad base. In our multi-analyst projects, we have been known to receive more than 700 sign-ups in about 2 weeks.

Moreover, a range of incentives could attract teams of analysts, such as gaining co-authorship and the chance to work on important questions or simply to collaborate with specialists. Further incentives and catalysts are easy to imagine. In a forthcoming special issue of the journal *Religion, Brain & Behavior*, several teams will each publish their own conclusions and interpretations of the research question addressed by the main article<sup>6</sup>, and this means each team's contribution is individually recognized. When a question is particularly urgent, journals, governments and philanthropists should actively recruit or support multi-analysis teams.

Yet another approach would be to incorporate multiple analyses into training programs, which would be both useful for the research community and eye-opening for statisticians. (At least one university has incorporated replication studies into its curricula<sup>11</sup>.) Ideally, participating in multiple analyses will be seen as part of being a good science 'citizen', and be rewarded through better prospects for hiring and promotion.

Whatever the mix of incentives and formats, the more that multiple analyses efforts are implemented and discussed, the easier they will become. What makes such multi-team efforts work well should be studied and applied to improve and expand the practice. As the scientific community learns how to run multi-team analyses and what can be learnt, acceptance and enthusiasm will grow.

We argue that rejecting the multi-analyst vision would be like Neo opting for the blue pill in the film *The Matrix*, and so continuing to dream of a reality that is comforting but false. Scientists and society will be better served by confronting the potential fragility of reported statistical outcomes. It is crucial for researchers and society to have an indication of such fragility from the moment the results are published, especially when these results have real-world ramifications. Recent many-analyst projects suggest that any single analysis will yield conclusions that are overconfident and unrepresentative. Overall, the benefit of increased insight will outweigh the extra effort.

## The authors

**Eric-Jan Wagenmakers** is a methodologist and **Alexandra Sarafoglou** a postdoctoral fellow at the University of Amsterdam, the Netherlands. **Balazs Aczel** is vice dean for science at Eötvös Loránd University in Budapest, Hungary. e-mails: ej.wagenmakers@gmail.com; alexandra.sarafoglou@gmail.com; balazs.aczel@gmail.com

1. Scientific Pandemic Influenza Group on Modelling. *SPH-M-O: Consensus statement on COVID-19*, 8 October 2020 (2020).
2. Botvinnik-Nezer, R. et al. *Nature* **582**, 84–88 (2020).
3. Menkveld, A. J. et al. Preprint at SSRN <https://doi.org/10.2139/ssrn.3961574> (2021).
4. Silberzahn, R. et al. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
5. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).
6. Hoogeveen, S. et al. Preprint at PsyArXiv <https://doi.org/10.31234/osf.io/pbfye> (2022).
7. Marek, S. et al. *Nature* **603**, 654–660 (2022).
8. Wagenmakers, E.-J. et al. *Nature Hum. Behav.* **5**, 1473–1480 (2021).
9. Aczel, B. et al. *eLife* **10**, e72185 (2021).
10. van Dongen, N. N. N. et al. *Am. Stat.* **73**, 328–339 (2019).
11. Button, K. *Nature* **561**, 287 (2018).

The authors declare no competing interests. Supplementary information accompanies this comment online (see [go.nature.com/37hdvkb](https://go.nature.com/37hdvkb)).

## 4.5. Consensus-based guidance for conducting and reporting multi-analyst studies<sup>51</sup>

Balazs Aczel<sup>1\*</sup>, Barnabas Szaszi<sup>1\*</sup>, Gustav Nilsson<sup>2,3</sup>, Olmo R van den Akker<sup>4</sup>, Casper J Albers<sup>5</sup>, Marcel A L M van Assen<sup>4,6</sup>, Jojanneke A Bastiaansen<sup>7,8</sup>, Dan Benjamin<sup>9,10</sup>, Udo Boehm<sup>11</sup>, Rotem Botvinik-Nezer<sup>12</sup>, Laura F Bringmann<sup>5</sup>, Niko A Busch<sup>13</sup>, Emmanuel Caruyer<sup>14</sup>, Andrea M Cataldo<sup>15,16</sup>, Nelson Cowan<sup>17</sup>, Andrew Delios<sup>18</sup>, Noah N N van Dongen<sup>11</sup>, Chris Donkin<sup>19</sup>, Johnny B van Doorn<sup>11</sup>, Anna Dreber<sup>20,21</sup>, Gilles Dutilh<sup>22</sup>, Gary F Egan<sup>23</sup>, Morton Ann Gernsbacher<sup>24</sup>, Rink Hoekstra<sup>5</sup>, Sabine Hoffmann<sup>25</sup>, Felix Holzmeister<sup>21</sup>, Juergen Huber<sup>21</sup>, Magnus Johannesson<sup>20</sup>, Kai J Jonas<sup>26</sup>, Alexander T Kindel<sup>27</sup>, Michael Kirchner<sup>21</sup>, Yoram K Kunkels<sup>7</sup>, D Stephen Lindsay<sup>28</sup>, Jean-Francois Mangin<sup>29,30</sup>, Dora Matzke<sup>11</sup>, Marcus R Munafö<sup>31</sup>, Ben R Newell<sup>19</sup>, Brian A Nosek<sup>32,33</sup>, Russell A Poldrack<sup>34</sup>, Don van Ravenzwaaij<sup>5</sup>, Jörg Rieskamp<sup>35</sup>, Matthew J Salganik<sup>27</sup>, Alexandra Sarafoglou<sup>11</sup>, Tom Schonberg<sup>36</sup>, Martin Schweinsberg<sup>37</sup>, David Shanks<sup>38</sup>, Raphael Silberzahn<sup>39</sup>, Daniel J Simons<sup>40</sup>, Barbara A Spellman<sup>33</sup>, Samuel St-Jean<sup>41,42</sup>, Jeffrey J Starns<sup>43</sup>, Eric L Uhlmann<sup>44</sup>, Jelte Wicherts<sup>4</sup>, Eric-Jan Wagenmakers<sup>11</sup>

<sup>1</sup>ELTE, Eotvos Lorand University, Budapest, Hungary, <sup>2</sup>Karolinska Institutet, Stockholm, Sweden, <sup>3</sup>Stockholm University, Stockholm, Sweden, <sup>4</sup>Tilburg University, Tilburg, The Netherlands, <sup>5</sup>University of Groningen, Groningen, The Netherlands, <sup>6</sup>Utrecht University, Utrecht, The Netherlands, <sup>7</sup>University of Groningen, University Medical Center Groningen, Groningen, The Netherlands, <sup>8</sup>Friesland Mental Health Care Services, Leeuwarden, The Netherlands, <sup>9</sup>University of California Los Angeles, Los Angeles, CA, USA, <sup>10</sup>National Bureau of Economic Research, Cambridge, MA, USA, <sup>11</sup>University of Amsterdam, Amsterdam, The Netherlands, <sup>12</sup>Dartmouth College, Hanover, NH, USA, <sup>13</sup>University of Münster, Münster, Germany, <sup>14</sup>University of Rennes, CNRS, Inria, Inserm, Rennes, France, <sup>15</sup>McLean Hospital, Belmont, MA, USA, <sup>16</sup>Harvard Medical School, Boston, MA, USA, <sup>17</sup>Department of Psychological Sciences, University of Missouri, MO, USA, <sup>18</sup>National University of Singapore,

---

<sup>51</sup> published as:

Aczel, B., Szaszi, B., Nilsson, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *ELife*, 10, e72185.

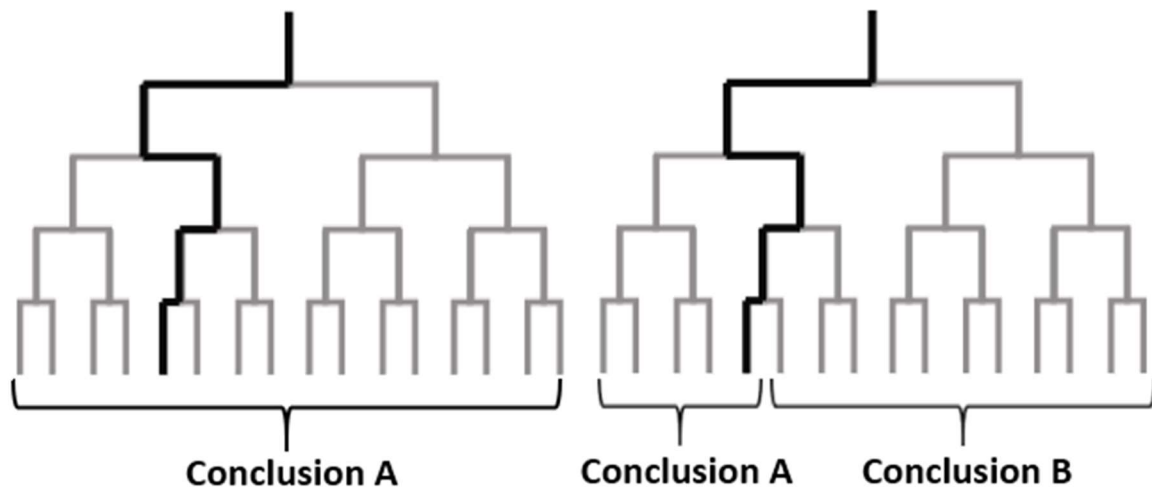
Singapore, <sup>19</sup>University of New South Wales, Sydney, Australia, <sup>20</sup>Stockholm School of Economics, Stockholm, Sweden, <sup>21</sup>University of Innsbruck, Innsbruck, Austria, <sup>22</sup>University Hospital Basel, Basel, Switzerland, <sup>23</sup>Monash University, Melbourne, Victoria, Australia, <sup>24</sup>University of Wisconsin-Madison, Madison, WI, USA, <sup>25</sup>Ludwig-Maximilians-University, Munich, Germany, <sup>26</sup>Maastricht University, Maastricht, The Netherlands, <sup>27</sup>Princeton University, Princeton, NJ, USA, <sup>28</sup>University of Victoria, Victoria, Canada, <sup>29</sup>Université Paris-Saclay, Paris, France, <sup>30</sup>Neurospin, CEA, France, <sup>31</sup>University of Bristol, Bristol, UK, <sup>32</sup>Center for Open Science, USA, <sup>33</sup>University of Virginia, Charlottesville, USA, <sup>34</sup>Stanford University, Stanford, USA, <sup>35</sup>University of Basel, Basel, Switzerland, <sup>36</sup>Tel Aviv University, Tel Aviv, Israel, <sup>37</sup>ESMT Berlin, Germany, <sup>38</sup>University College London, London, UK, <sup>39</sup>University of Sussex, Brighton, UK, <sup>40</sup>University of Illinois at Urbana-Champaign, USA, <sup>41</sup>University of Alberta, Edmonton, Canada, <sup>42</sup>Lund University, Lund, Sweden, <sup>43</sup>University of Massachusetts Amherst, USA, <sup>44</sup>INSEAD, Singapore

## Abstract

Any large dataset can be analyzed in a number of ways, and it is possible that the use of different analysis strategies will lead to different results and conclusions. One way to assess whether the results obtained depend on the analysis strategy chosen is to employ multiple analysts and leave each of them free to follow their own approach. Here, we present consensus-based guidance for conducting and reporting such multi-analyst studies, and we discuss how broader adoption of the multi-analyst approach has the potential to strengthen the robustness of results and conclusions obtained from analyses of datasets in basic and applied research.

## Introduction

Empirical investigations often require researchers to make a large number of decisions about how to analyze the data. However, the theories that motivate investigations rarely impose strong restrictions on how the data should be analyzed. This means that empirical results typically hinge on analytical choices made by just one or a small number of researchers, and raises the possibility that different – but equally justifiable – analytical choices could lead to different results (Figure 1).



**Figure 1. Analysis choices and alternative plausible paths.** The analysis of a large dataset can involve a sequence of analysis choices, as depicted in these schematic diagrams. The analyst first must decide between two options at the start of the analysis (top), and must make three additional decisions during the analysis: this leads to 16 possible paths for the analysis (grey lines). The left panel shows an example in which all possible paths lead to the same conclusion; the right panel shows an example in which some paths lead to conclusion A and other paths lead to conclusion B. Unless we can test alternative paths, we cannot know if the results obtained by following one particular path (thick black line) are robust, or if other plausible paths would lead to different results.

This "analytical variability" may be particularly high for datasets that were not initially collected for research purposes (such as electronic health records) because data analysts might know relatively little about how those data were collected and/or generated. The increasing availability of large, routinely-collected datasets (from, for example, administrative claims or electronic health records), offers the promises of "real-world" evidence and personalized treatment regimes. However, when analyzing such datasets – and when making decisions based on the results of such analyses – it is important to be aware that the results will be subject to higher levels of analytical variability than the results obtained from analyses of data from, say, clinical trials.

A recent example of the perils of analytical variability is provided by two articles in the journal *Surgery* that used the same dataset to investigate the same question: does the use of a retrieval bag during laparoscopic appendectomy reduce surgical site infections? Each paper used reasonable analysis, but there were notable differences between them in how they addressed

inclusion and exclusion criteria, outcome measures, sample sizes, and covariates. As a result of these different analytical choices, the two articles reached opposite conclusions: one paper reported that using a retrieval bag reduced infections (1), and the other reported that it did not (2; see also 3). This and other medical examples (4–6) illustrate how independent analysis of the same data can reach different, yet justifiable, conclusions.

The robustness of results and conclusions can be studied by evaluating multiple distinct analysis options simultaneously (e.g., vibration of effects (7) or multiverse analysis (8)) or by engaging multiple analysts to independently analyze the same data. Rather than exhaustively evaluating all plausible analyses, the multi-analyst method examines analytical choices that are deemed most appropriate by independent analysts. Botvinik-Nezer et al. (13), for example, asked 70 teams to test the same hypotheses using the same functional magnetic resonance imaging dataset. They found that no two teams followed the same data preprocessing steps or analysis strategies, which resulted in substantial variability in the teams' conclusions. This and other work (9–12, 14–18) confirms how results can depend on analytic choices

Although the multi-analyst approach will be new to many researchers, it has been in use since the 19th century. In 1857, for example, the Royal Asian Society asked four scholars to independently translate a previously unseen inscription to verify that the ancient Assyrian language had been deciphered correctly. The almost perfect overlap between the solutions indicated that “they have Truth for their basis” (19). The same approach can be used to analyze data today. With just a few co-analysts, the multi-analyst approach can be informative about the analytic robustness of results and conclusions. When the results of independent data analyses converge, more confidence in the conclusions is warranted. However, when the results diverge, confidence will be reduced, and scientists can examine the reasons for these discrepancies and identify potentially meaningful moderators of the results. With enough co-analysts, it is possible to estimate the variability among analysis strategies and attempt to identify factors explaining this variability.

The multi-analyst approach is still rarely used, but we argue that many disciplines could benefit from its broader adoption. To help researchers overcome practical challenges, we provide consensus-based guidance (including a checklist) to help researchers surmount the practical challenges of preparing, conducting, and reporting multi-analyst studies.

## Methods

To develop this guidance, we recruited a panel of 50 methodology experts who followed a preregistered ‘reactive-Delphi’ expert consensus procedure (20). We adopted this procedure to ensure that the resulting guidance represents the shared thinking of relevant experts and that it incorporates their topic-related insights. The applied consensus procedure and its reporting satisfy the recommendations of CREDES (21), a guidance on conducting and reporting Delphi studies. A flowchart of the Delphi expert consensus procedure is available at <https://osf.io/pzkcs/>.

## Preparation

### *Preregistering the project*

Before the start of the project, on 11 November 2020, a research plan was compiled and uploaded to a time-stamped repository at <https://osf.io/dgrua>. During the project, we followed the preregistered plan in all respects except implementing slight changes in the wording of the survey questions to improve comprehension and not using R to analyze our results. We declared that we would share the R code and codebook of our analyses, but the project ultimately did not require us to conduct analyses in R. Instead, we shared our code in Excel and ODS format at <https://osf.io/h36qy/>.

### *Creating the initial Multi-Analyst Guidance draft*

Before the expert consensus process, the first three authors and the last author (henceforth: proposers) created an initial multi-analyst guidance draft after brainstorming and reviewing all the previously published multi-analyst-type projects they were aware of (9–18). This initial document is available here: <https://osf.io/kv8jt/>

### *Recruiting experts*

The proposers contacted 81 experts to join the project. The contacted experts included all the organizers of previous multi-analyst projects known at the time (9–18), as well as the members of the expert panel from another methodological consensus project (22). The previous projects were identified by conducting an unsystematic literature search and by surveying researchers in social media. Of the 81 experts, 3 declined our invitation and 50 accepted the invitation and participated in the expert consensus procedure (their names are available at <https://osf.io/fwqvp/>), while 28 experts did not respond to our call.

### **Preparatory rounds**

Upon joining the project, the experts received a link to the preparatory online survey (available at <https://osf.io/kv8jt/>) which included the initial Multi-Analyst Guidance draft where they had the option to comment on each of the items and the overall content of the guidance.

