JOURNAL of ~~TRIAL~~ ~~TRIOL~~ TRIAL and ~~ERRO~~ ERROR

Journal of Trial and Error
Vol 4 N°2
ISSN 2667-1204

Journal of Trial and Error

Volume 4
Issue 2
December 23, 2024

ISSN 2667-1204
https://doi.org/10.36850/i4.2

Contents

Editorial from our Incoming Editor in Chief: Introducing Open Peer Review, Streamlined Review, and a Trial of the Registered Report Format

Sarahanne Field[1,6], Stefan D.M. Gaillard[2,6], David Joachim Grüning[3,4,6], Alex Job Visser[5,6]

[1]University of Groningen
[2]Institute for Science in Society, Radboud University Nijmegen, the Netherlands
[3]Department of Psychology, Heidelberg University, Heidelberg, Germany
[4]Department for Survey Design & Methodology, GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany
[5]Andersson Elffers Felix, Utrecht
[6]Center of Trial and Error, Utrecht, the Netherlands

Received
October 18, 2024
Accepted
November 10, 2024
Published
December 23, 2024
Issued
December 23, 2024

Correspondence
University of Groningen
s.m.field@rug.nl

Keywords *editorial, metascience, science reform*

In most industries, error is a problem. Take, for instance, the aviation industry. An error in the calculation of a plane's gross weight could be catastrophic, as how much a plane weighs directly influences how it needs to be configured for take-off and landing. Countless fatal aviation accidents underscore the importance of error-free weight calculations. The crash of Air Midwest flight 5481 in 2003, involving 22 human fatalities (National Transportation Safety Board, 2004) is just one example. Errors in communication between a pilot and an air-traffic controller can also lead to serious problems for safe aviation. Poor communication is listed as one of the primary causes of the accident with the highest number of human fatalities in the world to date – the Tenerife Airport disaster of 1977, in which a KLM aircraft struck a taxiing Pan Am jumbo, killing 583 people (McCreary et al., 1998).

Although the stakes and scope tend to be radically different in comparison with aviation, error can also be a problem in science. The nature of the error is important for understanding how detrimental its effect may be. While errors such as making mistakes in reporting statistical tests or mislabeling graph axes are undesirable and sloppy, it is *unreported* error that can truly undermine research quality. Unreported errors, such as inappropriate handling of missing data, errors in data processing, and undisclosed bias can lead to, and likely has led to, an unreliable literature body in all scientific disciplines. Consider well known examples such as the 1998 Lancet study on vaccines and autism, or the 2006 Potti Affair in cancer research (note, both of these articles have been retracted from the scientific literature). These errors and biases ultimately cause a flawed understanding of our universe and its inhabitants both for us and for future generations.

Consider the number of article retractions that occur every year. In 2023 alone, more than 10,000 articles were retracted from the scientific record — a record number, far exceeding previous years, and representing a trebling in numbers in the past decade (compare for instance, van Noorden's 2023 report that in 2022, just over 4,000 articles were retracted). Some see these numbers as evidence of a self-correcting science. Another perspective, however, is that these articles represent only a fraction of the literature that *should* be retracted from the scientific record – a fraction of the 'error' in the literature that still gets cited and still undermines the foundations of science.

Reframing Error as Constructive

But what if error could be informative? Certainly, the commercial aviation industry has constantly and greatly evolved through seeing error as being informative. As a result of the Air Midwest flight 5481 disaster, the American Federal Aviation Administration revised erroneous estimated weight values (which had not been reviewed since 1936; National Transportation Safety Board, 2004). As a consequence of the Tenerife disaster, new requirements for a standard aviation phraseology, as well as air traffic instruction and response, were implemented across the world, among other changes (see Helmreich et al., 2017). In the case of aviation, learning from error in this context comes at a great cost of human lives. Luckily that is not typically true in the research sphere.

Provided that mistakes are reported and used as learning opportunities, we at the *Journal of Trial & Error* (JOTE) argue that error should be embraced as a normal part of the scientific process. We argue that far from error hurting the scientific process, it can in fact enrich, and even move it forward. Error in science – when thoroughly and transparently documented and contextualized – can help us improve how we conduct scientific research (through finding out and reflecting on how the scientific process *shouldn't* work) and can even help us discover boundary conditions for effects and phenomena. While most journals only tend to publish error-free accounts of research, JOTE provides an outlet for publishing when things do go wrong, or do not go as expected. Our objective is to make these errors informative and prompt reflection for the scientific community, as well as to normalize discussion and disclosure of error. We approach the scientific enterprise pragmatically and critically, recognizing that the research process is rarely smooth, linear, and mistake-free.

In the same vein, JOTE also seeks to stimulate a fundamental aspect of cumulative science that has been de-emphasized (and even somewhat eliminated) in many other academic venues. Namely, the possibility of genuine scientific debate through a variety of publication formats, as if the scientific literature should function as a 'book of conversations' (Davis-Stober et al., 2024). In this effect, JOTE supports a plethora of different ways of doing and discussing science, be it from an empirical, conceptual, or meta-perspective standpoint. Furthermore, the journal seeks to go beyond the usual lens of published work by inviting reports on internal scientific processes, such as rejected grant applications or corrigenda and errata. JOTE regularly invites metascientists and humanities scholars to contribute their reflections (in the form of an independent article) on each recently published paper in JOTE. These reflections are also forwarded to the wider audience of scientists to add contributions. All these approaches are intended to fulfill the original idea of science on which its cumulative endeavor is based: an open, curious, and critical *conversation*, incorporating a diversity of rational perspectives (Field et al., 2024; Salmieri, 2024).

Changes at JOTE

Changes to the JOTE Team

Maura Burke served as JOTE's second editor-in-chief from 2020. In 2024, her developing career in academia led her to new challenges outside of the journal. In her place, JOTE is proud to present the new editor-in-chief, Dr. Sarahanne M. Field. In what follows, Dr. Field shares her plans for the journal going forward. Speaking of her involvement so far, she says:

> *In May 2023 I was proud to join the journal as the editor of the then-new metascience section; proud to be part of an organization which was trying to shift how we see error in science. I was proud in explicitly helping JOTE onto the metascience map. This year, I am honored to step into the role of the journal's Editor-in-Chief, to continue to bring it further in its aims alongside my colleagues. In stepping into this role, I have the opportunity to introduce three changes to how we operate at JOTE, all of which, in my opinion, allow the journal to better provide value not only to the metascience community, but to the broader scientific sphere.*

Registered Report Submission Format for JOTE

First, as of March 25, 2024, JOTE began a one-year trial of the registered report (RR) submission type. Doing an RR involves carefully documenting the plans for a research study and getting it peer-reviewed before the study is conducted. Once the plan receives positive peer review (typically after at least one review round), the study plan receives what is known as in-principle acceptance (IPA). A decision of IPA means that as long as the study adheres to the plan (or clearly describes deviations), the completed study will be published (Field et al., 2020).

Introducing the RR submission format strengthens our commitment to normalizing the discussion of error in science, in that it provides a way to publish well-planned studies regardless of the outcomes. It doesn't matter if the hypotheses aren't supported by the empirical findings – if you conduct a high-quality planned study and transparently report on what happened, it will be published. It should be emphasized that the RR format is no panacea. It does not fix all the problems we face in research, indeed, if people want to conduct poor-quality research, they can (although we will not be publishing it)! Nevertheless, it can be a powerful tool in the hands of researchers who are motivated to conduct more reliable and valid studies.