Based on the feedback received from the preparatory online survey, the proposers updated and revised the initial Multi-Analyst Guidance. This updated document was uploaded to an online shared document and was sent out to the experts who had the option to edit and comment on the content. Again, based on feedback, the proposers revised the content of the document, and this new version was included in the expert consensus survey.

### **Consensus survey**

The expert consensus questionnaire was sent out individually to each expert first on 8 February 2021 in the following Qualtrics survey available at <https://osf.io/wrpnq/>. The consensus survey approach had the advantage of minimizing potential biases in the experts' judgments: the questions were posed in a neutral way, experts all received the same questions, and experts did not see the responses of the other experts or any reaction of the project organizers. The survey contained the ten recommended practices grouped into the following five stages: i) recruiting co-analysts; ii) providing the dataset, research questions, and research tasks; iii) conducting the independent analyses; iv) processing the results; v) reporting the methods and results. The respondents were asked to rate each of the ten recommended practices on a nine-point Likert-type scale ('I agree with the content and wording of this guidance section' ranging from "1-Disagree" to "9-Agree"). Following each section, the respondents could leave comments regarding the given item.

The preregistration indicated consensus on the given item if the interquartile range of its ratings was two or smaller. It defined support for an item if the median rating was six or higher (as in 22).

Each recommended practice found support and consensus from the 48 experts who completed ratings in our first round. For each item, the median rating was eight or higher with an interquartile range of two or lower. Thus, following our preregistration, there was no need to conduct additional consensus-survey rounds; all of the items were eligible to enter the guidance with consensual support. This high level of consensus might have been due to the experts'



involvement in the preparatory round of the project. The summary table of the results is available at <https://osf.io/qc7a8/>.

### Finalising the manuscript

The proposers drafted the manuscript and supplements. All texts and materials were sent to the expert panel members. Each contributor was encouraged to provide feedback on the manuscript, the report, and the suggested final version of the guidance. After all discussions, minor wording changes were implemented, as documented at <https://osf.io/e39j4/>. No contributor objected to the content and form of the submitted materials and all approved the final item list.

### Multi-analyst Guidance

The final guidance includes ten recommended practices (Table 1) concerning the five main stages of multi-analyst studies. To further assist researchers in documenting multi-analyst projects, we also provide a modifiable reporting template (Supplementary file 1), as well as a reporting checklist (Supplementary file 2).

Table 1

Recommended Practices for the Main Stages of the Multi-Analyst Method

Stage	Recommended practices
Recruiting Co-analysts	<ol style="list-style-type: none"> <li>1. Determine a minimum target number of co-analysts and outline clear eligibility criteria before recruiting co-analysts. We recommend that the final report justifies why these choices are adequate to achieve the study goals.</li> <li>2. When recruiting co-analysts, inform them about (a) their tasks and responsibilities; (b) the project code of conduct (e.g., confidentiality/non-disclosure agreements); (c) the plans for publishing the research report and presenting the data, analyses, and conclusion; (d) the conditions for an analysis to be included or excluded from the study; (e) whether their names will be publicly linked to the analyses; (f) the co-analysts' rights to update or revise their analyses; (g) the project time schedule; and (h) the nature and criteria of compensation (e.g., authorship).</li> </ol>
Providing the Dataset, Research Questions, and Research Tasks	<ol style="list-style-type: none"> <li>3. Provide the datasets accompanied with a codebook that contains a comprehensive explanation of the variables and the datafile structure.</li> <li>4. Ensure that co-analysts understand any restrictions on the use of the data, including issues of ethics, privacy, confidentiality, or ownership.</li> </ol>

- |                                     |   |
|-------------------------------------|---|
| Conducting the Independent Analyses | <p>5. Provide the research questions (and potential theoretically derived hypotheses that should be tested) without communicating the lead team's preferred analysis choices or expectations about the conclusions.</p> <p>6. To ensure independence, we recommend that co-analysts should not communicate with each other about their analyses until after all initial reports have been submitted. In general, it should be clearly explained why and at what stage co-analysts are allowed to communicate about the analyses (e.g., to detect errors or call attention to outlying data points).</p>   |
| Processing the Results              | <p>7. Require co-analysts to share with the lead team their results, the analysis code with explanatory comments (or a detailed description of their point-and-click analyses), their conclusions, and an explanation of how their conclusions follow from their results.</p> <p>8. The lead team makes the commented code, results, and conclusions of all non-withdrawn analyses publicly available before or at the same time as submitting the research report.</p>   |
| Reporting the Methods and Results   | <p>9. The lead team should report the multi-analyst process of the study, including (a) the justification for the number of co-analysts; (b) the eligibility criteria and recruitment of co-analysts; (c) how co-analysts were given the data sets and research questions; (d) how the independence of analyses was ensured; (e) the numbers of and reasons for withdrawals and omissions of analyses; (f) whether the lead team conducted an independent analysis; (g) how the results were processed; (h) the summary of the results of co-analysts; (i) and the limitations and potential biases of the study.</p> <p>10. Data management should follow the FAIR principles (23), and the research report should be transparent about access to the data and code for all analyses (22).</p> |
- 

## Practical Considerations

In addition to the Multi-analyst Guidance and Checklist, we provide practical considerations that can support the organization and execution of multi-analyst projects. This section contains various clarifications, recommendations, practical tools, and optional extensions, covering the five main stages of a multi-analyst project.

## **Recruiting co-analysts**

### ***Choosing co-analysts***

The term co-analyst refers to one researcher or team of researchers working together in a multi-analyst project. Researchers can collaborate on the analyses, but if they do, we recommend that they submit the analyses as one co-analyst team, in order to ensure the independence of the analyses across teams. Researchers from the same lab or close collaborators should be able to submit separate reports in the multi-analyst project as long as they do not discuss their analyses with each other until the project rules allow that. The lead team may conduct an analysis themselves depending on the study goals and the design of the project (e.g., to set a performance baseline for comparing submitted models). Alternatively, the lead team may choose not to conduct an analysis themselves; in any case, they are expected to be transparent about their level of involvement as well as the timing (e.g., whether they conducted their analyses with or without knowing the results of the crowd of analysts).

Researchers should carefully consider both the breadth and depth of statistical and research-area expertise required for their project and should justify their choices about the required qualifications, skills, and credentials for analysts in the project. If the aim of the study is to explore what factors influence researchers' analytical choices, then it can be useful to seek "natural variation" (representativeness) within an expert community or to maximize diversity of the co-analysts along the dimensions where they might differ the most in their choices (e.g., experience, background, discipline, interest in the findings, intellectual allegiance to different theories, paradigmatic viewpoints).

### ***Deciding on the number of co-analysts***

To decide on the desired number of co-analysts, one has to consider which of the two main purposes of the multi-analyst method applies to the given project:

#### (A) Checking the robustness of the conclusions

The aim here is solely to check whether different analysts obtain the same conclusions. Confidence in the stability of the conclusions decreases with divergent results and increases with convergent results. Many projects can achieve this aim by recruiting only one additional analyst, or a handful of further analysts. For example, the above-mentioned two analyses of the same dataset published in the journal *Surgery* (1,2) were sufficient to detect that the analytical space allows for opposite conclusions.

## (B) Assessing the variability of the analyses

Those who wish to estimate the variability among the different analytical strategies often need to satisfy stricter demands. For example, studies that aim to assess how much the results vary among the analysts will require a larger number of co-analysts. When determining the number of co-analysts in such cases, the same factors need to be taken into consideration as in standard sample size estimation methods. For example, Botvinik-Nezer et al. (13) presented the analyses of 70 teams to demonstrate the divergence of results when analyzing a functional magnetic resonance imaging dataset.

### *Recruiting co-analysts*

Depending on the specific goal of the research, the recruitment of co-analysts can happen in several ways. Co-analysts can be recruited before or after obtaining the dataset. With stricter eligibility criteria, co-analysts can be invited individually from among topic experts or statistical experts. Follow-up open invitations can ask experts to suggest others to be invited. Alternatively, the lead team can open the opportunity to anyone to join the project as a co-analyst within the expert community (e.g., in professional society mailing lists and on social media), where expertise can be defined as the topic requires it.

It is important to note that whenever the co-authors' behavior is the subject of the study then they should be regarded similarly to human participants respecting ethical and data protection regulations. Useful templates for project advertisement and analyst surveys can be found in (12,24).

## **Providing the dataset, research questions, and research tasks**

### *Providing the dataset*

The lead team can invite the co-analysts to conduct data preprocessing (in addition to the main analysis). If the lead team decides to conduct the preprocessing themselves, showing their preprocessing methods can be informative to the co-analysts, but also has the potential to influence them if the preprocessing reflects some preference of methods or expectations of outcomes.

Before providing the dataset, the lead team should ensure that data management will comply with legal (e.g., the General Data Protection Regulation (GDPR) in the European Union) and ethical regulations applying to all teams (see 25). If the dataset contains personal information, a version should be provided where data can no longer be related to an individual. An alternative is to provide a simulated dataset and ask the co-analysts to provide code to analyze the data (26,27). The lead team can then run the code on the actual data.

It is important that the co-analysts understand not just the available dataset but also any ancillary information that might affect their analyses (e.g., prior exclusion of outliers or handling of missing data in the blinded dataset). Providing a codebook that is accessible and understandable for researchers with different backgrounds is essential (28).

#### Providing the research question

The provided research question(s) should motivate the analysis conducted by the co-analysts. The research questions should be conveyed without specifying preferred analysis choices or expectations about the conclusions. Depending on the purpose of the project, the research questions can be more or less specific. While more specific research questions limit the analytical freedom of the co-analysts, less specific ones better explore the ways researchers can diverge in their operationalization of their question. A research question (e.g., “Is happiness age-dependent?”) can be more specific when, for example, it is formulated as a directional hypothesis (e.g., “Are young people more happy than old ones?”) or when the constructs are better operationalized (e.g., by defining what counts as young and happy).

#### ***Providing the task***

The multi-analyst approach can leave the operationalization of the research question to the co-analysts so that they can translate the theoretical question into the measurement. Taking this approach can reveal the operational variations of a question, but it can also make it difficult to compare the statistical results.

Requesting results in terms of standardized metrics (e.g.,  $t$ -values, standardized beta, Cohen’s  $d$ ) makes it easier to compare results between co-analysts. The requested metric can be determined from the aim of the analysis (e.g., hypothesis testing, parameter estimation). It needs to be borne in mind, however, that this request might bias the analysis strategies towards using methods that easily provide such a metric. [A practical tool for instructions on reporting effect estimates: (29).]

Co-analysts should be asked to keep a record of any code, derivatives etc. that were part of the analysis, at least until the manuscript is submitted and all relevant materials are (publicly) shared.

As an extension, the co-analysts can be asked to record considered but rejected analysis choices and the reasoning behind their choices (e.g., by commented code, log-books, or dedicated solutions such as DataExplained (24)). These logs can reflect where and why co-analysts diverge in their choices.

Robustness, or multiverse analyses (in the sense that each team is free to provide a series of outcomes instead of a single one) can also be part of the task of the co-analysts so that multiple analyses are conducted under alternative data analysis preprocessing choices.

### ***Communication with co-analysts***

In projects with many co-analysts, keeping contact via a dedicated email address and automating some of the messages (e.g., automated emails when teams finished a stage in the process) can help streamline the communication and make the process less prone to human errors. For co-analyst teams with multiple members, it can be helpful for each team to nominate one member as the representative for communications.

If further information is provided to a co-analyst following specific questions, it can be useful to make sure the same information is provided to all teams, for example via a Q&A section of the project website, hosting weekly office hours where participants could ask questions, or via periodic email with updates.

## **Conducting the independent analyses**

### ***Preregistering the process and statistical analyses***

We can distinguish *meta-* and *specific preregistrations*. Meta-preregistrations concern the plan of the whole multi-analyst project. It is good practice for the lead team to preregister how they would process, handle, and report the results of the co-analysts in order to prevent result-driven biases. This can be done in the form of a Registered Report at journals that invite such submissions (30). Any metascientific questions, such as randomization of co-analysts to different conditions with variations in instructions or data, or covariates of interest for studying associations to analytic variability, should be specified.

Specific preregistrations concern the analysis plans of the co-analysts. Requiring co-analysts to prepare a specific preregistration for each analysis can be a strategy to prevent overfitting and undisclosed flexibility. It makes sense to require it from either all or none of the teams in order to maintain equal treatment among them (unless the effect of preregistration is a focus of the study).

Requiring specific preregistrations may be misaligned with the goals of the project when the aim is to explore how the analytic choices are formed during the analyses, independent of initial plans. Under such circumstances, requiring specific preregistrations may be counterproductive. Nevertheless, the lead team can record their meta-preregistration that lays down the details of the multi-analyst project.

There are alternative solutions to prevent researchers from being biased by their data and results. For example, co-analysts could be provided with blinded datasets (14,16,31), simulated datasets (27), or with a subset of the data (e.g., 11).

## **Processing the results**

### ***Collecting the results***

To facilitate summarizing the co-analysts' methods, results, and conclusions, the lead team can collect results through provided templates or survey forms that can structure analysts' reports. It is practical to ask the co-analysts at this stage to acknowledge that they did not communicate or cooperate with other co-analysts regarding the analysis in the project. It can also be helpful for the lead team if the co-analysts explain how their conclusions were derived from the results. In case preregistration was employed for any analyses, the template can also collect any deviations from the preregistered plan for inclusion in an online supplement.

To collect analytic code, it may be useful to require a container image (32,33) or a portable version of the code that handles issues like software package availability (34) (for a guideline see 35).

### ***Validating the results***

The lead team is recommended to ensure that each analyst's codes/procedures reproduce that analyst's submitted results. Computational reproducibility can be ascertained by running the code or repeating the analytic process by the lead team, but independent experts or the other co-analysts can also be invited to undertake this task (36,37).

The project can leverage the crowd by asking co-analysts to review others' analyses, or the lead team can employ external statistical experts to assess analyses and detect major errors. The lead team can decide to omit analyses with major errors. In that case, the reasons for omission should be documented, and for transparency, the results of the omitted analyses should be included in an online supplement.

After all the analyses have been submitted and validated, the co-analysts could have the option in certain projects to inspect the work of the other analysts and freely withdraw their own analyses. This can be appropriate if seeing other analyses makes them aware of major mistakes or shortcomings in their analytic procedures. A potential bias in this process is that co-analysts might lose confidence in their analyses after seeing other, more senior, or more expert co-analysts' work. One way to decrease this potential bias is to follow a multi-stage process: after the first round of analyses is submitted, co-analysts could be allowed to see each other's analysis steps/code without knowing the identity of the co-analyst or the results of their analysis. It is the lead team's decision whether they allow co-analysts to correct or update their analyses after an external analyst or the co-analysts themselves find issues in their analyses.

Importantly, it is a minimum expectation that from the start of the project, the co-analysts should know about the conditions for their analyses to be included in, or omitted from, the study. All withdrawals, omissions, and updates of the results should be transparent in subsequent publications, for example in the supplementary materials.

## **Reporting the methods and results**

### ***Recording contributorship***

Using CRediT taxonomy can transparently record organizers' and co-analysts' contributions to the study. Practical tools (e.g., tenzing 38) can make this task easier.

Co-analysts can be invited to be co-authors and/or be compensated for their contribution in other ways (e.g., prizes, honorariums). Expectations for contribution and authorship should be communicated clearly at the outset.

### ***Presenting the methods and results***

Beyond a descriptive presentation of results in a table or graph, the reporting of the results of multi-analyst projects is not straightforward and remains an open area of research. Published reports of multi-analyst projects have adopted several effective methods for presenting results.



For binary outcomes, Botvinik-Nezer et al. (39) used a table with color coding (i.e., a binary heat map) to visualize outcomes across all teams. They overlaid each teams' confidence in their findings and added additional information about analytical paths in adjacent columns (Supplementary Table 1). For a project with a relatively small number of effect sizes for continuous outcomes, Schweinsberg et al. (24) used interval plots combined with an indication of analytical choices underlying each estimate (Figure 3). Olsson Collentine et al. (40) (Figure 2) used funnel plots and Patel et al. (7) (Figures 1 and 2) used volcano plots to depict numerous, diverse outcomes with an intuitive depiction of clustering (akin to a multiverse analysis).

If the main purpose is to estimate variability of analyses, it is interesting to investigate and report factors that might influence variability in the chosen analytic approaches and in the results obtained by these analytical approaches. If, on the other hand, the main purpose is to investigate the robustness of conclusions by assessing the degree to which different analysts obtain the same results, it is advisable to focus more on methods that produce only a single answer to the research question of interest. When each analysis team can provide multiple, distinct responses to the same research question, it becomes more difficult to explore how conclusions depend on the analysis choices because the individual analyses are no longer independent of each other.

The analytical approach of each co-analyst can be divided into discrete choices concerning, for instance, data preprocessing steps and decisions in model specification. If it is possible to recombine the individual choices (which will not always be the case as certain data preprocessing steps or method choices may only make sense if the aim is to fit a certain class of models), it may be worthwhile to create a larger set of possible analytical approaches that is made up of all possible combinations. In this case, the descriptive results of the multi-analyst project can be combined with a multiverse type approach (e.g., vibration of effects 7, multiverse analysis 8, or specification curve 41) to quantify and compare the variability in results that can be explained by the different analytical choices (7,42). Additionally, this larger set of possible combinations can be helpful to present the results in an interactive user interface in which readers can explore how the results change as a function of certain analytical choices (42,43). Finally, dividing the co-analysts' analytical approaches into individual choices may ultimately help in providing a unique answer to the research question of interest while accounting for the uncertainty in the choice of the analytical approach. While there are so far no approaches that would allow the derivation of a unique result that integrates all uncertain decisions, it may be a

promising area of research to extend Bayesian approaches that account for model uncertainty (44) and measurement error (45).

To support the reporting of Multi-Analyst projects, we provide a freely modifiable *Reporting Template* available from here: <https://osf.io/h9mgy/>

## Limitations

The present work does not cover all aspects of multi-analyst projects. For instance, the multi-analyst approach outlined here entails the independent analysis of one or more datasets, but it should be acknowledged that other crowd-sourced analysis approaches might not require such independence of the analyses. Some of our practical considerations reflect disagreement and/or uncertainty within our expert panel, so they remain underspecified. Those include how to determine the number or eligibility of co-analysts for a project, how best to assess the validity of each analysis; and how to measure robustness of conclusions. Therefore, we emphasize that this consensus-based guidance is a first step towards the broader adoption of the multi-analyst approach in empirical research, and we hope and expect that our recommendations will be developed further in response to user feedback. Users of this guidance can provide feedback and suggestions for revisions at <https://forms.gle/2fVqZAD3KKHVUDKq7>.

## Conclusions

This guidance document aims to facilitate adoption of the multi-analyst approach in both basic and clinical research. Although the multi-analyst approach is at an incipient stage of adoption, we believe that the scientific benefits greatly outweigh the extra logistics required, especially for projects with high relevance for clinical practice and policy making. The approach should have particular relevance when it indicates that applying different analytical strategies to a given dataset may lead to conflicting results. The multi-analyst approach allows a systematic exploration of the analytical space to assess whether the reported results and conclusions are dependent on the chosen analytical strategy, ultimately improving the transparency, reliability, and credibility of research findings.

We hope that our guidance here and in guideline databases will make it easier for researchers to adopt this approach to empirical analyses. We encourage journals and funders to consider recommending or requesting independent analyses whenever it is crucial to know whether the conclusions are robust to alternative analytical strategies.

**Acknowledgements:** This research was not funded. AS was supported by a talent grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568). RB-N is an Awardee of the Weizmann Institute of Science – Israel National Postdoctoral Award Program for Advancing Women in Science. BAN was supported by grants from the John Templeton Foundation, Templeton World Charity Foundation, Templeton Religion Trust, and Arnold Ventures. SSt-J is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number BP-546283-2020] and the Fonds de recherche du Québec - Nature et technologies (FRQNT) [Dossier 290978]. JMW and ORvdA were supported by a Consolidator Grant (IMPROVE) from the European Research Council (ERC; grant no. 726361). YKK was supported by a grant from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (ERC-CoG-2015; No 681466 to M Wichers). DvR was supported by a Dutch scientific organization VIDI fellowship grant (016.Vidi.188.001). LFB was supported by a Dutch scientific organization VENI fellowship grant (Veni 191G.037). MJS was supported by the US National Science Foundation (1760052). ELU was supported by an R&D grant from INSEAD.

**Author contributors:** BA and BS are joined first authors and guarantors. BA, BS, GN, and E-JW were responsible for the study conception and design. ORvdA, CJA, MALMvA, JAB, DB, UB, RB-N, LFB, NB, EC, AMC, NC, A Delios, NNNvD, CD, JBvD, A Dreber, GD, GFE, MAG, RH, SH, FH, JH, MJ, KJJ, ATK, MK, YKK, DSL, J-FM, DM, MRM, BRN, BAN, RAP, DvR, JR, MJS, AS, TS, MS, DS, RS, DJS, BAS, SSt-J, JJS, ELU, and JW served as expert panel in the development of the guidance and checklist. All authors participated in drafting and critically revising the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Competing interests:** We have read the journal's policy and the authors of this manuscript have the following competing interests: BAN is Executive Director of the Center for Open Science, a non-profit technology and culture change organization with a mission to increase openness, integrity, and reproducibility of research. The other authors declare no competing interest.

**Data and materials availability:** All anonymized data as well as the survey materials are publicly shared on the Open Science Framework page of the project: <https://osf.io/4zvst/>. Our methodology and data-analysis plan were preregistered. The preregistration document can be accessed at: <https://osf.io/dgrua>.

**Transparency declaration:** The lead author affirms that the manuscript is an honest, accurate, and transparent account of the work being reported; that no important aspects of the study have been omitted; and that any discrepancies from the work as planned have been explained.

## References

1. Fields AC, Lu P, Palenzuela DL, Bleday R, Goldberg JE, Irani J, et al. Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? *Surgery*. 2019;165(5):953–7.
2. Turner SA, Jung HS, Scarborough JE. Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates. *Surgery*. 2019;165(6):1199–202.
3. Childers CP, Maggard-Gibbons M. Same Data, Opposite Results?: A Call to Improve Surgical Database Research. *JAMA Surg*. 2021;156(3):219–20.
4. de Vries M, Witteman CLM, Holland RW, Dijksterhuis A. The Unconscious Thought Effect in Clinical Decision Making: An Example in Diagnosis. *Med Decis Making*. 2010;30(5):578.
5. Jivanji D, Mangosing M, Mahoney SP, Castro G, Zevallos J, Lozano J. Association Between Marijuana Use and Cardiovascular Disease in US Adults. *Cureus*. 2020 12:e11868(12).
6. Shah S, Patel S, Paulraj S, Chaudhuri D. Association of Marijuana Use and Cardiovascular Disease: A Behavioral Risk Factor Surveillance System Data Analysis of 133,706 US Adults. *Am J Med*. 2021 May 1;134(5):614-620.e1.

7. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol.* 2015;68(9):1046–58.
8. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. *Perspect Psychol Sci.* 2016 Sep 1;11(5):702–12.
9. Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, et al. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J Psychosom Res.* 2020 Oct 1;137:110211.
10. Dongen NNN van, Doorn JB van, Gronau QF, Ravenzwaaij D van, Hoekstra R, Haucke MN, et al. Multiple Perspectives on Inference for Two Simple Statistical Scenarios. *Am Stat.* 2019 Mar 29;73(sup1):328–39.
11. Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A, et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc Natl Acad Sci.* 2020 Apr 14;117(15):8398–403.
12. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv Methods Pract Psychol Sci.* 2018 Sep 1;1(3):337–56.
13. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature.* 2020 Jun;582(7810):84–8.
14. Dutilh G, Annis J, Brown SD, Cassey P, Evans NJ, Grasman RPPP, et al. The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychon Bull Rev.* 2019 Aug 1;26(4):1051–69.
15. Fillard P, Descoteaux M, Goh A, Gouttard S, Jeurissen B, Malcolm J, et al. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *NeuroImage.* 2011 May 1;56(1):220–34.
16. Starns JJ, Cataldo AM, Rotello CM, Annis J, Aschenbrenner A, Bröder A, et al. Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Adv Methods Pract Psychol Sci.* 2019;2(4):335–49.

17. Maier-Hein KH, Neher PF, Houde J-C, Côté M-A, Garyfallidis E, Zhong J, et al. The challenge of mapping the human connectome based on diffusion tractography. *Nat Commun.* 2017;8:1349
18. Poline J-B, Strother SC, Dehaene-Lambertz G, Egan GF, Lancaster JL. Motivation and synthesis of the FIAC experiment: Reproducibility of fMRI results across expert analyses. *Hum Brain Mapp.* 2006;27(5):351–9.
19. Fox Talbot WH, Hincks E, Oppert J, Rawlinson HC. 1861. Comparative translations of the inscription of Tiglath Pileser I. *Journal of the Royal Asiatic Society of Great Britain & Ireland.* **18**:150–219. DOI: <https://doi.org/10.1017/S0035869X00013666>
20. McKenna HP. The Delphi technique: a worthwhile research approach for nursing? *J Adv Nurs.* 1994 Jun 1;19(6):1221–5.
21. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and REporting DELphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliat Med.* 2017;31(8):684–706.
22. Aczel B, Szaszi B, Sarafoglou A, Kekecs Z, Kucharský Š, Benjamin D, et al. A consensus-based transparency checklist. *Nat Hum Behav.* 2020 Jan;4(1):4–6.
23. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016 Mar 15;3:160018.
24. Schweinsberg M, Feldman M, Staub N, van den Akker OR, van Aert RCM, van Assen MALM, et al. Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organ Behav Hum Decis Process.* 2021 Jul 1;165:228–49.
25. Lundberg I, Narayanan A, Levy K, Salganik MJ. Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge. *Socius.* 2019 Jan 1;5:2378023118813023.
26. Drechsler J. Synthetic datasets for statistical disclosure control: theory and implementation. Vol. 201. Springer Science & Business Media; 2011.
27. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation.. *eLife.* 2020;9:e53275.

28. Kindel AT, Bansal V, Catena KD, Hartshorne TH, Jaeger K, Koffman D, et al. Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge. *Socius*. 2019;5:2378023118817378.
29. Parker TH, Fraser H, Nakagawa S, Fidler F, Gould E, Gould E, et al. Evolutionary Ecology Data. 2020 Mar 18 [cited 2021 Sep 28]; Available from: <https://osf.io/34fzc/>
30. Chambers CD. Registered reports: a new publishing initiative at Cortex. *Cortex*. 2013;49(3):609–10.
31. Gøtzsche PC. Blinding during data analysis and writing of manuscripts. *Control Clin Trials*. 1996 Aug;17(4):285–90; discussion 290-293.
32. Boettiger C. An introduction to Docker for reproducible research. *ACM SIGOPS Oper Syst Rev*. 2015 Jan 20;49(1):71–9.
33. Nüst D, Sochat V, Marwick B, Eglen SJ, Head T, Hirst T, et al. Ten simple rules for writing Dockerfiles for reproducible data science. *PLOS Comput Biol*. 2020 Nov 10;16(11):e1008316.
34. Liu DM, Salganik MJ. Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge. *Socius*. 2019 Jan 1;5:2378023119849803.
35. Elmenreich W, Moll P, Theuermann S, Lux M. Making simulation results reproducible—Survey, guidelines, and examples based on Gradle and Docker. *PeerJ Comput Sci*. 2019 Dec 9;5:e240.
36. Hurlin C, Pérignon C. Reproducibility Certification in Economics Research. HEC Paris Research Paper No. FIN-2019-1345. 2019 Jul 12.
37. Pérignon C, Gadouche K, Hurlin C, Silberman R, Debonnel E. Certify reproducibility with confidential data. *Science*. 2019 Jul 12;365(6449):127–8.
38. Holcombe AO, Kovacs M, Aust F, Aczel B. Documenting contributions to scholarly articles using CRediT and tenzing. *PLOS ONE*. 2020 Dec 31;15(12):e0244611.
39. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams [Internet]. 2019 Nov [cited 2021 Sep 27] p. 843193. Available from: <https://www.biorxiv.org/content/10.1101/843193v1>

40. Olsson-Collentine A, Aert RCM van, Bakker M, Wicherts J. Preprint - Meta-Analyzing the Multiverse: A Peek Under the Hood of Selective Reporting [Internet]. PsyArXiv; 2021 [cited 2021 Sep 27]. Available from: <https://psyarxiv.com/43yae/>
41. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. *Nat Hum Behav.* 2020;4(11):1208–14.
42. Liu Y, Kale A, Althoff T, Heer J. Boba: Authoring and visualizing multiverse analyses. *IEEE Trans Vis Comput Graph.* 2020;27(2):1753–63.
43. Dragicevic P, Jansen Y, Sarma A, Kay M, Chevalier F. Increasing the transparency of research papers with explorable multiverse analyses. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 2019. p. 1–15.
44. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. *Stat Sci.* 1999;14(4):382–417.
45. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol.* 1993;138(6):430–42.

#### Additional files

##### Supplementary file 1

Reporting template for multi-analyst studies.

<https://cdn.elifesciences.org/articles/72185/elifesciences-72185-suppl1-v2.docx>

##### Supplementary file 2

Reporting checklist for multi-analyst studies.

<https://cdn.elifesciences.org/articles/72185/elifesciences-72185-suppl2-v2.docx>



## 4.6. SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies<sup>52</sup>

Marton Kovacs<sup>1,2</sup>, Don van Ravenzwaaij<sup>3</sup>, Rink Hoekstra<sup>3</sup>, & Balazs Aczel<sup>1</sup>

<sup>1</sup> Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>2</sup> Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary

<sup>3</sup> University of Groningen, Groningen, The Netherlands

Author note: Marton Kovacs and Don van Ravenzwaaij are shared first authors.

---

<sup>52</sup> published as:

Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2022). SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies. *Advances in Methods and Practices in Psychological Science*, 5(1), 25152459211054060.

**Abstract**

Planning sample size often requires researchers to identify a statistical technique and to make several choices during their calculations. Currently, there is a lack of clear guidelines for researchers to find and use the applicable procedure. In the present tutorial, we introduce a web app and R package that offer nine different procedures to determine and justify the sample size for independent two-group study designs. The application highlights the most important decision points for each procedure and suggests example justifications for them. The resulting sample size report can serve as a template for preregistrations and manuscripts.

*Keywords:* sample size determination; power analysis; study design

**Introduction**

Social and behavioral sciences are known to be plagued by undersampling (Ioannidis, 2005). In the traditional statistical framework, even when the effect exists, undersampled studies yield either nonsignificant results or significant results due to overestimating the size of the effect. Because nonsignificant results are less likely to reach publications than significant ones, results of undersampled studies either remain unpublished or impose a substantial bias on our body of published empirical findings. In addition, the low informational value of undersampled studies may not justify the cost or potential risk they induce (Halpern, Karlawish, & Berlin, 2002). To mitigate these issues, authors are increasingly expected to plan and justify the sample size of their study (Maxwell, 2004). However, such sample size justifications are only meaningful if they provide sufficient information to the readers to judge the adequacy of the author's decisions.

In the statistical literature, a few methods have been proposed to determine and justify sample size. In practice, however, authors are short of practical guides on how to navigate among the different sample size methods. The aim of our tutorial is to point out for each method the essential decision points that a researcher has to face during this process. We provide a short description of each method and the corresponding parameters, but we avoid listing their advantages and disadvantages. As there are disagreements between the experts of the field regarding the correct use of some of the methods, we intentionally try to remain impartial and do not favor any of the presented methods. Researchers who want to know more about each method can find a number of useful references in the description of the methods. We also provide a collection of ready-to-use analysis code and a ShinyApp that helps researchers use and report the main sample size estimation techniques for different scenarios. The tutorial is

focused exclusively on the scenario of the comparison of two independent groups (i.e., the independent  $t$ -test design) with a one-sided test.

### Sample Size Determination and Justification

A lot of factors go into the determination of the sample size for an independent two group study design. In this section, we will first provide a birds-eye view of the most important decisions. Next, we will go into more detail on the specific inference tool that results from the combination of the larger choices.

It is crucial to not just state how we determined our planned sample size but to also give the reader insight into the reasons behind our choices. In a recent overview, Lakens (2021) lists six types of general approaches to justify sample size in quantitative empirical studies: (1) Measure entire population; (2) Resource constraints; (3) A-priori power analysis; (4) Accuracy; (5) Heuristics; and (6) No justification. For the first approach, no quantitative justification is necessary; and for the second approach, the researcher has no freedom to increase the sample size. Power analysis, or more generally the estimation of true positive rate, is used when one plans to conduct hypothesis testing; accuracy justifications are used when one plans to conduct parameter estimation. Our tutorial mainly focuses on approaches two, three, and four, and is aimed at providing a hands-on approach for the mechanical part of the sample size determination (i.e., the calculation). For a deeper discussion of justification of these approaches, or for other approaches (i.e., using heuristics or not providing justification), we refer the reader to Lakens (2021).

### Choosing a method in case of sample size justification

In an ideal world, the choice for the number of participants would be solely determined by scientific considerations and depending on the chosen technique the collection of data would continue until either the desired sample size or a desired outcome has been reached. In practice, researchers are limited by time (collecting data is quite demanding), money (participants or people collecting the data may be paid, and the same may hold for renting space or equipment), or availability of participants (the population may be relatively small, and/or the participation rate quite low).