Motivated researchers benefit from not having to worry about whether they find support for their hypotheses or not, can avoid the undue influences of hindsight and confirmation bias, and can test whether their planned study might be good enough to be part of the scientific record before they do the work of conducting the study (Field et al., 2020). The literature body benefits as well, from articles that tell the whole story of research rather than just the nice, perfect parts. Members of the public, in turn receive trustworthy information about themselves and the world in which they exist, so long as the media accurately represent study results – a contingency that unfortunately cannot be relied upon. Finally, the benefits of the RR format can be conferred to other researchers. Researchers can avoid conducting studies that will end up being uninformative and wasting time and money (resources that are scarce and hard-come-by in academia). To do so, the mismatch between what is researched and what is published should be kept to a minimum (the overarching goal of JOTE, and a likely outcome of more and more RRs in the scientific record; see Devine et al., 2020, and Field et al., 2020, respectively).

Streamlined Review

The second change we will be making to submission at JOTE is to introduce the streamlined review submission type, which is also being implemented in *Collabra: Psychology* (the first scientific publishing outlet to do so, to our knowledge). Authors can now request a streamlined review process for articles they submit to JOTE. The streamlined review process allows for articles that have undergone peer review at another journal, but which have been rejected by that journal for reasons unrelated to rigor or methodological soundness or which have been withdrawn by the author for improper journal conduct. Usually, such articles will have been rejected for lack of novelty or impact, or for reporting inconclusive results.

JOTE's approach to research dissemination attempts to break the mold of traditional publishing in many ways. Part of this involves publishing articles that do not fulfill traditional criteria, such as being groundbreaking or following the expected trajectory of supporting the stated hypotheses exactly as expected. To the editors of JOTE, an article has earned a place in the research record if it is scientifically sound, provides scientific value (including informative failures to replicate, or failures to support study hypotheses) and, crucially, communicates these qualities effectively and truthfully to the research community.

When authors submit an article for streamlined review, they must submit all documentation of the previous review process, including editorial decisions, the reviews, and responses to reviews. These may be submitted along with the manuscript itself. We also require a cover letter describing what has occurred so far in terms of the previous peer review process and what is known about why the article was rejected. Additionally, authors should provide the name of the previous journal, and whether

or not the editor(s) and reviewers have given permission for the review documents to be made openly available at JOTE alongside the article, if it is published. The editor-in-chief will assess streamlined review submissions on a case-by-case basis, and the emphasis will be on reusing the existing reviews, rather than engaging with the reason for why the article did not proceed to publication at the previous outlet. More details about the process can be found on JOTE's submission guidelines page.

JOTE Introduces Fully Open Peer Review

The third change that we are implementing concerns open peer review. It is a smaller-scale shift in our operation, but nonetheless an important one. Until now, the default peer review option in our system was double-blind, masking both the identity of the authors and reviewers. Double-blind peer review undoubtedly has its benefits. For instance, complete anonymity is thought to reduce bias (though in practice, depending on how small or specialized a scientific discipline is, complete anonymity is often an unachievable goal). However, a lack of transparency leads to a lack of accountability for both authors and peer reviewers. We are committed to transparency in the peer review process and aim to allow reviewers to get recognition for their work. While our system has yet to be configured to formally require that peer-reviewers identify themselves, this change is in progress and will be realized.

To implement open peer review, we will require that both authors' and reviewers' identities are known during the peer review process. This aligns well with our existing practices of allowing readers to comment on articles via PubPub, encouraging the use of a preprint server for the unpublished manuscript, and requiring that authors either share their empirical data or give us permission to share it on their behalf (in line with the GDPR and other ethical guidelines). We are also working on making all peer reviews and response letters open, accompanying published articles. In our view, transparency and accountability are crucial parts of normalizing error and the human element that underpins the scientific enterprise.

A Closing Comment from the New Editor-in-Chief

We close this editorial with a final remark from our new editor-in-chief, Dr. Sarahanne M. Field:

> *I follow an inspiring and successful predecessor, Maura Burke, who has served as the EiC of JOTE since 2020. Maura's initial work at JOTE is an incredibly hard act to follow, and the scientific community at large owes her and the rest of our colleagues at JOTE a debt for the work they have already done, paving the way for normalizing and embracing error, as the first journal of its kind in our industry. I thank my colleagues at JOTE for giving me the opportunity to take up this role, and for trusting me with the heavy responsibility it entails. I hope that I can serve both the journal and the wider research community well at the helm of this unique and avant-garde publishing outlet in the coming years.*

For full transparency, and in the spirit of scientific integrity, I wish to disclose that I receive a small, taxed financial stipend for my work in the role of editor-in-chief for this journal.

References

Devine, S., Bautista-Perpinya, M., Delrue, V., Gaillard, S. D. M., Jorna, T. F. K., van der Meer, R. M., Millett, L., Pozzebon, C., & Visser, J. (2020). Science fails. Let's publish. *Journal of Trial and Error*, *1*(1), 1-5. https://doi.org/10.36850/ed1

Field, S. M., Kiers, H., & Derksen, M. (2024). Exploring the constellation of communities shaping science reform and its future progress. Preprint under review.

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*(4), Article 181351. https://doi.org/10.1098/rsos.181351

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (2017). The evolution of crew resource management training in commercial aviation. In R. K. Dismukes (Ed.), *Human error in aviation*

(pp. 275-288). Routledge. https://doi.org/10.4324/9781315092898-15

McCreary, J., Pollard, M., Stevenson, K., & Wilson, M. B. (1998). Human factors: Tenerife revisited. *Journal of Air Transportation World Wide, 3*(1), 23-32.

National Transportation Safety Board. (2004). *Loss of pitch control during takeoff Air Midwest Flight 5481* (NTSB/AAR-04/01). Federal government of the United States of America. https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR0401.pdf

Potti, A., Dressman, H. K., Bild, A., Riedel, R. F., Chan, G., Sayer, R., Cragun, J., Cottrill, H., Kelley, M. J., Petersen, R., Harpole, D., Marks, J., Berchuck, A., Ginsburg, G. S., Febbo, P., Lancaster, J. M., Nevins, J. R., & Nevins, J. R. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, *12*(11), 1294-1300. https://doi.org/10.1038/nm1491 (Retraction published 2011, *Nature Medicine*, *17*(1), 135)

Salmieri, G. (2024). Free speech as a right and a way of life. In T. Smith (Ed.), *The first amendment: Essays on the imperative of intellectual freedom* (pp. 193-239). Ayn Rand Institute Press.

van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023 — a new record. *Nature, 624*, 479-481. https://doi.org/10.1038/d41586-023-03974-8

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103), 637-641. https:/ / doi.org/ 10.1016/ s0140-6736(97)11096-0 (Retraction published 2004, *The Lancet*, *363*(9411), 750; 2010, *The Lancet*, *375*(9713), 445)

Preclinical Assessment of a Cannabinoid CB2 Receptor Antagonist in a Murine Model of Cerebral Malaria

Ana Borrego Escartín[1], María Gómez-Cañas[2], Soledad García Gómez-Heras[3], Patricia Marín-García[4], Javier Fernández-Ruiz[1], Amalia Diez[2]