When constrained by limited resources, it is important to be transparent about those limitations. It is also important to be open about scientific considerations. Depending on the nature of the study (perhaps it is an initial exploration?), small sample sizes need not be a dealbreaker. So

although more data are always preferred from an informational point of view, by owning the limitations of our study, we improve future readers' understanding of the process leading up to the eventual paper, and we also answer in advance to those who think the chosen sample size was insufficient.

Whether or not authors have limited resources, two important choices need to be made: (1) whether they are interested in *statistical testing* or in *parameter estimation*; and (2) whether they want to conduct their statistical inference within the *frequentist* framework or within the *Bayesian* framework. Starting with the first decision, statistical testing is the primary framework when one is interested in establishing whether an underlying population effect is equal to, different from, larger than, or smaller than a certain value. In essence, statistical testing lends itself to binary decision making. Typically, testing is concerned with a fixed point null hypothesis (e.g., there is no difference between two groups), although using intervals for testing is also possible. Alternatively, one might be interested in parameter estimation that is less interested in establishing the existence of a difference and instead is concerned with establishing the magnitude of the difference.

The second important decision concerns the statistical framework. Choosing to conduct statistical tests within a frequentist framework, one is usually interested in balancing the type I (false positive) and type II (false negative) error rates. Practitioners choosing to conduct statistical tests within a Bayesian framework are typically interested in being able to quantify the relative probability of hypotheses or models being true given the data and in including prior information.

Within the realm of statistical testing, there are some other factors that affect the preferred inference tool: Do you prefer to test for equivalence (no difference in mean) or for superiority (mean of one group larger than mean of other group), are you interested in calculating a required sample size for a specific hypothetical effect size or for a range of possible values, and do you wish to employ sequential testing (applicable to Bayesian testing)? In case of testing, some of the methods are designed to find support for the null hypothesis (e.g., TOST, ROPE), while others are designed to find support for the alternative hypothesis (e.g., traditional null hypothesis testing), and some methods are designed to find support for either (e.g., BFDA). For frequentist estimation, the preferred inference tool might differ depending on whether we evaluate uncertainty for each group separately or jointly. We will describe these specific factors when we go into detail about each of the preferred methods. A flow-chart representing all of these choices is given in Figure 1.

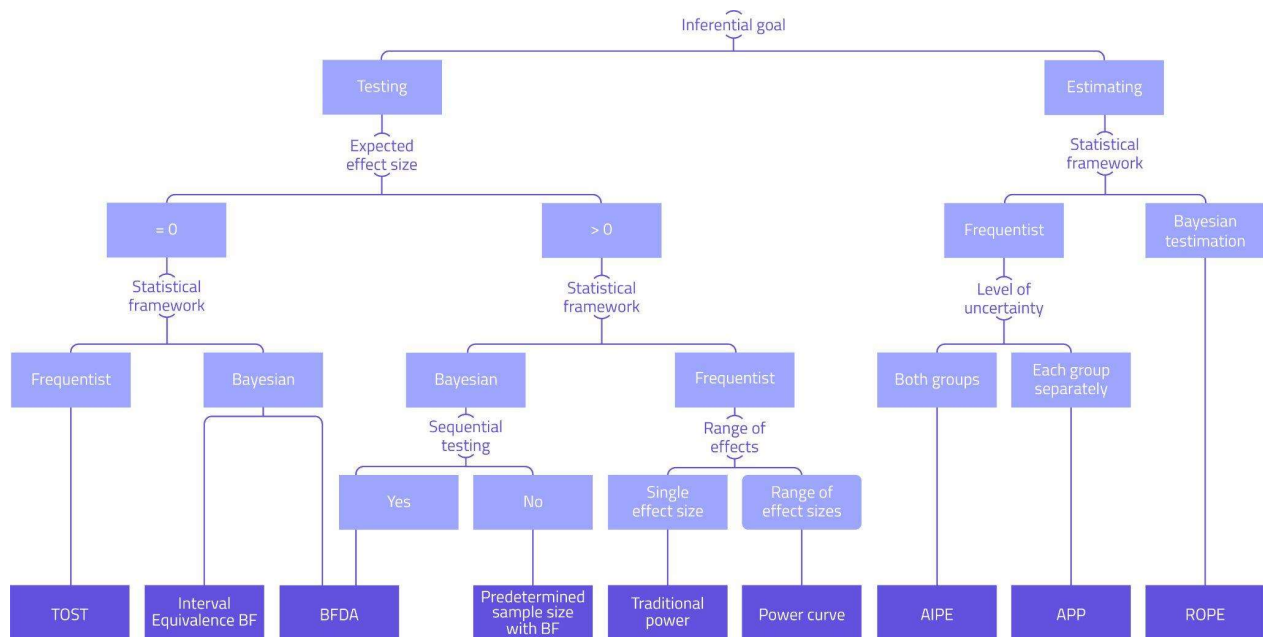


Figure 1. The figure depicts the decisions that one faces when choosing among sample size estimation methods. The nine sample size estimation methods discussed in this paper are listed in the bottom row. Some decisions are determined by the investigated question and the design of the study while others are based on the preferred statistical framework.

### How to use this guide

In the next section, we will illustrate the specific inference tools and resulting sample size calculations in more detail using a ShinyApp and an R package we have developed. Throughout this section, we recurrently use two terms that have different meanings for different techniques. These are the *true positive rate* (TPR), and the *equivalence band* (EqBand). The TPR reflects the long-run probability of concluding there is an effect, given that it does exist. For traditional null hypothesis testing, this is typically referred to as power, but related concepts exist for different inference tools. The EqBand refers to an effect size region, typically around zero, that is deemed clinically insignificant or irrelevant. Different names are given to this region depending on the technique that employs them, such as statistical effect size of interest (SESOI) or region of practical equivalence (ROPE). For both TPR and EqBand, we explain the specific meaning in context of the relevant inference tool below.

For each method, only the main parameters can be adjusted with a certain range of values in the ShinyApp by using a slider. These parameters are presented in the text in bold. Other parameters are set to preset values in the application but can be adjusted in the accompanying R package to any sensible value. These parameters are highlighted in italics in the tutorial. Both

the app and the package allow the users to save or copy a text template with the results of the sample size determination. We offer a list of possible justifications at the decision points for each method (indicated between square brackets), but users are able to provide their own justification as free-text. It is important to note that the listed justifications are meant to provide guidance for the user, and they are not sufficient without further details provided by the researcher in the context of the given study. For example, previously reported values should always be accompanied by a theoretical justification of why these values make sense. The provided justification text could serve as a stub for the description of the chosen sample size in a paper, a preregistration or registered report, or a grant proposal.

Throughout, we will use the example story of Mary the educational psychologist. Mary has come up with a new set of games that challenge spatial insight. She would like to test whether distributed and targeted engagement with these games for a period of six months for children in the age range of 8 to 12 will lead to lasting improvements on their IQ score as measured through Raven's progressive matrices test (population mean 100, population SD 15). Mary collects data for a control sample that gets regular education and for an experimental sample and plans to compare those samples. Mary has good reason to be skeptical about the effectiveness of training on increasing performance as there are several studies questioning the existence of such effects (Owen et al., 2010; Simons et al., 2016). For illustrative purposes, in some of the upcoming examples Mary expects a null effect and in others Mary expects a positive effect in order to highlight the different research scenarios for each sample size planning method. We will also present a justification text for each sample size planning method based on Mary's choices described in the example research scenario for the given method.

The ShinyApp is available on <https://martonbalazskovacs.shinyapps.io/SampleSizePlanner> and the R package can be installed by running the following command in R devtools::install\_github("marton-balazs-kovacs/SampleSizePlanner"). There is more information about the R package and the ShinyApp on the projects' Github page <https://github.com/marton-balazs-kovacs/SampleSizePlanner>, or on the website <https://marton-balazs-kovacs.github.io/SampleSizePlanner/>.

## 1. Testing

### 1.1. Effect size = 0.

#### 1.1.1. Two One-Sided Tests (TOST).

Study context.

Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is practically equivalent to the mean IQ score of the control group. She tests this assumption in a frequentist framework, and considers a population effect size between -0.2 and 0.2 to be 'practically equivalent' to no difference. This would correspond to IQ scores between 97 ( $100+15*(-.2)$ ) and 103 ( $100+15*.2$ ).

Description.

TOST is a frequentist equivalence testing approach that adopts two one-sided hypotheses to designate an interval hypothesis (Schuirmann, 1987). The lower and upper boundaries of the interval are determined by the equivalence band (i.e. SESOI) around the expected population effect size (e.g., 0). Lakens, Scheel, and Isager (2018) lists several methods that can be used to determine the SESOI. In case of TOST, the two null hypotheses state that the effect size is equal to the lower and upper equivalence band values, whereas the alternative hypotheses state that the effect size is significantly smaller than the upper equivalence band value and significantly larger than the lower equivalence band value. In case both one-sided tests reject the null-hypothesis at a given significance level, the group means are considered to be practically equivalent. See Lakens, Scheel, and Isager (2018), for further reading.

Parameters.

**Delta:** The expected population effect size. In most cases, this value will be zero.

**TPR:** The desired long run probability of obtaining a significant result with TOST, given Delta.

**EqBand:** The chosen width of the region for practical equivalence, i.e. the SESOI.

*Alpha:* The level of significance. The alpha level in the application is preset to 0.05.

How to use the package.

```
SampleSizePlanner::ssp_tost(tpr = 0.8, eq_band = 0.2, delta = 0)
```

How to report your sample size estimation.

In order to calculate an appropriate sample size for testing whether the two groups are practically equivalent, we used the Two One-Sided Tests of Equivalence [TOST; Schuirmann (1987)] method. We used an alpha of 0.05. We set the aimed TPR to be 0.8, because [1] it is

the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero, because [1) previous studies reported the choice of a similar equivalence band; 2) of the following substantive reasons: ...]. The expected delta was 0 because [1) we expected no difference between the groups]. Based on these parameters, a sample size of 429 per group was estimated in order to reach a TPR of 0.8 with our design.

### 1.1.2. Equivalence interval Bayes factor.

#### Study context.

Mary would like to know what sample size she needs to have a long-run probability of .80 of obtaining a Bayes factor larger than 10. Mary would like to test whether the mean IQ score of the experimental group's population is practically equivalent to the mean IQ score of the control group. Mary hypothesizes that there is no difference (i.e.,  $H_0$  is true). Mary tests this assumption in a Bayesian framework. Mary considers a population effect size between -0.2 and under 0.2 to be 'practically equivalent.' This would correspond to IQ scores between 97 ( $100+15*(-.2)$ ) and 103 ( $100+15*.2$ ).

#### Description.

Equivalence interval Bayes factors contrast an equivalence hypothesis to a non-equivalence hypothesis and quantify the evidence with Bayes factors. Typically,  $H_0$  constitutes the equivalence interval (comparable to SESOI in the TOST framework), and  $H_a$  constitutes the complementary non-equivalence regions. Formally, the Bayes factor is calculated by dividing the fraction *posterior area inside the interval/posterior area outside the interval* (i.e., the posterior odds) by the fraction *prior area inside the interval/prior area outside the interval* (i.e., the prior odds). The resulting value quantifies how much more likely it is that the data occurred under a population effect size deemed 'equivalent' relative to the data having occurred under a population effect size deemed non-equivalent. The current implementation uses a default Cauchy prior on effect size with the possible scale parameters of medium ( $r = 1/\sqrt{2}$ ), wide ( $r = 1$ ), or ultra-wide ( $r = \sqrt{2}$ )  $\sqrt{2}$ . For further reading, see Morey and Rouder (2011), van Ravenzwaaij, Monden, Tendeiro, and Ioannidis (2019), and Linde, Tendeiro, Selker, Wagenmakers, and van Ravenzwaaij (2020).

#### Parameters.

**Delta:** The expected population effect size.

**TPR:** The desired long-run probability of obtaining a Bayes factor at least as high as the



Threshold, given Delta.

**EqBand:** The chosen width of the equivalence region.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to:  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

*Threshold:* Critical threshold for the Bayes factor. The threshold level in the application can be set to 10, 6, or 3.

How to use the package.

```
SampleSizePlanner::ssp_eq_bf(tpr = 0.8, delta = 0, eq_band = 0.2,
  thresh = 10, prior_scale = 1/sqrt(2))
```

How to report your sample size estimation.

In order to estimate the sample size, we used the interval equivalent Bayes factor (Morey & Rouder, 2011; Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019) method. We used a Cauchy prior distribution centered on zero with a scale of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8, because [1) it is the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero, because [1) previous studies reported the choice of a similar equivalence region; 2) of the following substantive reasons: ...]. The expected delta was 0 because [1) we expected no difference between the groups]. Our Bayes factor threshold for concluding equivalence was 10. Based on these parameters, a minimal sample size of 144 per group was estimated in order to reach 0.8 TPR for our design.

1.2. Effect size  $>0$ .

1.2.1. Frequentist.

1.2.1.1. Classical power analysis.

Study context.

Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is significantly higher than the mean IQ score of the control group. She tests this assumption in a frequentist framework for a hypothetical population effect size of 0.5. This corresponds to a mean IQ score of 107.5 in the experimental group ( $100+15*.5$ ), assuming a mean IQ score of 100 in the control group.

Description.

The classical power analysis approach allows one to calculate the required sample size in order to obtain a significant result for the null hypothesis test a certain proportion of times in the long run given an assumed population effect size.

Parameters.

**Delta:** The expected population effect size.

**TPR:** The desired long-run probability of obtaining a significant result with a one-sided  $t$ -test, given Delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

*Alpha:* The level of significance. Alpha is preset to 0.05 in the application.

How to use the package.

```
SampleSizePlanner::ssp_power_traditional(tpr = 0.8, delta = 0.5,  
max_n = 5000, alpha = 0.05)
```

How to report your sample size estimation.

We used a power analysis to estimate the sample size. We used an alpha of 0.05. We set the aimed TPR at 0.8, because [1) it is the common standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1) previous results published in ...; 2) of the following substantive reasons: ...]. Based on these parameters, a minimal sample size of 51 per group was estimated in order to reach 0.8 TPR for our design.

1.2.1.2. Power curve.

Study context.

Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is significantly higher than the mean IQ score of the control group. She tests this assumption in a frequentist framework. However, she is reluctant to commit to a single hypothetical population effect size a-priori, preferring to calculate required sample size for a range of hypothetical deltas between 0.1 and 0.9.

Description.

The power curve method is similar to a classical power analysis but instead of calculating the appropriate sample size for one hypothesized population effect size, the method calculates the required sample size for a range of plausible population effect sizes.

Parameters.

**Delta:** A range of hypothetical population effect sizes.

**TPR:** The desired long-run probabilities of obtaining a significant result with a one-sided  $t$ -test, given each value of Delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

*Alpha:* The level of significance. Alpha is preset to 0.05 in the application.

How to use the package.

```
# Determine the sample sizes for each delta
```

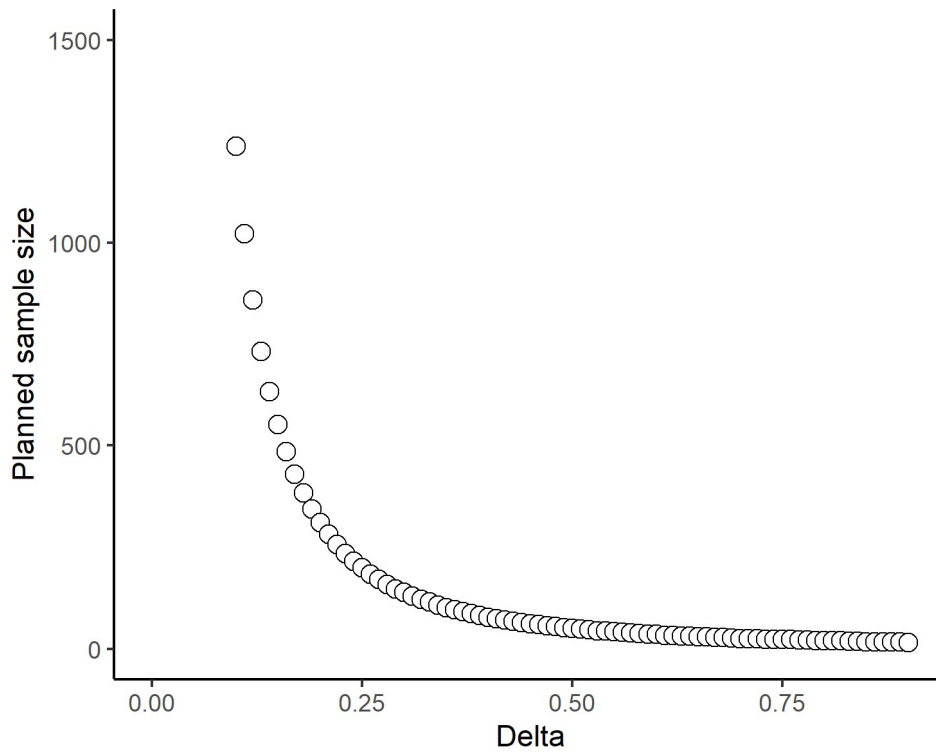
```
curve_data <- SampleSizePlanner::ssp_power_curve(tpr = 0.8, delta = seq(0.1,  
0.9, 0.01), max_n = 5000)
```

```
# Plot the power curve
```

```
SampleSizePlanner::plot_power_curve(delta = curve_data$delta,  
n1 = curve_data$n1, animated = FALSE)
```

How to report your sample size estimation.

We used a power analysis to estimate the sample size. We used an alpha of 0.05. We set the aimed TPR at 0.8, because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. Because [1] we have no clear expectation of the magnitude of delta 2) we expected the delta to be around...], we include power calculations for delta ranging from 0.1 to 0.9. Based on these parameters, minimal sample sizes per group for different hypothetical effect sizes to reach 0.8 TPR can be found in Figure 2.



*Figure 2.* The figure shows the resulting power curve created by the application. The X-axis shows the range of deltas from the example, while the Y-axis shows the corresponding sample sizes determined by the power curve method.

### 1.2.2. Bayesian.

#### 1.2.2.1. Predetermined sample size with Bayes factor.

##### Study context.

Mary would like to test whether the mean IQ score of the experimental group's population is higher than the mean IQ score of the control group. She'd like to know what sample size she needs to have for a long-run probability of .80 of obtaining a Bayes factor larger than 10. Mary plans to collect all her data in one batch without testing sequentially. Mary expects the population effect size to be 0.5. This corresponds to a mean IQ score of 107.5 ( $100+15 \cdot .5$ ) in the experimental group, assuming a mean IQ score of 100 in the control group.

##### Description.

The present method calculates the corresponding default Bayes factor for a  $t$ -test statistic with Cauchy prior distribution centered on zero with scale parameter of either  $1/\sqrt{2}$ , 1, or  $\sqrt{2}$  for several sample sizes (the so-called Jeffrey-Zellner-Siow Bayes factor, see e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009). The function returns the optimal sample size needed

to reach the TPR for a given Bayes factor threshold to detect an expected population effect size. If a range of possible population effect sizes are plausible under the given hypothesis, the function can calculate the optimal sample sizes for the given range of effect sizes and present the results in a figure (analogous to the Power Curve method). This method is designed to determine the sample sizes for the existence of an effect (i.e.,  $\Delta > 0$ ).

Parameters.

**Delta:** The expected population effect size or a range of expected effect sizes.

**TPR:** The long-run probability of obtaining a Bayes factor at least as high as the critical threshold favoring superiority, given Delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to:  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

**Threshold:** Critical threshold for the Bayes factor. Three threshold levels are available in the app: 3, 6, and 10.

How to use the package.

```
SampleSizePlanner::ssp_bf_predetermined(tpr = 0.8, delta = 0.5,
  thresh = 10, max_n = 5000, prior_scale = 1/sqrt(2))
```

How to report your sample size estimation.

We used the Jeffrey-Zellner-Siow Bayes factor method to estimate the sample size. We used a Cauchy prior distribution centered on zero with a scale of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8, because [1) it is the common standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1) previous results published in ...; of the following substantive reasons: ...]. Our evidence threshold was 10. Based on these parameters, a minimal sample size of 105 per group was estimated in order to reach a 0.8 TPR for our design.

#### 1.2.2.2. Bayes Factor Design Analysis (BFDA).

Study context.

Mary would like to know what sample size she needs to have a long-run probability of .80 of obtaining a Bayes factor larger than 10. Mary would like to test whether the mean IQ score of

the experimental group's population is higher than the mean IQ score of the control group in a Bayesian framework. Mary plans to collect all her data incrementally and as such is interested in using the advantage of not testing more than strictly necessary offered by sequential testing in her Bayesian analysis. Mary expects the population effect size to be 0.5. This corresponds to a mean IQ score of 107.5 in the experimental group ( $100 + 15 * .5$ ), assuming a mean IQ score of 100 in the control group.

#### Description.

The description of the BFDA method is functionally identical to the one provided in section 'Predetermined sample size with Bayes factor,' but gains in TPR due to the addition of sequential testing. In the app,  $H_0$  and  $H_a$  indicate the proportion of times sequential testing leads to Bayes factors providing evidence with the given threshold for the null hypothesis and for the alternative hypothesis, respectively. Users of the Shiny app and R package should set Delta to 0 if they wish to determine the sufficient sample size for rejecting an effect, and use  $\Delta > 0$  if they wish to find support for the existence of an effect. For further reading, see Schönbrodt and Wagenmakers (2018) and Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017).

#### Parameters.

**Delta:** The expected population effect size.

**TPR:** The long run probability of obtaining a Bayes factor at least as high as the critical threshold favoring superiority, given Delta.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to:  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

**Threshold:** Critical threshold for the Bayes factor. Three threshold levels are available in the app: 3, 6, and 10.

#### How to use the package.

```
SampleSizePlanner::ssp_bfda(tpr = 0.8, delta = 0.5, thresh = 10,
  n_rep = 1000, prior_scale = 1/sqrt(2))
```

#### How to report your sample size estimation.

We used the BFDA method to estimate the sample size. We used a Cauchy prior distribution centered on zero with a scale of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8, because [1] it is the common

standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1) previous results published in ...; 2) of the following substantive reasons: ...]. Our evidence threshold was 10. Based on these parameters, a minimal sample size of 81 per group was estimated in order to reach a 0.8 TPR for our design.

## 2. Estimation

### 2.1. Frequentist.

#### 2.1.1. Accuracy In Parameter Estimation (AIPE).

##### Study context.

Mary would like to know what sample size she needs, such that the 95% confidence interval for the population effect size has an expected width of 0.4. She estimates the population effect size to be 0.2.

##### Description.

Accuracy in parameter estimation aims to determine the sufficient sample size to obtain a confidence interval with a desired width (precision) around the expected effect size (Kelley & Rausch, 2006). Note that the width of the calculated confidence interval will depend on the sample variance. As a result, it is possible that for a given sample the variance is relatively large, leading to a resulting confidence interval that is larger than the width of the desired interval for a given sample. Thus, the AIPE method aims to establish the expected value of the calculated confidence interval, which can be thought of as the 50% long-run probability of obtaining a confidence interval no wider than the provided width.

##### Parameters.

**Delta:** The expected population effect size.

**Width:** The desired width of the confidence interval, given Delta.

**Confidence level:** The desired level of confidence.

##### How to use the package.

```
SampleSizePlanner::ssp_aipe(delta = 0.5, width = 0.2, confidence_level = 0.8)
```

##### How to report your sample size estimation.

In order to estimate the sample size, we used the accuracy in parameter estimation [AIPE; Kelley and Rausch (2006)] method. We aimed for a 95% confidence level, because [1) it is the common standard in the field; 2) it is the journal publishing requirement]. The desired width was 0.4 because [1) previous studies reported the choice of a similar region of practical equivalence; 2) of the following substantive reasons: ...]. We expected an underlying population effect size of 0.3, because [1) previous results published in ...; 2) of the following substantive reasons: ...]. Based on these parameters, a minimal sample size of 195 per group was estimated for our design.

### 2.1.2. A Priori Precision (APP).

Study context.

Mary would like to know the sample size for which she will have a 95% long-run probability that the sample means in both the experimental and the control group lie within 0.2 standard deviations (3 IQ points) of the true population mean.

Description.

APP aims to determine the sample size needed to have a certain long-run probability of both sample means being within a certain range of their respective population means, expressed in terms of standard deviations (Trafimow & MacDonald, 2017). As a result, APP is not reliant on the expected effect size.

Parameters.

**Closeness:** The desired closeness of the sample mean to the population mean defined in standard deviation.

**Confidence:** The desired probability of obtaining the sample mean with the desired closeness to the population mean.

How to use the package.

```
SampleSizePlanner::ssp_app(closeness = 0.2, confidence = 0.95)
```

How to report your sample size estimation.

In order to estimate the sample size, we used the a-priori precision [APP; Trafimow and MacDonald (2017)] method. Before data collection, we wanted to be 95% confident that both



sample means lie within 0.2 SD of the true population means. Based on these parameters, the resulting minimum sample size was 126 per group for our design.

## 2.2. Bayesian estimation.

### 2.2.1. Region of Practical Equivalence (ROPE).

#### Study context.

Mary would like to conduct parameter estimation to see whether the mean IQ score of her experimental group's population is practically equivalent to 100. She would like to know what sample size she needs to have a long-run probability of .80 of obtaining a 95% highest density interval that is contained within her predefined region of practical equivalence (ROPE). Mary hypothesizes that there is no difference (i.e.,  $H_0$  is true). She considers a population effect size between -0.2 and under 0.2 to be 'practically equivalent.' This would correspond to IQ scores between 97 ( $100+15*-.2$ ) and 103 ( $100+15*.2$ ).

#### Description.

The highest density interval region of practical equivalence technique (HDI-ROPE, often just referred to as ROPE) shares some features with the equivalence interval Bayes factor procedure. Both define an equivalence interval, construct a prior for the population effect size, and update to a posterior after the data comes in. The equivalence interval Bayes factor procedure then focuses on the posterior and prior odds under complementary hypotheses. The ROPE procedure, on the other hand, identifies the 95% highest density interval (HDI; other percentages are permissible as well) and determines whether or not the HDI is fully contained within the equivalence interval. For further reading, see Kruschke (2018) and Kruschke (2011).

#### Parameters.

**Delta:** The expected population effect size.

**TPR:** The desired long run probability of having the HDI fully contained within the ROPE interval, given Delta.

**EqBand:** The chosen ROPE interval.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to:  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

How to use the package.

```
SampleSizePlanner::rope(tpr = 0.8, delta = 0.5,  
  thresh = 10, max_n = 5000, prior_scale = 1/sqrt(2))
```

How to report your sample size estimation.

In order to estimate the sample size, we used the Region of Practical Equivalence (Kruschke, 2018) method. We used a Cauchy prior distribution centered on zero with a scale of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8, because [1) it is the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero, because [1) previous studies reported the choice of a similar region of practical equivalence; 2) of the following substantive reasons: ...]. The expected delta was 0 because [1) we expected no difference between the groups]. Based on these parameters, a minimal sample size of 517 per group was estimated in order to reach a 0.8 TPR for our design.

### Summary

Justifying the decisions made during the sample size planning process presents valuable information when one evaluates the inferences drawn from a study. The Shiny app and R package presented in this paper aim to help researchers to choose and employ their sample size estimation method. In addition, the tool provides assistance in reporting the process and justification behind sample size choices. We encourage users and experts of the field to provide feedback and recommendations towards further developments.

### Authors contribution

**Conceptualization:** Marton Kovacs, Don van Ravenzwaaij, Rink Hoekstra, and Balazs Aczel.

**Methodology:** Don van Ravenzwaaij.

**Project Administration:** Marton Kovacs.

**Software:** Marton Kovacs and Don van Ravenzwaaij.

**Supervision:** Balazs Aczel.

**Writing - Original Draft Preparation:** Marton Kovacs, Don van Ravenzwaaij, Rink Hoekstra, and Balazs Aczel.

**Writing - Review & Editing:** Marton Kovacs, Don van Ravenzwaaij, Rink Hoekstra, and Balazs Aczel.

### Notes

## Glossary

*Accuracy in Parameter Estimation (AIPE):* A sample size estimation method used for parameter estimation. The approach aims to find the required sample size, such that the confidence interval has a certain expected width.

*Priori Procedure (APP):* The approach aims to plan a sample size based on how close the researcher wishes both sample means to be to their respective population parameter, and how confident the researcher wants to be in this.

*Bayesian inference:* A general framework for updating one's prior beliefs in light of new data.

*Bayes Factor Design Analysis (BFDA):* This technique provides an expected sample size such that compelling evidence in the form of a Bayes factor can be collected for a given effect size with a certain long-run probability when allowing for sequential testing.

*Testing vs. Estimation:* Two schools of inference, focusing on establishing whether or not an effect exists versus establishing the magnitude of an effect, respectively.

*Equivalence band (EqBand):* The region of effect sizes considered practically equivalent to zero. In our paper, SESOI and ROPE are subsumed under EqBand.

*Frequentist inference:* A general framework in which probabilities are defined as frequencies in hypothetical repeated events. In the context of statistical testing, frequentist inference is concerned with long-run error rates of rejecting the null hypothesis for the observed or more extreme parameters in a given design when the model assumptions (e.g., independence of observations) are true.

*Statistical power:* The long-run probability of finding a significant effect given a certain population effect size.

*True positive rate (TPR):* The long-run probability of finding evidence for an effect, given that it exists. In our paper, statistical power is subsumed under TPR.

*Classical power analysis:* This method is used to estimate the minimum sample size that a design needs to reach a certain level of statistical power, given a desired significance level and expected effect size.

*Power-curve:* This curve shows how changes in effect size modify the statistical power of a test.

*Region Of Practical Equivalence (ROPE):* The region of effect sizes considered practically equivalent to zero under the HDI-ROPE method.

*Smallest Effect Size Of Interest (SESOI):* The region of effect sizes considered practically equivalent to zero under the TOST method.

*Sequential testing:* The practice of incrementally testing as data comes in, typically until some pre-determined level of evidence is obtained.

*Two One-Sided Tests (TOST):* A frequentist statistical testing approach aimed at establishing equivalence between two groups.

*Equivalence interval BF:* A Bayesian statistical testing approach aimed at establishing equivalence between two groups.

## References

- Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3), 358–362.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11(4), 363–385.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3), 299–312.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
- Lakens, D. (2021). Sample size justification. Retrieved from <https://doi.org/10.31234/osf.io/9d3yf>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259–269.
- Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E., & van Ravenzwaaij, D. (2020, November 10). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. <https://doi.org/10.31234/osf.io/bh8vu>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406–419.
- Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, 465(7299), 775–778.
- Ravenzwaaij, D. van, Monden, R., Tendeiro, J. N., & Ioannidis, J. P. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, 19(1), 1–12.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237.
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142.
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339.
- Schuirman, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680.
- Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do “brain-training” programs work?. *Psychological Science in the Public Interest*, 17(3), 103–186.
- Trafimow, D., & MacDonald, J. A. (2017). Performing inferential statistics prior to data collection. *Educational and Psychological Measurement*, 77(2), 204–219.

## 5. SUMMARY AND CONCLUSIONS

### Summary

The aim of this thesis was to demonstrate that psychologists can play an important role in the development of metascience and their perspective and methodology are indispensable for the understanding and improvement of science. The present work focused on three lines of studies conducted on the following metascientific topics: (1) *Problems in the publication practice*; (2) *Lack of transparency*; and (3) *Issues in statistical practice*. Table 2 and the present chapter provide an overview to the studies presented in this thesis.

**Table 2** Summary Table of the Studies Presented in this Thesis

Chapter	Topic	Output
2.1.	Researchers' contribution to the publication system through peer review	Global estimates of reviewers' time and salary-based contribution to publishers
2.2.	Documenting authorship information in scholarly articles	A web-application to assist authors in collecting and reporting required authorship and funding information
2.3.	Researchers working from home	Survey data on the benefits and challenges of home-office
3.1.	Reporting transparency in social sciences	A checklist and application to report transparency-related aspects of studies
3.2.	Researchers' experience with preregistration	Survey data on how preregistration can help the workflow of empirical studies
3.3.	Transparency practices in statistics	Seven recommendations to promote transparency in statistical practice
3.3.1.	Demonstration of good transparency and statistical practices in psychology	A preregistered multi-lab replication project on a moral dilemma paradigm
4.1.	Surveying mistakes in research data management	Identification of the most frequent and most serious data management mistakes
4.2.	The strength of evidence for the null hypothesis in psychology	Quantification of the evidence in non-significant results in psychological studies
4.3.	how to conduct and report Bayesian analyses	Consensus-based thinking guideline and reporting template for Bayesian analyses
4.4.	The importance of exploring alternative statistical analyses	Discussion of the results of multi-analyst studies in various disciplines
4.5.	Conducting and reporting multi-analyst studies	Consensus-based guidance on how to prepare and run multi-analyst studies
4.6.	Estimating and justifying sample sizes for two-group studies	An application for calculating and justifying required sample sizes

*Contributions to the ‘problems with the publication practice’ topic*

The study of **Chapter 2.1.** (Aczel, Szaszi, & Holcombe, 2021), highlights a rarely discussed expense of the academic publication system, the time and effort that researchers spend with reviewing journal articles. This labour is generally unrecognised. Most reviewers complete their job anonymously without any monetary compensation for their work. Although peer review is the gate keeper of science, reviewers give science’s certification to submission, from the side of the publishers, this is free labour. How much free labour goes to the publication system every year? –was the question of our investigation. Since in the publication system not all numbers are available on a global scale, we had to work with rough estimations, but in cases of uncertainty, we always used conservative estimates. With the use of average rejection and acceptance rates, we could calculate the required values from the known number of accepted articles. With this methodology, we estimated that for the year 2020, there must have been around 4.7M accepted submissions and 3.8M rejections after peer review. Working with 3 reviews per eventually accepted submission and 2 reviews per eventually rejected submissions, we estimated 21.8M was the number of reviews that year. Previous estimations indicate that the average time that reviewers spend on a manuscript is 6 hours. These estimates produce over 130M hours for the total time that reviewers spent on peer review in 2020, which is equivalent of ~15 thousand years. For the academically most productive countries (USA, China, UK), we calculated the salary-based contribution to journal peer review system. For the USA, this was over 1.5 billion USD. For China-based reviewers, the estimate was over 600 million USD, for UK-based reviewers it was close to 400 million USD. Since we had to work with rough estimates, the true values might be somewhat different from our figures. Nevertheless, the study managed to show the magnitude of the cost of peer review on the reviewers’ side. In addition to our calculations, we list a number of initiatives that could improve this arrangement and reduce inefficiencies in the system. With these estimates, we aim to draw attention to the mostly unrecognised contribution of reviewers, and we encourage an open discourse between the scientific community and the stakeholders for a fairer and more efficient publication system.