[1]Departamento de Bioquímica y Biología Molecular, Universidad Complutense de Madrid, Facultad de Medicina, 28040 Madrid, Spain

[2]Departamento de Bioquímica y Biología Molecular, Universidad Complutense de Madrid, Facultad de Veterinaria, 28040 Madrid, Spain

[3]Departamento de Ciencias Básicas de la Salud, Facultad de Ciencias de la Salud, Universidad Rey Juan Carlos, 28933, Madrid, Spain

[4]Departamento de Especialidades Médicas y Salud Pública, Universidad Rey Juan Carlos, Facultad de Ciencias de la Salud, 28933, Madrid, Spain

Received
April 9, 2021
Accepted
April 8, 2022
Published
August 11, 2024
Issued
December 23, 2024

Correspondence
Departamento de Bioquímica y Biología Molecular, Universidad Complutense de Madrid, Facultad de Medicina, 28040 Madrid, Spain
anaborregoescartin@gmail.com

Malaria is a most important parasitic disease due to its highest impact worldwide. It results in around 200 million clinical cases and 0,5-1 million deaths per year, mainly due to cerebral malaria (CM), a life-threatening neurological syndrome that predominantly affects children under five years old. CM follows neurological alterations leading to the death if left untreated, and, even when it is treated, it is fatal in 15-20% of cases. Moreover, among the survivors, more than 10% of the children develop neurological sequelae. Consequently, there is an urgent need to find therapies to attenuate these neurological signs. Recent evidence has proposed the endocannabinoid system, which plays an important neuromodulatory function in the central nervous system (CNS), also including immunomodulation preferentially exerted by CB_2 receptor. Previous studies have shown that the genetic ablation of this receptor improved mice survival against CM, suggesting a potential for the pharmacological treatment of CM with selective antagonists of this receptor. Considering this background, we investigated CM therapy by a classic CB_2 antagonist SR144528 in a murine model of the disease. First, we carried out binding studies with SR144528 to confirm its pharmacodynamic profile (binding affinity [Ki] value = 2.34 ± 0.61 *nM*; and efficacy [IC_{50}] = 96.17 ± 1.41 *nM*, at the CB_2 receptor). Second, *P. berghei* ANKA infected C57BL/6 mice were treated daily with SR144528 and assessed for parasitemia growth and neurological alterations. 30% of the treated mice showed partial recovery of CM symptoms with 20% increased survival, but finally succumbing to hyperparasitemia and severe anemia. These preliminary preclinical results suggest that, although part of the CM course might be modulated by the pharmacological blockade of the CB_2 receptor, other elements trigger the lethal outcome. Thus, while our hypothesis could not be completely validated in this CM model, we detail here all obtained results for further research.

Keywords *cerebral malaria, CB_2 receptor, antimalarial, antagonist, pharmacodynamics*

Malaria is a persistent major global health challenge, affecting more than one third of the world's population. In fact, malaria is one of the three most important infectious diseases worldwide in its impact, particularly in terms of deleterious economic consequences morbidity and mortality, affecting particularly to young children under five (Hunt et al., 2006).

Malaria in humans is a mosquito-borne disease caused by infection with one of the following protozoan *Plasmodium parasite species: P. falciparum, P. vivax, P. malariae, P. ovale* and *P. knowlesi.* The infection causes a wide variety of clinical symptoms, ranging in severity from asymptomatic or flu-like illness to life-threatening complications leading to death. In fact, clinical malaria disease can be classified as uncomplicated or severe (Bartoloni & Zammarchi, 2012). Uncomplicated malaria is mainly accompanied by fever with mild symp-

Take-home Message

Studies have shown that the inactivation of the cannabinoid type-2 receptor (CB_2) could have therapeutic value against cerebral malaria (CM). We investigated this potential in vivo with SR144528, a selective CB_2 antagonist. Although further studies would be necessary, our results suggest that SR144528 could be used as an adjuvant therapy.

toms resolving spontaneously in 10 to 30 days without complications. On the other side, severe malaria, mostly caused by infection with *P. falciparum*, is a potentially fatal disease with a quick progression in which most patients (mainly children, pregnant women, and the elderly) need to be assessed and treated quickly. Clinical presentations of severe malaria vary, but include altered consciousness, respiratory distress, acidosis, severe anemia, multi-organ failure, and CM including coma. Particularly, the latter is considered one of the most life-threatening complications.

CM occurs predominantly in patients with little or no background immunity, including children aged 2-6 years growing up in endemic areas of Africa, or adults who have not acquired immunity to malarial infection or have lost their immunity to the disease (Grau et al., 1987). This severe complication represents an enormous burden of disease due to the high prevalence of infection (Medana & Turner, 2006) and it is considered the most important parasitic CNS disease worldwide, accounting for 80% of all fatal cases of malaria (Linares et al., 2013). If left untreated, CM is nearly always fatal and, even when treated, this neurological syndrome has an approximate 15-30% of mortality rate (Bartoloni & Zammarchi, 2012). Among survivors, moreover, more than 10% of the affected children have neurological sequelae. Long-term cognitive impairments have been reported in one of every four child survivors of CM (Mariotti & Bertini, 2011).

CM is a complex condition whose pathogenesis is not completely understood. Several hypotheses have been raised to explain the pathophysiology of the disease, but, none of them explain the pathogenesis by themselves. Many authors suggest that the combination of mechanical obstruction of microvessels by parasitized red blood cells (pRBC) and exacerbated neuroinflammation are the main two mechanisms triggering the neurological syndrome (Combes et al., 2006).

Given the available knowledge on CM pathogenesis, active treatments for this neurological condition are still limited. Recent evidence suggests that cannabinoids and, in general, the modulators of the endocannabinoid system (ECS) may have therapeutic value in this disease. The ECS is an intercellular communication system, widely distributed in the organism, constituted by cannabinoid receptors, endogenous ligands and the enzymatic machinery for their biosynthesis and hydrolysis. To date, two major cannabinoid receptors are known and well-characterized, which are called CB_1 and CB_2 receptors. Both belong to the G-protein-coupled receptor (GPCR) superfamily. The wide distribution of cannabinoid receptors in the body, and the multitude of cellular signaling mechanisms in which this system is involved, suggest that ECS has critical physiological and pathophysiological significance. Among the many ECS regulatory physiological control of different systems organs and tissues, the involvement of this system in neuroprotection processes has been vastly studied, resulting as a great drug target in the treatment of neurodegenerative diseases and acute neuronal damage. It has been observed that the ECS modulates several events which take place in this kind of pathologies, including excitotoxicity, mediated by CB_1 activation; neuroinflammation, mainly modulated by CB_2 receptor in activated glia cells; decrease of reactive oxygen species (ROS) release; and neurogenesis increase (Fogaça et al., 2013). These properties are particularly appealing for the potential treatment of pathogenic dysfunctions caused by cerebral malaria.

Moreover, in the last decade, cannabinoids have emerged as putative modulators of the CNS immune neuroprotection, and plastic events as well as behavioral and cognitive functions. Modulation of the ECS has led to beneficial results for several neurodegenerative diseases such as Huntington's, Alzheimer's, Parkinson's, amyotrophic lateral sclerosis, multiple sclerosis, and hypoxia-ischemia (Fernández-Ruiz et al., 2015; Pazos et al., 2012). Regarding CM, some reports have

been published highlighting the therapeutic potential of cannabinoids and other modulators of the ECS in CM (Alferink et al., 2016; Campos et al., 2015). Particularly, the CB_2 receptor has been proposed as an important modulator of susceptibility in experimental cerebral malaria (ECM). Mice with a deletion of the CB_2 encoding gene ($CB_2^{-/-}$) inoculated with *P. berghei* ANKA erythrocytes exhibited a longer survival and a diminished blood-brain-barrier disruption. Moreover, treatments in wildtype mice with a specific CB_2 antagonist also seems to confer increased ECM resistance, whereas the same happened with the phytocannabinoid cannabidiol (CBD) (Campos et al., 2015), which may act, among other things, as a negative allosteric modulator for the CB_2 receptor (Martínez-Pinilla et al., 2017). Thus, CB_2 might be a promising therapeutic target to fight CM. More precisely, targeting CB_2 through selective antagonists could lead to an increased resistance to the development of this neurological complication.

Method

Considering the previous background, selective CB_2 receptor antagonists may have therapeutic potential in CM, and the objective of this study has been to further explore this potential. In this context, we confirmed first the pharmacodynamic profile of SR144528 as a selective ligand with antagonist activity at the CB_2 receptor (Portier et al., 1999; Rinaldi-Carmona et al., 1998), followed by its evaluation as disease-modifying agent in an experimental model of CM.

Radioligand binding assays for CB_1 and CB_2 receptors

We first proceeded to evaluate the affinity of SR144528 at the CB_1 and CB_2 receptors by conducting competition assays. To this end, SR144528 was evaluated *in vitro* for their ability to displace [^{3}H]-CP55,940 from human cannabinoid CB_1 and CB_2 receptors transfected into HEK293 EBNA cells. SR144528 was first subjected to a preliminary screening at saturating conditions of 4 µM for both CB_1 and CB_2 receptors. Subsequently, a complete concentration-occupancy curve rom 10^{-4} to 10^{-12} *M* was carried out to obtain the inhibitor constant K_i values given that SR144528 displaced the radioligand by more than 50% in the preliminary screening.

Membranes purified from transfected cells with human CB_1 or CB_2 receptors (RB-HCB1M400UA and RBXCB2M400UA) were supplied by Perkin-Elmer Life and Analytical Sciences (Boston, MA). The final membrane protein concentration was 0.8 *µg*/well and 0.4 *µg*/well respectively for the CB_1 and the CB_2 receptor assays. The radioligand [^{3}H]-CP55940 (PerkinElmer) was used at 0,4 *nM* for CB_1 and 0,53 *nM* for CB_2. The final incubation volume was 200 *µL* for both CB_1 and CB_2 binding. 96-well plates and the tubes necessary for the experiment were previously siliconized with Sigmacote (Sigma).

Membranes were resuspended in the corresponding buffer (50 *mM* TrisCl, 5 *mM* $MgCl_2.H_2O$, 2.5 *mM* EDTA, 0.5 *mg/mL* BSA and *pH* = 7.4 for CB_1 and 50 *mM* TrisCl, 5 *mM* $MgCl_2.H_2O$, 2.5 *mM* EGTA, 1 *mg/mL* BSA and *pH* = 7.5 for CB_2) and were incubated with the radioligand and SR144528 for 90 min at 30°C. Non-specific binding was determined with 10 *µM* WIN55212-2, a well-known CB_1/CB_2 agonist (*Ki* CB_1 = 9,9 *nM*, K_i CB_2 = 16.2 *nM*) (Rinaldi-Carmona et al., 1994) used as reference compound to determine non-specific binding in the laboratory radioligand binding assays protocols for CB_1 and CB_2 receptors (Cumella et al., 2012; Deiana et al., 2016).

Total radioligand binding to the membrane was determined by its incubation with the membranes in absence of any compound. Filtration was performed by a Harvester® filtermate (Perkin-Elmer) with Filtermat A GF/C filters pretreated with polyethylenimine 0.05%. After filtering, the filter was washed nine times with binding buffer, dried and a melt-on scintillation sheet (Meltilex™ A, Perkin Elmer) was melted onto it. Then, radioactivity was quantified by a liquid scintillation spectrophotometer (Wallac MicroBeta Trilux, Perkin-Elmer). Competition binding data were analyzed by using GraphPad Prism program and K_i values are expressed as mean ± SEM of at least three experiments performed in triplicate for each point.

[^{35}S]-GTPγS binding analysis

Once the binding assays performed, we proceeded to evaluate the functional activity of

Table 1 Binding affinity of SR144528 to CB_1 and CB_2 receptors. Values were obtained from competition studies using [^{3}H]-CP 55,940 as radioligand for CB_1 and CB_2 receptors. They are expressed as the mean ± SEM of 3 different experiments each performed in triplicates. CB_2 selectivity values are expressed as the ratio: K_iCB_1 / K_iCB_2.

Compound	K_i CB_1 (nM)	K_i CB_2 (nM)	CB_1 /CB_2 ratio
SR144528	84.72 ± 21.26	2.34 ± 0.61	36.2

Table 2 IC_{50} and E_{max} values of SR144528. Both values were calculated using nonlinear regression analysis. Data are expressed as the mean ±SEM of 3 independent experiments (n=3), each one run in triplicate.

Compound	IC_{50}(nM)	E_{max}(%)	Amplitude
SR144528	96.17 ± 1.41	-63.23 ± 3.14	62.57

SR144528 at the CB_2 receptor by carrying out [^{35}S]-GTPγS binding assays. To this end, CB_2 receptor-containing membranes (HTS020M2, Eurofins Discovery Services St. Charles, MO, EEUU) were used. These membranes (5 µg/well) were permeabilized by addition of saponin (1:1 v/v), then mixed with 0,3 *nM* [^{35}S]-GTPγS (Perkin-Elmer) and 10 *µM* GDP (Sigma-Aldrich) in 20 *mM* HEPES buffer (10mM MgCl2, 10 *mM* NaCl, pH 7.4). Increasing concentrations of SR144528, from 10^{-11} to 10^{-4} *M*, were added in a final volume of 100 *µL* and incubated for 30 min at 30ºC. The non-specific signal was measured with 10 *µM* non-labeled GTPγS (Sigma-Aldrich). All 96-well plates and the tubes necessary for the experiment were previously siliconized with Sigmacote (Sigma-Aldrich). The reaction was terminated by rapid vacuum filtration with a filter mate Harvester apparatus (Perkin-Elmer) through Filtermat A GF/C filters. The filters were washed nine times with ice-cold filtration buffer (10 *mM* sodium phosphate, pH 7.4) and dried, and a melt-on scintillation sheet (Meltilex™ A, Perkin Elmer) was melted onto it. The bound radioactivity was measured with a Luminiscence counter Wallac MicroBeta TriLux (Perkin-Elmer). [^{35}S]-GTPγS binding data were analyzed to determine the IC50 values by using an interactive curve fitting procedure with the GraphPad Prism version 5.02 (Graph-Pad Software Inc.). IC50 values are expressed as mean ± SEM of at least three experiments performed in triplicate for each point.