The scope of the study of **Chapter 2.2.** (Holcombe et al., 2020), is on how to decrease the waste in researchers’ time when trying to publish an article. Our paper presents a tool that we developed to help researchers with the preparation of their manuscript submission. When preparing a submission, authors have a number of tasks to complete. They have to format the text according to the requirements of the journal, add all the authors and their affiliation to the front page, create a section for the acknowledgement of fundings etc. An additional expectation

in an increasing number of journals is to document the contributions of each author. As the average number of authors has increased on scientific papers throughout the year (Fanelli & Larivière, 2016; Regalado, 1995), it became harder and harder to keep track of the ‘who did what’ and to report all these information. The solution that we developed is an app, *tenzing* (Figure 2), that helps teams keep track of their activities following the *Contributor Role Taxonomy* (CRediT, Allen et al., 2014) and to collect all the required information of the authors.

### 1. Create your contributors table


Duplicate and edit the [contributors table template](#)

### 2. Load your contributors table

Local file    URL

Choose the spreadsheet on your computer

Browse...    No file selected

Use the spreadsheet 

### 3. Download the output

Show author contributions text

Show author list with affiliations

Show XML file (for publisher use)

Show *papaja* YAML

Show funding information

Figure 2. A screenshot of the *tenzing* app. Available from: <https://tenzing.club/>

We put great effort into making the procedure as simple as possible. As a first step, authors can record the team’s contributions in a Google Spreadsheet template that is available from the app. In this template they can record the name, affiliations, funding information, email address etc. of each author and tick the CRediT category to which they contributed. Once all the information is added, one can easily generate the front page of the paper, the funding section, or the authors’ contribution section with one click. As a last step, the generated text can be added to the manuscript anytime. Since its publication, the application gained widespread popularity and incentivised the development of similar solutions in academic practice (Kovacs, Holcombe, et al., 2021).

Another way to increase researchers' efficiency when preparing a publication is to find the right environment for their work. In the past, for researchers, as for other workers, the workplace was the dedicated place for completing all their work-related activities. In other words, the work-life boundary was well-defined by time (working time vs. free time) and location (office vs. home). The technological mobilisation that has been accelerated by the easy availability of home computers, laptops, continuous access to stronger internet etc. decreased the necessity of this arrangement and for most professionals, working from home became an option. In **Chapter 2.3.** (Aczel, Kovacs, et al., 2021), we explored researchers' perspective on which location supports better different aspects of their research work (e.g., analysing data; working on the manuscript). The Covid pandemic-related lockdowns provided unique opportunities to explore this question as researchers in great numbers had to experience the benefits and challenges of working from home. One of our key findings is that while the lockdown decreased the efficiency of half of our respondents, around a quarter of them found working from home more efficient than working from the office. Importantly, 70% of them thought that after the pandemic they would find working from home beneficial for their research work. Of course, the advantage of remote working depends a lot on the circumstances. For example, people living with children, especially single parents found working from home less efficient when the children are also at home. Another level of differentiation comes when we look at the various aspects of research work. For example, our respondents reported that certain aspects of their work are, on average, still more efficient in the office, for example: 'sharing thoughts with the colleagues', 'collecting data', 'keeping in touch with the team'. We conclude that the pandemic only accelerated the disintegration of the traditional work-life boundary and researchers are moving towards a hybrid arrangement between their office and their home as the location of their work. We emphasise that academics need to develop skills and tactics for managing the boundary between work and personal life in these new arrangements giving efficiency as well as personal wellbeing due respect.

### *Contributions to the 'lack of transparency' topic*

In the study of **Chapter 3.1.** (Aczel, Szaszi, et al., 2020), we introduced a checklist to improve and document the transparency of research reports in social sciences. Although, some specialised reporting guidelines exist, but they are not comprehensive to research fields and the different aspects of transparency. For the development of the checklist, we recruited 45 behavioural and social science journal editors-in-chief and associate editors, as well as 18 open-science advocates to serve as members of our expert panel. The content of the checklist was



developed using a preregistered Delphi expert consensus methodology in which the initial set of items was iteratively evaluated and improved by the panel members until all the items reached a sufficiently high level of acceptability and consensus. The final version of the Transparency Checklist contains 36 items that cover four components of a study: preregistration; methods; results and discussion; and data, code and materials availability. In addition to the checklist, we developed an online app<sup>53</sup> by which all the checklist items can be easily answered (see Figure 3). The app also generates a report that can be submitted along with the article. This procedure comes with a number of benefits: (1) since the checklist is filled out before the submission of the article, most transparency-related aspects of the text (e.g., the explanation of the participant eligibility criteria) can be still improved. (2) the report of the Transparency Checklist can help editors, reviewers, and readers to gain insight into the transparency-related aspects of the study. (3) The guideline can be used for educational purposes, as students who follow the checklist with their research assignments will directly learn about these openness standards. (4) Finally, funding agencies can also improve the research culture if they expect the use of this checklist in their funded projects.

The manuscript...

(26) distinguishes explicitly between "confirmatory" (i.e., prespecified) and "exploratory" (i.e., not prespecified) analyses	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>
(27) describes how violations of statistical assumptions were handled.	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>
(28) justifies all statistical choices (e.g., including or excluding covariates; applying or not applying transformations; use of multi-level models vs. ANOVA).	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>
(29) reports the sample size for each cell of the design.	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>
(30) reports how incomplete or missing data were handled.	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>
(31) presents protocols for data preprocessing (e.g., cleaning, discarding of cases and items, normalizing, smoothing, artifact correction).	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> N/A	<b>1</b>

Figure 3. A screenshot from a section of the Transparency Checklist app. The full version of the checklist can be found at <http://www.shinyapps.org/apps/TransparencyChecklist/>

<sup>53</sup> <http://www.shinyapps.org/apps/TransparencyChecklist/>

In **Chapter 3.2.**, I presented one of our recent studies (Sarafoglou et al., 2022) on preregistration. The concept of preregistration goes back to the 19<sup>th</sup> century (Peirce, 1878) and centres around the expectation that the hypothesis should be stated before data collection. This principle of transparency aims to prevent the statistical analyses to be contaminated by confirmation and hindsight bias. Hypothesising after the results are known (known as HARKing, Kerr, 1998) is regarded to be one of the main sources of the credibility crisis as this questionable research practice can inflate the presence of false positive findings (Simmons et al., 2011). In addition to preregistering the hypothesis, the research procedure and the analysis plan are also expected to be laid out before the analysis of the data. Today, several journals recommend and value preregistration and online repositories can publicly store these documents with timestamps (Nosek & Lindsay, 2018). Preregistration, however, received criticism from various aspects. These voices about the procedure sometimes highlight that the replicability problem stems from some deeper issues (e.g., weak theories in social sciences (Szollosi & Donkin, 2021)), or that it curtails creativity (McDermott, 2022). In our paper, however, we didn't aim to address these concerns, instead we explored how preregistration affects the research workflow from the researchers' perspective. The motivation of our work originated from the anecdotal evidence that preregistration makes researcher think through the conceptual, analytical, and practical aspects of the project which, in turn, improved its quality. We asked psychologists who had and those who did not have experience with preregistration how preregistering the project affected their analysis plan, research hypothesis, experimental design, preparatory work, research data management, project workflow, collaboration, work-related stress, and total project duration. The results show evidence for the secondary benefits for preregistration, at least those who tried it found it beneficial in most stages of their workflow. Our project also identified some potential disadvantages, such as increased work-related stress and lengthened project duration. With our results, we hope to convince researchers to give a change to preregistration even if they don't believe that it is the ultimate solution for the credibility and replicability problems of social sciences.

In the study of **Chapter 3.3.** (Wagenmakers et al., 2021), we recommend seven procedures to enhance transparency, a fair acknowledgement of uncertainty, and openness to alternative interpretations in statistical practice. These procedures concentrate on (1) visualising data; (2) quantifying inferential uncertainty; (3) assessing data preprocessing choices; (4) reporting multiple models; (5) involving multiple analysts; (6) interpreting results modestly; and (7) sharing data and code. We argue that these procedures are directly linked to Merton's

(1973) ethos of science as they are driven by the norms of communalism, universalism, disinterestedness, and organized scepticism. For each procedure, we describe the associated benefits through examples. We also reflect on the current status and limitations of each procedure, and we provide reference to practical guidelines for their realisation.

**Chapter 3.3.1.** (Bago et al., 2022) presents an empirical project to demonstrate how these transparent practices can be used in psychological studies. The aim of this project was to test the universality of people's responses to a moral dilemma task, the trolley problem, by replicating its original design (Greene et al., 2009). As a start, we aimed to widen our sample and go beyond the original US sample, so we teamed up with the *Psychological Science Accelerator*<sup>54</sup> through which we could collect data in collaboration with 146 labs from 52 countries. To allow the improvement of the study before data collection, we submitted the research plan as a registered report to an interdisciplinary journal. Registered reports are assessed in two rounds: once before and once after data collection. Should the submission pass the review rounds before data collection, it can gain in-principle acceptance, meaning that if the submitted plan is followed then the journal would accept the study, irrespective of its results. A great advantage of this format is that it decreases the bias in the publication system towards positive findings. Another benefit comes from the opportunity to improve any part of the plan, in light of the review, before data collection. Importantly, our replication plan has been vetted by the lead author of the original study as well. We employed Bayesian analyses in our plan and set the decision threshold high;  $BF_{10}$  to  $> 10$  for  $H_1$  and  $< 1/10$  for  $H_0$ . These thresholds mean that we could claim evidence for or against any of our hypotheses only if the data gives at least 10 times more support to one, compared to the other hypothesis. Leveraging the benefits of the Bayesian framework, we proposed a sequential data collection, analysing the data after each cluster of participants and stopping data collection only when the results reach our pre-set evidence threshold. To estimate our required sample size, we conducted Monte Carlo simulations with the criterion to keep the probability of correct inference above 95%. As the results can always end up being inconclusive, our final incorrect inference rate was estimated to be  $< 0.001\%$ . Our preregistered analysis also promised to probe the robustness of our conclusions to different priors, meaning that our final results would not be sensitive to the set parameters of our hypothesis (Dienes, 2019). Eventually, the data collection was concluded

---

<sup>54</sup> <https://psysciacc.org/>

after 27,502 participants and the results provided robust support for the cultural universality of the original theory.

The Open Science movement introduced a great number of new terms and tools, radically changing the way how we think about and do science. Yet, as a side-effect, the burgeoning list of new terms can also create barriers to novices in the field or to those who are less involved in the recent developments in Open Science. Terms, such as CARKing, PARKing, or paradata are less known even among those who try to keep the pace with the new changes. For these reasons, I proposed a glossary for the terms related to Open Science. With the coordination of ‘Framework for Open and Reproducible Research Teaching’ (FORRT) community, we worked together with over 100 contributors so that the glossary would reflect the perspectives of a diverse range of disciplines. At the end, this crowd-sourced Open Scholarship Glossary contained more than 250 terms<sup>55</sup>, 30 of which are not published (Parsons et al., 2022). The Glossary project highlighted that it is not enough to constantly advance Open Science, the community should not lose touch with those who are not that much involved in the movement.

### *Contributions to the ‘issues with statistical practice’ topic*

Over the last decade, my colleagues and I developed a strong interest in the study of statistical practice from a descriptive, prescriptive, and supportive perspective. Our original motivation came from the experience of facing questionable statistical analyses in publications. Whenever the given methodology was within our expertise, we published comments on the original study by reanalysing the original data (e.g., Aczel et al., 2015; Kekecs et al., 2016) or the data we collected (Aczel et al., 2012, 2016; Aczel, Szollosi, & Bago, 2018; Aczel, Szollosi, Palfi, et al., 2018).

When the empirical issues of social sciences are discussed, it is common to blame the researcher with intentional bias or malpractice. It’s important to see, however, that humans are not machines; we can easily make honest mistakes. Data management is an area in which accidents easily happen. Most empirical researchers have memories of accidentally overwriting data-files, analysing the wrong dataset, or copying and pasting the wrong test statistics. These mistakes originate not from any bad intention but from the sheer fallibility of human functioning. In **Chapter 4.1**. (Kovacs, Hoekstra, et al., 2021), we show the results of a survey

---

<sup>55</sup> <https://forrt.org/glossary/>

that we conducted among researchers in order to explore the honest mistakes made during the data management process in psychological research. As it was the first known empirical project to explore researchers' data management mistakes, we had to develop a methodology for such exploration. As a result, we present a data management mistake taxonomy, concerning the causes (e.g., 'lack of planning'; 'inattention'), types (e.g., 'ambiguous naming'; 'version control error'), and outcomes (e.g., 'frustration', 'money loss') of these mistakes. With the help of our methodology, we could estimate not just the general frequency and severity of the data management mistakes in psychology labs, but we could identify what mistakes occur most frequently in psychological projects. To assist the field to reduce these errors, we provide a list of available solutions to prevent these mistakes for each major type of mistake.

In another approach of our work, we aimed to leverage the potentials of Bayesian statistics. An objective of a number of these metascientific projects was to assess the strength of evidence in psychological results. In one of our projects (Aczel et al., 2017), we calculated Bayes factors for 287,424 findings of 35,515 articles published in 293 psychological journals to see how strong evidence psychology provide in its publications. Overall, 55% of all analysed positive results were found to provide strong ( $BF > 10$ ) evidence while more than half of the remaining results do not pass the upper level of anecdotal evidence ( $BF = 3$ ). In the study of **Chapter 4.2.** (Aczel, Palfi, et al., 2018), we explored the strength of nonsignificant results in favour of the null hypothesis in psychological literature and found that fewer than 5% of the findings provided strong evidence for the null. In this study, we also explored how nonsignificant results are interpreted in high-profile psychological journals and found that the great majority of them did not follow the recommended practice. The methodology that we developed for this project has been replicated to explore other fields, such as gerontology (Brydges & Bielak, 2020), cognitive development (Legg et al., 2021), audiology (Brydges & Gaeta, 2019), rehabilitation research (Kinney et al., 2021), animal cognition (Farrar et al., 2022), and across many fields (Lyu et al., 2020).

The study of **Chapter 4.3.** (Aczel, Hoekstra, et al., 2020) asks why there is no consensual way of conducting Bayesian analyses. The promotion of Bayesian statistics in psychology is certainly limited if researchers find Bayesian experts disagree on how to prepare an analysis, how to calculate Bayes factors, and how to report or interpret the results. To ease this situation, I teamed up with international experts of Bayesian statistics in social sciences and identified the key points of the ongoing debates. Then, we discussed the seven main points and their sub-points and realised that there is no disagreement in the main principles. In the

paper, we emphasise that statistical inference is making choices. We agreed that the key is that no procedure can be ritualised; instead, researchers have to use common sense on a case-by-case basis. To assist analysts, we provide a thinking guideline that showed that questions, we think, should be considered when conducting Bayesian statistics.

**Chapter 4.4.** (Wagenmakers et al., 2022) presents a *Nature* article in which we draw attention to a neglected type of uncertainty in statistical analysis, the analytical variability. As discussed in the Introduction, there are various equally justifiable ways to analyse the same question on the same dataset. As most empirical reports present the results of a single analytical path, there remains the uncertainty whether alternative analyses would have presented the same results and conclusions. We demonstrate that such important questions as how much the COVID-19 virus is spreading in the population can have different but correct statistical answers, depending on the analyst's data handling and model choices. In order to gain robust answers to our empirical questions, we argue, journals should actively support multi-analyst project where more than one analyst provide independent analyses within the same study.