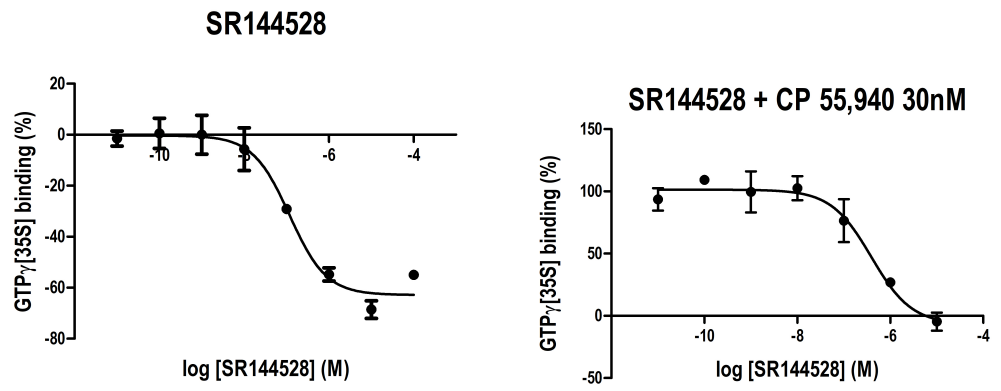
In a complementary assay, SR144528 was also investigated to elucidate their ability to antagonize the effect of an agonist. To this purpose, a second [^{35}S]-GTPγS study was performed incorporating CP 55,940 (30 *nM*) to the assay conditions. All the material, procedures of incubation, filtration, radioactivity counting, and data analysis were identical to the experimental conditions described previously.

Assessment of SR144528 in a cerebral malaria in vivo model

All experiments with animals were conducted at the Universidad Complutense de Madrid in accordance with national and international regulations for animal experimentation and were performed with fully adherence to their corresponding deontological and ethical guidelines. The study protocol was approved by the "Committee of Animal Experimentation" of the Universidad Complutense de Madrid.

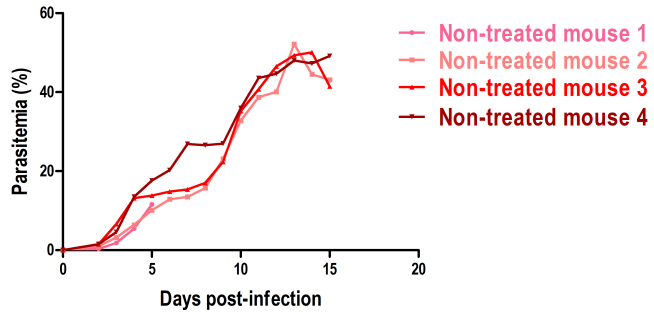
The experimental mice used were 4-weeks old pathogen-free male C57BL/6 mice, as a CM susceptible strain. The animals were purchased from Harlan Ibérica (Barcelona, Spain) and kept in our facilities at the Universidad Complutense de Madrid, with free access to food and water. Ten mice were infected by intraperitoneal injection of $5x10^6$ parasitized red bloods cells (pRBC) obtained from *P. berghei* ANKA previously infected mice. Four of them were designated as non-treated mice, whereas the remaining six animals were treated daily with intraperitoneal injections of 25 *µg* SR144528. The SR144528 doses were prepared according to previous studies (Alferink et al., 2016) by resuspending the compound in a Tween 80:saline (1:16 v/v) solution with a 1.25% DMSO. To minimize animal suffering, only two of the four non-treated mice were injected with the vehicle that included the same proportion of DMSO in Tween 80: saline. In accordance with preliminary experiments supported in previous reports, injected and non-injected control mice have identical phenotypic and histological characteristics (Marín-García et al., 2009).

Infection progress was monitored daily by staining blood smears with Wright's eosin

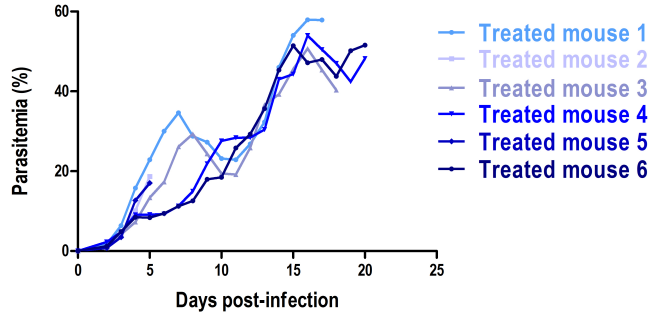


(a) (b)

Figure 1



(a)



(b)

Figure 2

methylene blue solution (Merck) followed by counting pRBC under the microscope and quantitative determination of percent parasitemia using the PlasmoScore software (version 1.3). Moreover, the neurological performance of the animals was evaluated daily in individual mice by recording clinical symptoms including ruffled fur, abnormal gait, tremor, reduced grip strength, affected startle, abnormal visual response, reduced motility, head deviation, hemi- or paraplegia, tendency to roll over on stimulation, back elevation, ataxia, and convulsions. According to previous standardizations of the experimental cerebral malaria (Linares et al., 2013; Martinez et al., 2013), infected animals were grouped in disease stages from I to IV depending on the severity of their neurological signs as follows: stage I, no neurological symptoms in the first few days of infection; stage II, incipient symptoms of CM; stage III, appreciable neurological symptoms; and stage IV, severe symptoms. After the death of the animals, their brain tissue was immediately removed, fixed in 10% formaldehyde, and embedded in paraffin. Series of 5 *μm*-thick were cut and stained with hematoxylin and eosin to elucidate and assess histopathological damages.

Results and Discussion

CB_1/CB_2 receptor binding studies for SR144528

The CB_1 and CB_2 receptor binding affinities of SR144528 were evaluated by radioligand binding assays carried out by competition with [^{3}H]-CP 55,940 and the data of receptor affinities and selectivity are shown in Table 1. These data indicate that SR144528 is able to bind at the CB_1 receptor with a *Ki* in the high *nM* range, but that it has a higher affinity for the CB_2 receptor, reaching a K_i value in the low nM range and with a CB_1/CB_2 ratio of 36.2 (Table 1).

Determination of the functional activity of SR144528 at the CB_2 receptor

Subsequently, we investigated the efficacy (agonism *versus* antagonism and/or inverse agonism) of SR144528 at the CB_2 receptor by conducting [^{35}S]-GTPγS binding assays. As expected, SR144528 behaved as inverse agonist of the CB_2 receptor (Fig. 1a) with significant values of IC_{50} and efficacy (Emax) described in the Table 2. We also conducted supplementary assays to determine the antagonistic capacity of SR144528 in the presence of the CB_1/CB_2 agonist CP55,940 (30 *nM*), which revealed a strong capability for this compound to antagonize the effects of CB_2 agonists (Fig. 1b).

In vivo assessment of SR144528 therapeutic potential in ECM

The preclinical assessment of the therapeutic effect of SR144528 as a CB_2 receptor antagonist was carried out in the *P. berghei* ANKA ECM murine model. Several studies have widely highlighted the phenotypic similarities between the experimental CM model of *P. berghei* ANKA infecting C57BL/6 mouse and the neurological disease progression in humans (De Souza et al., 2010; Hunt & Grau, 2003; Lou et al., 2001; Medana & Turner, 2006). To follow disease and infection progress in this model, peripheral blood parasitemia, and neurological phenotype alterations were monitored on a daily basis in treated and non-treated mice.