**Chapter 4.5.** (Aczel, Szaszi, Nilsonne, et al., 2021) extends this topic and presents consensus-based guidance for conducting and reporting such multi-analyst studies. Without doubt, the multi-analysts approach requires extra effort and comes with practical challenges. The aim of this guidance is to assist researchers in five stages of the workflow: (1) Recruiting Co-analysts; (2) Providing the Dataset, Research Questions, and Research Tasks; (3) Conducting the Independent Analyses; (4) Processing the Results; and (5) Reporting the Methods and Results. To further assist researchers in documenting multi-analyst projects, we also provide a modifiable reporting template, as well as a reporting checklist. The guidance documents were developed following a preregistered expert consensus procedure. At its first stage, the expert panel could suggest changes to an initial list of recommendations or modify them. In the next round, they could rate their agreement with each item. The final document contains those items that gain a pre-set level of support from the members of the panel.

In **Chapter 4.6.** (Kovacs et al., 2022), we provide support to researchers to calculate and report the sample size of their study. As discussed in the Introduction, low power is one of the problems in empirical social sciences. Under-sampled studies can easily produce waste by unpublishable results or impose bias to the evidence accumulation of a field. The calculation of sample sizes is a general expectation (Maxwell, 2004) but if the authors don't provide justification for the choices they make during their calculations, the whole procedure can lose its credibility, as there are too many decision points in the calculation that can be set

opportunistically. It is, however, insufficient to say that researchers should justify their sample size, the *what* and *how* should also be explained. The work provides guidance for all of these questions. First, it helps analysts to choose the method they need for their question. As a first step, the researchers should decide whether the aim is testing or estimation. Depending on that, our decision-tree offers a step-by-step guidance on choosing what sample size estimation method could be the most appropriate for the given case. Once the method is identified, the web application<sup>56</sup> that were launched along with the article, helps researchers to calculate the required sample size after all the parameters are set. Our app provides estimations for the following methods: TOST (Two One-Sides Tests); Interval Equivalence Bayes Factor; Bayes Factor Design Analysis; Predetermined Sample Size with Bayes Factor; Traditional Power Analysis; Power Curve; AIPE (Accuracy in Parameter Estimation); APP (A priori Procedure); and Region Of Practical Equivalence (ROPE) for independent two-group study designs. In addition to running the calculations, the app provides editable boilerplate justification texts that the author can adjust within the app and copy the resulting justification text to the article.

### ***Conclusions***

This thesis started with a reference to Francis Bacon's fantasy for the ideal organisation of a future scientific community. Even before the foundations of academic organisations, it was clear that specialisation of labour, roles, duties, and the supply of required equipment are key to the future of human knowledge and discovery. Undoubtedly, a great part of Bacon's dream became reality. Scientific communities are well-developed and organised around the world. Research institutes, publishers, scientific communities set the routines and standards of scientific practice. The progress in science and technology in the last few hundred years is a success story of humanity. Nevertheless, today's science is far from faultless. Many argue that in the prevailing incentive system quantity outweighs quality, story-telling brings more prestige than the honest presentation of facts, openness and transparency are non-rewarding. As a result, the published results lose their robustness and trustworthiness; science loses its credibility.

The widely recognised need to reform scientific practice led researchers from many disciplines to focus on metascience. In this thesis, I argued that psychologists can play an important role in the development of metascience and their perspective and methodology are indispensable for the understanding and improvement of science. The need for psychologists to

---

<sup>56</sup> <https://martonbalazskovacs.shinyapps.io/SampleSizePlanner>

contribute to our understanding of science has been repeatedly pressed in the past. Mahoney's words clearly express this need: "... we are still left with very meager understanding of the psychology of the scientist. / Our relative ignorance about *homo scientus* is not, in my opinion, a harmless mission which simply disturbs the esthetic balance of our current knowledge. The oversight is not benign. Our continues neglect of the scientist could well be the most costly blunder in the history of empiricism." (2004, p. xxx).

This thesis is non-traditional in the sense that it does not explore a single empirical question, rather it presents lines of studies to demonstrate how psychologists' perspective, methodology, and interdisciplinarity can play a pivotal role in the development of various areas of metascience. In particular, I showed how the publication system, research transparency, and the statistical practice can be understood and supported by the work of psychologists. Most of these projects have been conducted with international and inter-disciplinary collaborations.

Beyond their success in the targeted topics, these projects also shed light on some challenges that metascientists face when developing new tools or solutions to practical questions. One challenge is that traditional psychological training is insufficient for these projects as an array of other skills are required such as advanced programming knowledge, usability and user experience perspectives. Even though tools are essential and benefit many further, tool development doesn't fit in the traditional academic career schemes and there is a shortage of funding schemes available for tool development in academia. A further difficulty is that tools and guidelines are not in the traditional scope of journals and their presentation does not fit in the usual article formats. The success of new tools or practices in scientific practice depends on activities that go beyond the framework of the tool-development project. So that researchers start using these solutions, continuous dissemination, marketing, webhosting, maintenance, and updating are required. These ongoing demands can build up serious burden on the developers without no definite end. Finally, researchers are ignorant of best practices for citing tools such as software, so even when the tools are used, they remain unrecognised.

Despite all the difficulties of metascience, psychologists can find a lot of merit and reward in their contribution to the understanding and improvement of scientific practice. After all the pessimism of the credibility crisis, many find the zeal of the renaissance (Nelson et al., 2018; O'Connor, 2021) to join forces with other disciplines in setting new standards for a more transparent and reliable science. I emphasise, however, that while constantly advancing Open Science, the community should not lose touch with those who are less involved in the movement.



**Introduction and Summary References**

- Aalbersberg, Ij. J., Appleyard, T., Brookhart, S., Carpenter, T., Clarke, M., Curry, S., Dahl, J., DeHaven, A. C., Eich, E., Franko, M., Freedman, L., Graf, C., Grant, S., Hanson, B., Joseph, H., Kiermer, V., Kramer, B., Kraut, A., Karn, R. K., ... Vazire, S. (n.d.). *Making Science Transparent By Default; Introducing the TOP Statement*.  
<https://doi.org/10.31219/osf.io/sm78t>
- Aczel, B., Bago, B., & Foldes, A. (2012). Is there evidence for automatic imitation in a strategic context? *Proceedings of the Royal Society B: Biological Sciences*, 279(1741), 3231–3233.
- Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., & Vanpaemel, W. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, 4, 561–563.
- Aczel, B., Kovacs, M., Van Der Lippe, T., & Szaszi, B. (2021). Researchers working from home: Benefits and challenges. *PloS One*, 16(3), e0249127.
- Aczel, B., Palfi, B., & Szaszi, B. (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE*, 12(8), e0182651.  
<https://doi.org/10.1371/journal.pone.0182651>
- Aczel, B., Palfi, B., Szaszi, B., Szollosi, A., & Dienes, Z. (2015). Commentary: Unlearning implicit social biases during sleep. *Frontiers in Psychology*, 6, 1428.
- Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., Bergh, D. van den, & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation: *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245918773742>

- Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1), 1–8.
- Aczel, B., Szaszi, B., Nilsson, G., van den Akker, O. R., Albers, C. J., van Assen, M. A., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N., Donkin, C., van Doorn, J. B., ... Wagenmakers, E.-J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *ELife*, 10, e72185. <https://doi.org/10.7554/eLife.72185>
- Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4(1), 4–6. <https://doi.org/10.1038/s41562-019-0772-6>
- Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning*, 22(1), 99–117.
- Aczel, B., Szollosi, A., & Bago, B. (2018). The Effect of Transparency on Framing Effects in Within-Subject Designs. *Journal of Behavioral Decision Making*, 31(1), 25–39. <https://doi.org/10.1002/bdm.2036>
- Aczel, B., Szollosi, A., Palfi, B., Szaszi, B., & Kieslich, P. J. (2018). Is action execution part of the decision-making process? An investigation of the embodied choice hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(6), 918.
- Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, 508(7496), 312–313.

- Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. (2011). Public availability of published research data in high-impact journals. *PloS One*, *6*(9), e24357.
- American Psychological Association. (2001). *Publication manual 5th edition*. American Psychological Association Washington, DC.
- Avey, M. T., Moher, D., Sullivan, K. J., Fergusson, D., Griffin, G., Grimshaw, J. M., Hutton, B., Lalu, M. M., Macleod, M., & Marshall, J. (2016). The devil is in the details: Incomplete reporting in preclinical animal research. *PloS One*, *11*(11), e0166733.
- Azevedo, F., Parsons, S., Micheli, L., Strand, J. F., Rinke, E., & Guay, S. (2019). Introducing a Framework for Open and Reproducible Research Training (FORRT). *Preprint at <https://doi.org/10.31219/osf.io/bnh7p>*.
- Bacon, F. (1627). *New Atlantis and the great instauration*. John Wiley & Sons.
- Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albaloooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., ... Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*(6), 880–895. <https://doi.org/10.1038/s41562-022-01319-5>
- Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. <https://doi.org/10.1177/0956797616647519>
- Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.
- Barber, B. (1961). Resistance by Scientists to Scientific Discovery: This source of resistance has yet to be given the scrutiny accorded religious and ideological sources. *Science*, *134*(3479), 596–602.

- Bartlett, A., & Mercer, G. (2000). Reconceptualising Discourses of Power in Postgraduate Pedagogies. *Teaching in Higher Education*, 5(2), 195–204.  
<https://doi.org/10.1080/135625100114849>
- Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., ... Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, 137, 110211.  
<https://doi.org/10.1016/j.jpsychores.2020.110211>
- Baum, M. A., Braun, M. N., Hart, A., Huffer, V. I., Meßmer, J. A., Weigl, M., & Wennerhold, L. (2022). The first author takes it all? Solutions for crediting authors more visibly, transparently, and free of bias. *British Journal of Social Psychology*, 00, 1–16.  
<https://doi.org/10.1111/bjso.12569>
- Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–425. <https://doi.org/10.1037/a0021524>
- Berger, J. O., & Wolpert, R. L. (1988). The likelihood principle. *Lecture Notes-Monograph Series*, 6, iii–199.
- Bishop, D. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, 73(1), 1–19.  
<https://doi.org/10.1177/1747021819886519>
- Bishop, D., & Gill, E. (2020). Robert Boyle on the importance of reporting and replicating experiments. *Journal of the Royal Society of Medicine*, 113(2), 79–83.

- Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, 7(4), 914–923. <https://doi.org/10.1016/j.joi.2013.09.001>
- Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLoS ONE*, 5(6), e11273. <https://doi.org/10.1371/journal.pone.0011273>
- Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Kryptos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, 87, 46–75. <https://doi.org/10.1016/j.jmp.2018.09.004>
- Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1), 1–15. <https://doi.org/10.1057/s41599-021-00903-w>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Boyle, R. (1911). *The Sceptical Chymist* (1661), London: JM Dent & Sons, Ltd; New York: E. P. Dutton & Co.
- Brierley, L., Nanni, F., Polka, J. K., Dey, G., Pálffy, M., Fraser, N., & Coates, J. A. (2022). Tracking changes between preprint posting and journal publication during a pandemic. *PLOS Biology*, 20(2), e3001285. <https://doi.org/10.1371/journal.pbio.3001285>

- Brydges, C. R., & Bielak, A. A. (2020). A Bayesian analysis of evidence in support of the null hypothesis in gerontological psychology (or lack thereof). *The Journals of Gerontology: Series B*, 75(1), 58–66.
- Brydges, C. R., & Gaeta, L. (2019). An analysis of nonsignificant results in audiology using Bayes Factors. *Journal of Speech, Language, and Hearing Research*, 62(12), 4544–4553.
- Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.
- Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., & Liu, L. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, 61(4), 349.
- Callaway, E. (2017). Heavyweight funders back central site for life-sciences preprints. *Nature*, 542(7641), 283–284. <https://doi.org/10.1038/nature.2017.21466>
- Callaway, E., & Powell, K. (2016). Biologists urged to hug a preprint. *Nature*, 530(7590), 265–265. <https://doi.org/10.1038/530265a>
- Campbell, P. (2013). Announcement: Reducing our irreproducibility. *Nature*, 496, 398.
- Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, 63(1), 289–300. <https://doi.org/10.1016/j.neuroimage.2012.07.004>
- Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Jama*, 291(20), 2457–2465.
- Christie, A. P., White, T. B., Martin, P. A., Petrovan, S. O., Bladon, A. J., Bowkett, A. E., Littlewood, N. A., Mupepele, A.-C., Rocha, R., Sainsbury, K. A., Smith, R. K., Taylor,

- N. G., & Sutherland, W. J. (2021). Reducing publication delay to improve the efficiency and impact of conservation science. *PeerJ*, 9, e12245. <https://doi.org/10.7717/peerj.12245>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145–153.
- Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, 601(7894), 505–507.
- Counsell, A., & Harlow, L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/Psychologie Canadienne*, 58(2), 140–147.
- Crawford, W. (2019). Gold open access 2013–2018: Articles in journals (GOA4). Livermore, CA.: Cites & Insights Books.
- Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science. *Zeitschrift Für Psychologie*, 227(4), 237–248. <https://doi.org/10.1027/2151-2604/a000387>
- Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- De Finetti, B. (2017). *Theory of probability: A critical introductory treatment* (Vol. 6). John Wiley & Sons.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3), 274–290. <https://doi.org/10.1177/1745691611406920>

- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z. (2019). How do I know what my theory predicts? Accessed via: <https://psyarxiv.com/yqaj4>. *Advances in Methods and Practices in Psychological Science*, 2(4), 364–377.
- Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, 4(2), 160254. <https://doi.org/10.1098/rsos.160254>
- Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Kryptos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., ... Donkin, C. (2019). The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychonomic Bulletin & Review*, 26(4), 1051–1069. <https://doi.org/10.3758/s13423-017-1417-2>
- Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., & Gamble, C. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS One*, 3(8), e3081.
- Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *The Lancet*, 350(9074), 326–329.
- Einav, L., & Yariv, L. (2006). What's in a Surname? The Effects of Surname Initials on Academic Success. *Journal of Economic Perspectives*, 20(1), 175–187. <https://doi.org/10.1257/089533006776526085>



- Else, H. (2020). Nature journals reveal terms of landmark open-access option. *Nature*, 588(7836), 19–20. <https://doi.org/10.1038/d41586-020-03324-y>
- Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, 16(4), 779–788. <https://doi.org/10.1177/1745691620970586>
- Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1(1), 60–69.
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25(1), 219–234. <https://doi.org/10.3758/s13423-017-1317-5>
- Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian Inference for Psychology. *Psychonomic Bulletin & Review*, 25(1), 5–34. <https://doi.org/10.3758/s13423-017-1262-3>
- European Commission. (2012). Towards better access to scientific information: Boosting the benefits of public investments in research. In *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee, and the Committee of the Regions, COM (2012) 401 Final*. European Commission Brussels.
- Evans, J. A., & Reimer, J. (2009). Open Access and Global Participation in Science. *Science*, 323(5917), 1025–1025. <https://doi.org/10.1126/science.1154562>
- Evans, T. (2022). Developments in open data norms. *Journal of Open Psychology Data*, 10(1).
- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, 4(5), e5738.
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904.

- Fanelli, D., & Larivière, V. (2016). Researchers' individual publication rate has not increased in a century. *PloS One*, *11*(3), e0149504.
- Farrar, B., Vernouillet, A., Garcia-Pelegrin, E., Legg, E. W., Brecht, K., Lambert, P., Elsherif, M., Francis, S., O'Neill, L., & Clayton, N. (2022). *Reporting and interpreting non-significant results in animal cognition research*. PsyArXiv. <https://doi.org/10.31234/osf.io/g9ja2>
- Feist, G. J. (2008). *The psychology of science and the origins of the scientific mind*. Yale University Press.
- Feist, G. J. (2011). Psychology of science as a new subdiscipline in psychology. *Current Directions in Psychological Science*, *20*(5), 330–334.
- Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift Für Psychologie/Journal of Psychology*, *217*(1), 27–37.
- Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. <https://doi.org/10.1177/1948550615612150>
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory & Psychology*, *12*(6), 825–853.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.
- Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. <https://doi.org/10.1126/science.aao0185>

- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505. <https://doi.org/10.1126/science.1255484>
- Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: Mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33–41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
- Galilei, G. (2016). *Sidereus Nuncius, or the sidereal messenger*. University of Chicago Press.
- Gelman, A. (2016). What has happened down here is the winds have changed [Blog post]. Retrieved from: <http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/>.
- Getz, M. (2005). Open Scholarship and Research Universities. In *Vanderbilt University Department of Economics Working Papers* (No. 0517; Vanderbilt University Department of Economics Working Papers). Vanderbilt University Department of Economics. <https://ideas.repec.org/p/van/wpaper/0517.html>
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198–218.
- Gilbert, G. E., & Prion, S. (2016). Making Sense of Methods and Measurement: The Danger of the Retrospective Power Analysis. *Clinical Simulation In Nursing*, 12(8), 303–304. <https://doi.org/10.1016/j.ecns.2016.03.001>
- Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, 20(1), 1–16.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45(3), 135–140. <https://doi.org/10.1053/j.seminhematol.2008.04.003>

- Goodman, S. N., & Berlin, J. A. (1994). The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. *Annals of Internal Medicine*, *121*(3), 200–206. <https://doi.org/10.7326/0003-4819-121-3-199408010-00008>
- Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.
- Grossmann, A., & Brembs, B. (2021). *Current market rates for scholarly publishing services* (10:20). F1000Research. <https://doi.org/10.12688/f1000research.27468.1>
- Guo, X., Li, X., & Yu, Y. (2021). Publication delay adjusted impact factor: The effect of publication delay of articles on journal impact factor. *Journal of Informetrics*, *15*(1), 101100. <https://doi.org/10.1016/j.joi.2020.101100>
- Hall, J., & Martin, B. R. (2019). Towards a taxonomy of research misconduct: The case of business school research. *Research Policy*, *48*(2), 414–427. <https://doi.org/10.1016/j.respol.2018.03.006>
- Hardwicke, T. E., & Goodman, S. N. (2020). How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey. *PLOS ONE*, *15*(10), e0239598. <https://doi.org/10.1371/journal.pone.0239598>
- Hardwicke, T. E., & Ioannidis, J. P. (2018). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLoS One*, *13*(8), e0201856.
- Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37.

- Harter, S. P., & Kim, H. J. (1997). ARCHIVE: electronic journals and scholarly communication: A citation and reference study. *Journal of Electronic Publishing*, 3(2).  
<https://doi.org/10.3998/3336451.0003.212>
- Hartgerink, C. H., van Aert, R. C., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, 4, e1935. <https://doi.org/doi.org/10.7717/peerj.1935>
- Himmelstein, D. S., & Powell, K. (2021). Analysis for “the history of publishing delays” blog post v1.0 (2016). URL <https://doi.org/10.5281/Zenodo.45516>.
- Hoekstra, R., Johnson, A., & Kiers, H. A. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, 72(6), 1039–1052.
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157–1164. <https://doi.org/10.3758/s13423-013-0572-3>
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 19–24.
- Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, 8(4), 201925. <https://doi.org/10.1098/rsos.201925>
- Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLOS ONE*, 15(12), e0244611. <https://doi.org/10.1371/journal.pone.0244611>
- Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Kwame Appiah, O., Atkinson, Q. D., Baimel,

- A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartoš, F., Becerra, M., Beffara, B., ... Wagenmakers, E.-J. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, 1–47. <https://doi.org/10.1080/2153599X.2022.2070255>
- Horbach, S. P. J. M. (2020). Pandemic publishing: Medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, 1(3), 1056–1067. [https://doi.org/10.1162/qss\\_a\\_00076](https://doi.org/10.1162/qss_a_00076)
- Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1(1), 70–85.
- Hua, F., Shen, C., Walsh, T., Glenny, A.-M., & Worthington, H. (2017). Open Access: Concepts, findings, and recommendations for stakeholders in dentistry. *Journal of Dentistry*, 64, 13–22. <https://doi.org/10.1016/j.jdent.2017.06.012>
- Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3), 944–960. <https://doi.org/10.1111/ecin.12992>
- Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Jama*, 279(4), 281–286.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Jahn, N., Matthias, L., & Laakso, M. (2022). Toward transparency of hybrid open access through publisher-provided metadata: An article-level study of Elsevier. *Journal of the Association for Information Science and Technology*, 73(1), 104–118. <https://doi.org/10.1002/asi.24549>

- Jain, V. K., Iyengar, Karthikeyan. P., & Vaishya, R. (2020). Article processing charge may be a barrier to publishing. *Journal of Clinical Orthopaedics and Trauma*, *14*, 14–16.  
<https://doi.org/10.1016/j.jcot.2020.10.039>
- Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of Editorial Peer Review A Systematic Review. *JAMA*, *287*(21), 2784–2786.  
<https://doi.org/10.1001/jama.287.21.2784>
- Jiang, Y., Lerrigo, R., Ullah, A., Alagappan, M., Asch, S. M., Goodman, S. N., & Sinha, S. R. (2019). The high resource impact of reformatting requirements for scientific papers. *PLOS ONE*, *14*(10), e0223976. <https://doi.org/10.1371/journal.pone.0223976>
- Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing*, *23*(3), 258–263.
- Jiroutek, M. R., & Turner, J. R. (2017). Why it is nonsensical to use retrospective power analyses to conduct a postmortem on your study. *The Journal of Clinical Hypertension*, *20*(2), 408–410. <https://doi.org/10.1111/jch.13173>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. <https://doi.org/10.1177/0956797611430953>
- Jüni, P., Holenstein, F., Sterne, J., Bartlett, C., & Egger, M. (2002). Direction and impact of language bias in meta-analyses of controlled trials: Empirical study. *International Journal of Epidemiology*, *31*(1), 115–123.
- Kahneman, D. (2012). *A proposal to deal with questions about priming effects*. [https://www.nature.com/news/polopoly\\_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf](https://www.nature.com/news/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf)

- Kekecs, Z., Szollosi, A., Palfi, B., Szaszi, B., Kovacs, K. J., Dienes, Z., & Aczel, B. (2016). Commentary: Oxytocin-gaze positive loop and the coevolution of human–dog bonds. *Frontiers in Neuroscience, 10*, 155.
- Keren, G., & Lewis, C. (1993). *A handbook for data analysis in the behavioral sciences: Methodological issues*. L. Erlbaum Associates Hillsdale, NJ.
- Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196–217.  
[https://doi.org/10.1207/s15327957pspr0203\\_4](https://doi.org/10.1207/s15327957pspr0203_4)
- Khan, A., Montenegro-Montero, A., & Mathelier, A. (2018). Put science first and formatting later. *EMBO Reports, 19*(5), e45731.
- Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., & Hess-Holden, C. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456. <https://doi.org/10.1027/1864-9335/a000178>
- Kinney, A. R., Middleton, A., & Graham, J. E. (2021). A Bayesian analysis of non-significant rehabilitation findings: Evaluating the evidence in favour of truly absent treatment effects. *Annals of Physical and Rehabilitation Medicine, 64*(4), 101425.
- Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsson, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology, 4*(1), 20.  
<https://doi.org/10.1525/collabra.158>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and*



- Practices in Psychological Science*, 1(4), 443–490.  
<https://doi.org/10.1177/2515245918810225>
- Koch, C., & Jones, A. (2016). Big science, team science, and open science for neuroscience. *Neuron*, 92(3), 612–616.
- Kovacs, M., Hoekstra, R., & Aczel, B. (2021). The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management. *Advances in Methods and Practices in Psychological Science*, 4(4), 25152459211045930.
- Kovacs, M., Holcombe, A., Aust, F., & Aczel, B. (2021). Tenzing and the importance of tool development for research efficiency. *Information Services & Use*, 41(1–2), 123–130.
- Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2022). SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies. *Advances in Methods and Practices in Psychological Science*, 5(1), 25152459211054060.
- Kozlov, M. (2022). NIH issues a seismic mandate: Share data publicly. *Nature*, 602(7898), 558–559. <https://doi.org/10.1038/d41586-022-00402-1>
- Kožnjak, B. (2017). Kuhn Meets Maslow: The Psychology Behind Scientific Revolutions. *Journal for General Philosophy of Science*, 48(2), 257–287.  
<https://doi.org/10.1007/s10838-016-9352-x>
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.
- Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The development of open access journal publishing from 1993 to 2009. *PloS One*, 6(6), e20961.
- Lakens, D. (2022). *Improving Your Statistical Inferences*.  
<https://doi.org/10.5281/zenodo.6409077>
- Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-

- based articles. *PLOS ONE*, 16(6), e0253129.  
<https://doi.org/10.1371/journal.pone.0253129>
- Larsen, P., & Von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3), 575–603.
- Leahey, T. H. (2004). *A history of psychology: Main currents in psychological thought* (6th ed.). Prentice-Hall, Inc.
- Legg, E. W., Farrar, B., Lazić, A., Thiele, M., Kampis, D., Mani, N., Sesar, K., Klapwijk, E., Schlingloff, L., Reindl, E., Li, K., Birovljević, G., Attwood, M., Tatone, D., Yuniarto, L. S., & Ostojic, L. (2021). *Assessing the reporting and interpretation of non-significant results in the study of cognitive development: A systematic review*. PsyArXiv.  
<https://doi.org/10.31234/osf.io/mxac9>
- Lewis, D. W. (2012). The inevitability of open access. *College & Research Libraries*, 73(5), 493–506.
- Lilienfeld, S. O., & Waldman, I. D. (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley & Sons.
- Lobo, F., & Crawford, P. (2003). Time, closed timelike curves and causality. In *The Nature of Time: Geometry, Physics and Perception* (pp. 289–296). Springer.
- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M., & Epskamp, S. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, 88, 1–17.
- Lyu, X.-K., Xu, Y., Zhao, X.-F., Zuo, X.-N., & Hu, C.-P. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, 14. <https://doi.org/10.1017/prp.2019.28>
- Mahoney, M. J. (2004). *Scientist as subject: The psychological imperative*. ISD LLC.

- Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331–348.
- Maslow, A. H. (1966). *The Psychology of Science: A reconnaissance*. Harper & Row.
- Maunsell, J. (2008). Neuroscience Peer Review Consortium. *Journal of Neuroscience, 28*(4), 787–787.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.
- Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology, 6*(3), 279–284.
- McDermott, R. (2022). Breaking free: How preregistration hurts scholars and science. *Politics and the Life Sciences, 41*(1), 55–59.
- Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Razen, M., Weitzel, U., Abad, D., Abudy, M. (Meni), Adrian, T., Ait-Sahalia, Y., Akmansoy, O., Alcock, J., Alexeev, V., Aloosh, A., Amato, L., Amaya, D., Angel, J., ... Bao, L. (2021). *Non-Standard Errors* [SSRN Scholarly Paper]. <https://doi.org/10.2139/ssrn.3961574>
- Merton, R. K. (1963). Resistance to the Systematic Study of Multiple Discoveries in Science. *European Journal of Sociology / Archives Européennes de Sociologie, 4*(2), 237–282. <https://doi.org/10.1017/S0003975600000801>
- Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press. <http://dx.doi.org/10.1063/1.3128814>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med, 6*(7), e1000097.

- Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Morey, C. C., Newman, D. P., & Schönbrodt, F. D. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.
- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., & Antfolk, J. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. <https://doi.org/10.1038/s41562-016-0021>
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2018). Beyond WEIRD psychology: Measuring and mapping scales of cultural and psychological distance. *Available at SSRN 3259613*.
- Nature. (2006). Peer review and fraud. *Nature*, *444*(7122), 971–972. <https://doi.org/10.1038/444971b>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231*, 289–337.

- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301. <https://doi.org/10.1037/1082-989X.5.2.241>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, 31, 19–21.
- Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4), 1205–1226.
- Oakes, M. W. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Wiley.
- O'Connor, D. B. (2021). Leonardo da Vinci, preregistration and the architecture of science: Towards a more open and transparent research culture. *Health Psychology Bulletin*, 5(1), 39–45.
- Page, M. J., Sterne, J. A., Higgins, J. P., & Egger, M. (2021). Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods*, 12(2), 248–259.
- Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., ... Aczel, B. (2022). A community-sourced glossary of open scholarship

- terms. *Nature Human Behaviour*, 6(3), 312–318. <https://doi.org/10.1038/s41562-021-01269-4>
- Pashler, H., & Wagenmakers, E. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- Peh, W. C. (2022). Peer review: Concepts, variants and controversies. *Singapore Med J*, 63(2), 55–60.
- Peirce, C. S. (1878). *The probability of induction*. In *The Collected writings of Charles Sanders Peirce, 1930–1934. Vol. II*. Cambridge, Eng.: Univer. of Cambridge Press.
- Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375. <https://doi.org/10.7717/peerj.4375>
- Pléh, C. (2010). *A lélektan története*. Osiris.
- Poldrack, R. A. (2018). *Statistical Thinking for the 21st Century*. Russell Poldrack. <https://statstinking21.github.io/statstinking21-core-site/>
- Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, 27(7), 1036–1042. <https://doi.org/10.1177/0956797616645672>
- Psychonomic Society. (2012). *New Statistical Guidelines for Journals of the Psychonomic Society*. <https://www.psychonomic.org/page/statisticalguidelines>
- Regalado, A. (1995). Multiauthor papers on the rise. *Science*, 268(5207), 25–25.
- Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use—A literature review. *Research Evaluation*, 25(2), 161–169.

- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Ross-Hellauer, T., Deppe, A., & Schmidt, B. (2017). Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers. *PLOS ONE*, 12(12), e0189311. <https://doi.org/10.1371/journal.pone.0189311>
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21(2), 301–308. <https://doi.org/10.3758/s13423-014-0595-4>
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is There a Free Lunch in Inference? *Topics in Cognitive Science*, 8(3), 520–547. <https://doi.org/10.1111/tops.12214>
- Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, 2(1).
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15), 8398–8403. <https://doi.org/10.1073/pnas.1915006117>
- Sarafoglou, A., Kovacs, M., Bakos, B. E., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, 9(7), 211997.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(18), 1. <https://doi.org/10.1186/1741-7015-8-18>

- Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., ... Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. <https://doi.org/10.1016/j.obhdp.2021.02.003>
- Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309–316.
- Shi, D., Rousseau, R., Yang, L., & Li, J. (2017). A journal's impact factor is influenced by changes in publication delays of citing journals. *Journal of the Association for Information Science and Technology*, *68*(3), 780–789.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., ... Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- 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*(4), 178–182. <https://doi.org/10.1177/014107680609900414>



- Smith, R. (2015, May 28). Ineffective at any dose? Why peer review simply doesn't work (May 28). *Times Higher Education*. <https://www.timeshighereducation.com/content/the-peer-review-drugs-dont-work>
- Spier, R. (2002). The history of the peer-review process. *TRENDS in Biotechnology*, 20(8), 357–358.
- Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., & Dobbins, I. G. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, 2(4), 335–349.
- Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4), 717–724.
- Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., & Lukner, H. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1), 1–11.
- Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS One*, 10(8), e0134826.
- Time to break academic publishing's stranglehold on research. (2018). *New Scientist*. <https://www.newscientist.com/article/mg24032052-900-time-to-break-academic-publishings-stranglehold-on-research/>
- Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, 6. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00726>

- Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author Sequence and Credit for Contributions in Multiauthored Publications. *PLOS Biology*, 5(1), e18. <https://doi.org/10.1371/journal.pbio.0050018>
- United Nations, U. G. (1948). Universal declaration of human rights. *UN General Assembly*, 302(2), 14–25.
- Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2), 217–239. <https://doi.org/10.1037/met0000100>
- Van Noorden, R. (2013). Half of 2011 papers now free to read. *Nature*, 500(7463), 386–387. <https://doi.org/10.1038/500386a>
- van Zwet, E. W., & Cator, E. A. (2021). The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*, 75(4), 437–452. <https://doi.org/10.1111/stan.12241>
- Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article Commentary: On the Persistence of Low Power in Psychological Science. *Quarterly Journal of Experimental Psychology*, 67(5), 1037–1040. <https://doi.org/10.1080/17470218.2014.885986>
- Veronese, M., Rizzo, G., Belzunce, M., Schubert, J., Searle, G., Whittington, A., Mansur, A., Dunn, J., Reader, A., & Gunn, R. N. (2021). Reproducibility of findings in modern PET neuroimaging: Insight from the NRM2018 grand challenge. *Journal of Cerebral Blood Flow & Metabolism*, 41(10), 2778–2796.
- Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, 24(1), 94–97.
- Wagenmakers, E.-J., Sarafoglou, A., Aarts, S., Albers, C., Algermissen, J., Bahník, Š., van Dongen, N., Hoekstra, R., Moreau, D., van Ravenzwaaij, D., Sluga, A., Stanke, F., Tendeiro, J., & Aczel, B. (2021). Seven steps toward more transparency in statistical

- practice. *Nature Human Behaviour*, 5(11), 1473–1480. <https://doi.org/10.1038/s41562-021-01211-8>
- Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, 605(7910), 423–425.
- Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, 47(4), 913–917. <https://doi.org/10.3758/s13428-014-0517-4>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Watson, C. (2022). Many researchers say they’ll share data—But don’t. *Nature*, 606(7916), 853–853. <https://doi.org/10.1038/d41586-022-01692-1>
- Watson, D. L. (1938). *Scientists are human*. Watts.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158–177.
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, 6(3), 291–298. <https://doi.org/10.1177/1745691611406923>
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7), 726–728.
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01832>

- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- World Medical Association. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4), 373–374.
- Yohe, G. W. (1980). Current Publication Lags in Economics Journals. *Journal of Economic Literature*, 18(3), 1050–1055.
- Zuccala, A. (2010). Open Access and Civic Scientific Information Literacy. *Information Research: An International Electronic Journal*, 15(1). <https://eric.ed.gov/?id=EJ881439>