Peripheral blood parasitemia values are shown in Figure 2 for both, non-treated (Fig. 2a) and SR144528 treated mice (Fig. 2b). Values are individually plotted to show potential outliers responses, if any. Comparing non-treated and treated animals, a certain level of individual heterogeneity was observed in both groups. Thus, some animals showed a marked blood hyperparasitemia from day 10 post-infection, whereas others died early on in the first 5 days after infection with low parasitemia levels (around 15%).

Behavioral characterization of neurological alterations underlying CM progression, allowed the conclusion that, based on the non-treated mice results, the *P. berghei* ANKA-C57BL/6 model can lead to the acquisition of either non-cerebral malaria (NCM) with high levels of peripheral blood parasitemia or CM. Thus, considering the parasitemia curves and the corresponding phenotypes, we classified disease progress into three different clinical groups: a first group of CM, a second of NCM, and a third designated as 'mixed group'. Consequently, we separately reanalyzed the data according to these three clinical groups (Fig. 3).

The **experimental group of CM** (Fig. 3a) included the non-treated mouse 1 and the treated mice 2 and 5. These three animals

Original Purpose

Nowadays, the mainstay of the treatment of severe cerebral malaria is the immediate onset of parenteral antimalarial treatment. Available drugs include quinolones, antifolates and artemisin-combination therapies. However, despite the effectiveness of antimalarial drugs on treating most malaria related symptoms, their efficacy on promoting survival and preventing neurological damage, mainly in children, is contested. In this scenario, the combination of antimalarial drugs and CNS-acting compounds would be an interesting strategy for the improvement of the neurological outcome in CM patients. Previous studies have shown that the CB2 receptor could be an important modulator of susceptibility in experimental cerebral malaria (ECM). Moreover, therapeutic application of a specific CB2 antagonist also seems to confer increased ECM resistance in wild type mice. Thus, targeting CB2 and blocking its intracellular signaling might be promising for the development of alternative treatment regimens for CM.

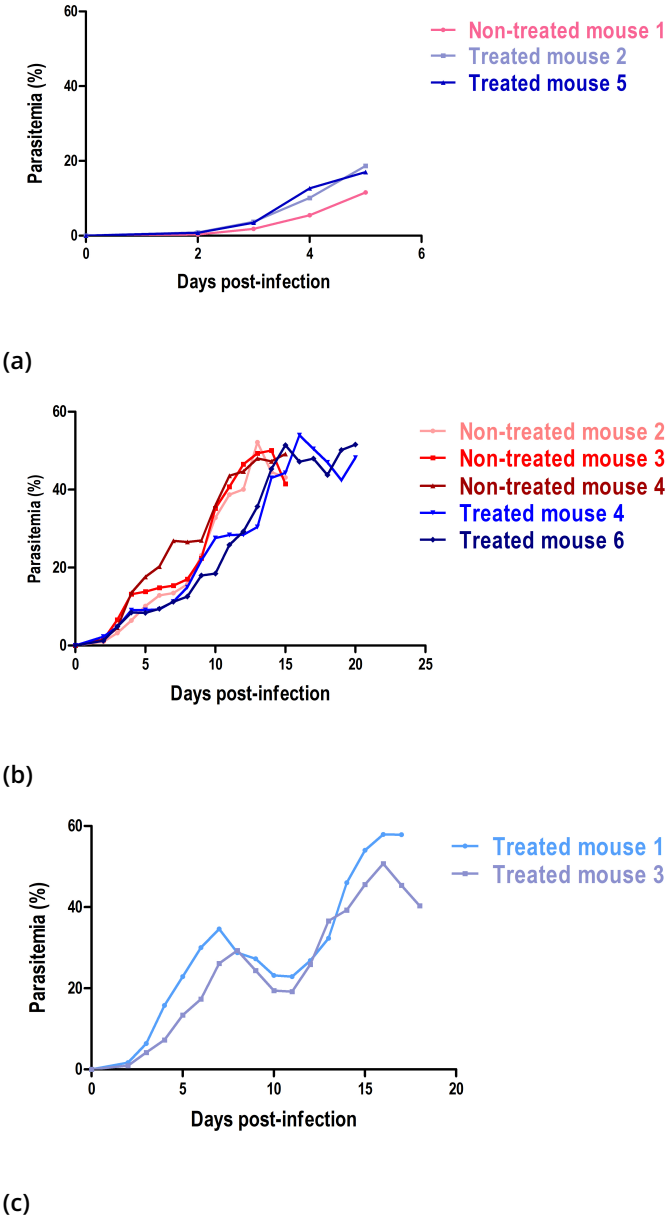


Figure 3

were classified as CM cases, considering their phenotypic alterations, blood parasitemia levels, and disease progression. These mice showed phenotypic features of brain pathology, that can be assigned to a disease stage II on day 5 post-infection, with incipient neurological alterations such as ruffled fur, head deviation, tremor, and increased respiratory rate. These mice died on day 6 post-infection with low parasitemia levels around 11-18%.

The **NCM group** (Fig. 3b) included the non-treated mice 2, 3, and 4 and the treated mice 4 and 6. These animals were classified as NCM cases, based on parasitemia development and disease progression, showing no signs or behaviors of neurological disorder. All individuals in this group evolved similarly and died within the same time interval between days 15 and 21 post-infection due to a strong hyperparasitemia around 49-53% that caused a strong weight loss and weakness of the animals.

The **'mixed group'** (Fig. 3c) enclosed treated mice 1 and 3. In these two mice, the progression of the disease took place in three different phases, clearly associated with their parasitemia level profiles: a first stage of CM between days 1-7 post-infection; a second phase of recovery between days 8 and 11; and a third stage in which the animals acquired NCM, from day 12 post-infection until death. In phase I, the animals developed a phenotypic and behavioral evolution similar to the CM group. On day 7 post-infection, concurring with a parasitemia peak around 25-30% (Fig. 3c) showed clinical manifestations of stage III CM, including ruffled fur, back elevation, head deviation, ataxia, slight hemi-paralysis of the hindlimbs, and lack of response to stimuli. Nevertheless, these two animals did not progress to coma and death but from day 8 post-infection, the animals entered what can be considered a phase II with a tendency to recovery by the observed decrease in parasitemia levels to 19-22%. This reduction in parasitemia values was also reflected in the neurological signs with recovered mobility, response to stimuli and no tremor or head deviation. Finally, from day 12 post-infection, their clinical and parasitemia evolution matched with the mice included in the NCM group, acquiring very high levels of parasitemia, around 50-55%, without neurological alterations but dying at day 17-19 post-infection due to critical weakness, marked weight loss and severe anemia.

According to this clustering, SR144528 administration to the CM and NCM groups had no therapeutic effect. The evolution of the infection did not change between treated and non-treated mice in CM and NCM groups, finally ending in death at the same time, either by CM or by severe anemia, respectively (Figs. 3a-3b).

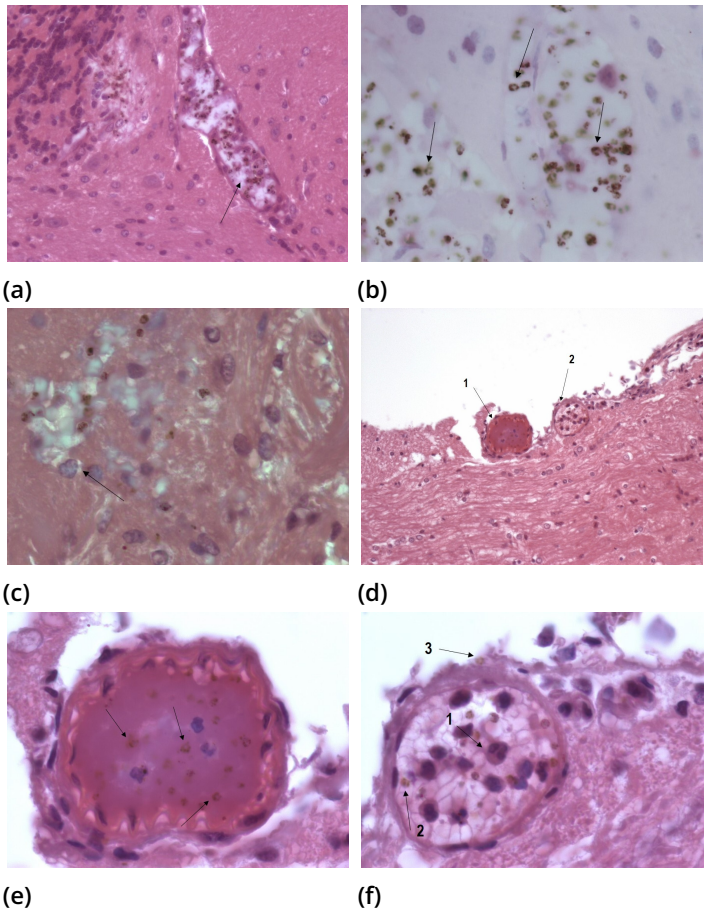


Figure 4

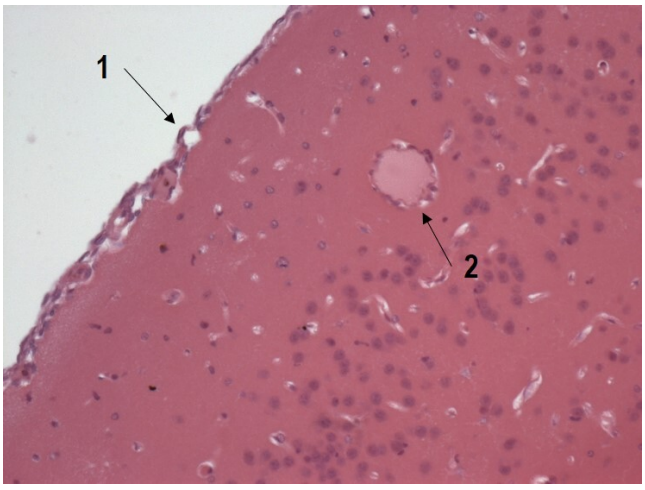
However, in the 'mixed group', it can be observed that the treatment might have a slight and transient favorable effect, reducing temporarily parasitemia with remission of neurological signs associated with CM (Fig. 3c).

Furthermore, post-mortem studies showed histopathological features that agreed the classification into three groups based on the clinical manifestations above established. Mice included in the CM group showed clear signs of cerebral alterations at meningeal, cerebellar, and cortical level (Fig. 4). Remarkable clusters of inflammatory cells can be observed in several blood vessels at the white and gray matter and meninges, as well as sequestration of pRBC with mature forms of the parasite in the cerebral vasculature (Figs. 4a-4b). Interestingly, typical erythrocyte Rosseting (Fig. 4c), i.e., aggregations of noninfected erythrocytes and pRBCs contributing to the microvascular obstruction, support further signs of severe malaria (Ross et al., 2007) in this CM group. Moreover, there were signs of blood brain barrier (BBB) disruption with leukocyte and parasitic extravasation to the brain tissue, diffuse microhemorrhages and areas of edema and thrombosis (Figs. 4d-4f). No distinctive histopathological differences were observed between the brain sections of the treated and non-treated animals included in the CM group.

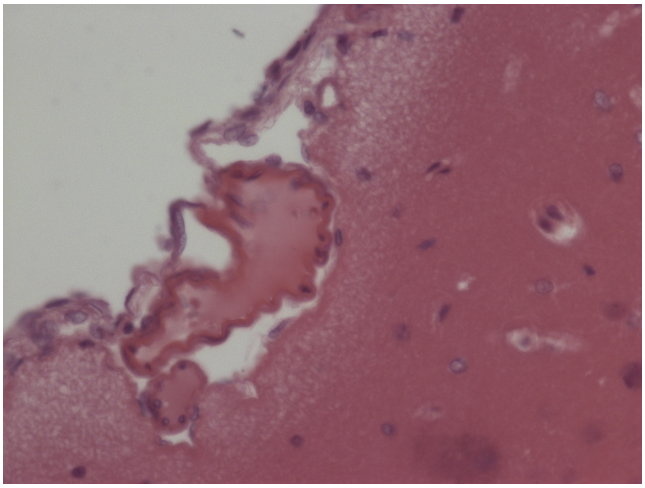
Mice classified within the NCM group, either SR144528 treated or non-treated, have minimal brain damage with diffuse microhemorrhages but without meningeal involvement (Fig. 5a). No parasite sequestration or clusters of circulating inflammatory cells were identified (Fig. 5b). There was no evidence of disruption of the BBB. As in the previous group, no differences were observed in brain sections between non-treated animals (non-treated mice 2, 3, and 4) and treated ones (treated mice 4 and 6) that undergone NCM form upon *P. berghei* infection.

Mice included in the 'mixed group', showed some diffuse microhemorrhages, small infiltrations of inflammatory cells, and parasitic sequestration in microvessels (Fig. 6b-6d) but with no meningeal involvement (Fig. 6a) or signs of BBB disruption. These observations suggest that the treatment with SR144528 in these mice slowed down brain damage during the infectious process, as also evidenced by both, the parasitemia curves (Fig. 3c) and the neurological phenotypes shown. Thus, our results agree with previous studies where CB_2 knockout mice ($Cnr2^{-/-}$) exhibit enhanced survival with reduced parasite load in the brain and a diminished BBB disruption (Alferink et al., 2016). SR144528 treatment in these two mice resulted in an increase in their survival time, reaching day up to day -20 post-infection. However, the compound did not show an antimalarial effect sufficient to eliminate circulating parasites, since parasitemia increased again (Fig. 3c) from day -10 onwards, with no neurological manifestations, but to finally cause death by hyperparasitemia and severe anemia.

Due to the low number of animals, the experimental model is inconclusive and, therefore, our initial hypothesis could not be validated. However, these partial results suggest that CM severity could be modulated by the



(a)



(b)

Figure 5

pharmacological blockade of the CB_2 receptors. In addition, other factors might be involved in the further disease evolution. In the first instance, activation of the CB_2 receptor has been associated with a decrease of exacerbated inflammation in several pathologies (Fernández-Ruiz et al., 2006, 2015; Morales et al., 2016; Turcotte et al., 2016), in other studies it has been demonstrated that this receptor could also contribute to tissue damage (Pacher & Mechoulam, 2011). Different reports have postulated that modulation of the endocannabinoid signaling through the CB_2 receptor, both by agonists and inverse agonists/antagonists, could have a relevant therapeutic potential in a large number of diseases by reducing the inflammatory and chemotactic response and attenuating the endothelial activation and cell adhesion (Pacher & Mechoulam, 2011). Regarding CM, in addition to the presumable anti-inflammatory activity that might counteract the exacerbated inflammation of this brain process, inhibition of the CB2 receptor by the SR144528 antagonist could also contribute to the maintenance of the integrity of the BBB. In fact, activation of CB_2 receptor triggers (among other things), MAPK signaling, which has been shown involved in BBB disruption and in vivo neurological symptoms of severe malaria (Picone & Kendall, 2015; Rhee & Sang-Keun, 2002). By blocking CB_2 receptor, the MAPK pathway would be inhibited, thus maintaining BBB integrity.

Conclusion

In this study, we hypothesized modulation and therapeutic potential of CB_2 receptor in the pathophysiology of experimental CM. Through blockade by SR144528 (a CB_2 receptor selective antagonist) in an in vivo murine model of CM, 30% of the treated mice showed partial recovery of CM symptoms with 20% increased survival while finally succumbing to hyperparasitaemia and severe anemia. Although other factors seem to be involved in controlling the infection and our results are inconclusive, the present observations provide valuable experimental information for the development of alternative treatment regimens for CM by combining classic antimalarial drugs and neuroprotective compounds targeting CB_2. Thus, we suggest that SR144528 could be used as an adjuvant in the treatment of CM for a rescue therapy that could prevent or eliminate neurological sequelae in individuals who survive the infection. Further experimental pharmacological studies would be interesting to elucidate optimal candidates.

References

Alferink, J., Specht, S., Arends, H., Schumak, B., Schmidt, K., Ruland, C., Lundt, R., Kemter, A., Dlugos, A., Kuepper, J. M., Poppensieker, K., Findeiss, M., Albayram, N., Otte, D. M., Marazzi, J., Gertsch, J., Förster, I., Maier, W., Scheu, S., … Zimmer, A. (2016). Cannabinoid receptor 2

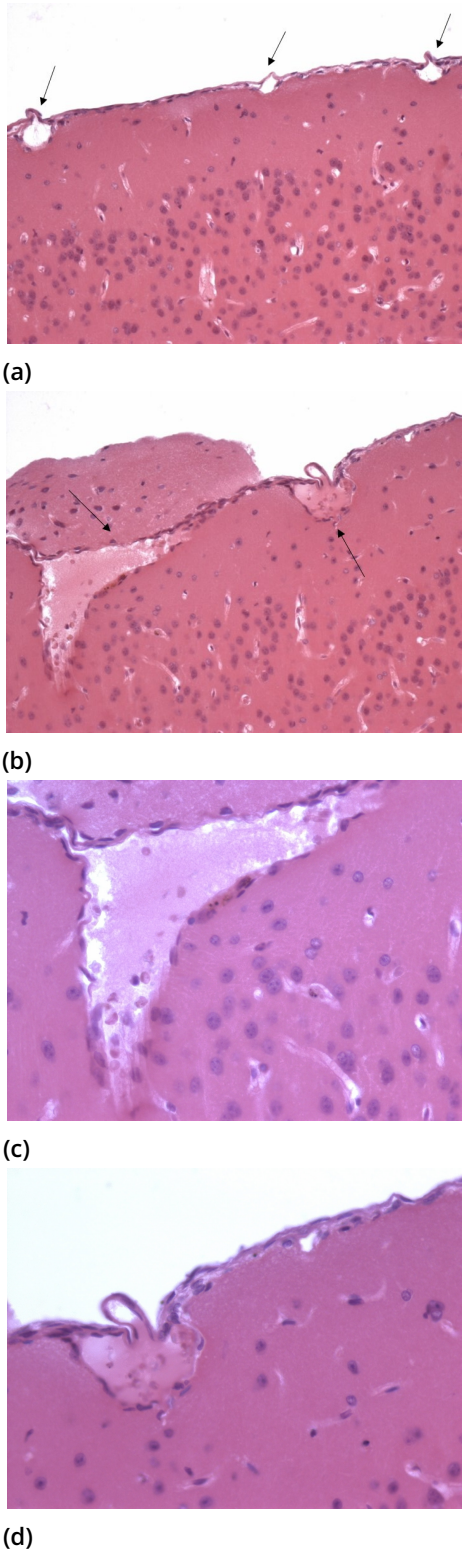


Figure 6

modulates susceptibility to experimental cerebral malaria through a ccl17-dependent mechanism. *Journal of Biological Chemistry*, *291*(37), 19517–19531. https://doi.org/10.1074/jbc.M116.746594 (see pp. 8, 9, 13).

Bartoloni, A., & Zammarchi, L. (2012). Clinical aspects of uncomplicated and severe malaria. *Mediterranean journal of hematology and infectious diseases*, *4*(1), e2012026. https://doi.org/10.4084/MJHID.2012.026 (see pp. 6, 7).

Campos, A. C., Brant, F., Miranda, A. S., Machado, F. S., & Teixeira, A. L. (2015). Cannabidiol increases survival and promotes rescue of cognitive function in a murine model of cerebral malaria. *Neuroscience*, *289*, 166–180. https://doi.org/10.1016/j.neuroscience.2014.12.051 (see p. 8).

Combes, V., Coltel, N., Faille, D., Wassmer, S. C., & Grau, G. E. (2006). Cerebral malaria: Role of microparticles and platelets in alterations of the blood-brain barrier. *International Journal for Parasitology*, *36*(5), 541–546. *https://doi.org/10.1016/j.ijpara.2006.02.005* (see p. 7)

Cumella, J., Hernández-Folgado, L., Girón, R., Sánchez, E., Morales, P., Hurst, D. P., Gómez-Cañas, M., Gómez-Ruiz, M., Pinto, D. C. G. A., Goya, P., Reggio, P. H., Martin, M. I., Fernández-Ruiz, J., Silva, A. M. S., & Jagerovic, N. (2012). Chromenopyrazoles: Non-psychoactive and selective cb 1 cannabinoid agonists with peripheral antinociceptive properties. ChemMedChem, *7*(3), 452–463. https://doi.org/10.1002/cmdc.201100568 (see p. 8).

De Souza, J., Hafalla, J. C. R., Riley, E. M., & Couper, K. N. (2010). Cerebral malaria: Why experimental murine models are required to understand the pathogenesis of disease. https://doi.org/10.1017/S0031182009991715 (see p. 11).

Deiana, V., Gómez-Cañas, M., Pazos, M. R., Fernández-Ruiz, J., Asproni, B., Cichero, E., Fossa, P., Muñoz, E., Deligia, F., Murineddu, G., García-Arencibia, M., & Pinna, G. A. (2016). Tricyclic pyrazoles. part 8. synthesis, biological evaluation and modelling of tricyclic pyrazole carboxamides as potential cb2 receptor ligands with antagonist/inverse agonist properties. *European Journal of Medicinal Chemistry*, *112*, 66–80. https://doi.org/10.1016/J.EJMECH.2016.02.005 (see p. 8).

Fernández-Ruiz, J., Romero, J., & Ramos, J. A. (2015). Endocannabinoids and neurodegenerative disorders: Parkinson's disease, hunt-

ington's chorea, alzheimer's disease, and others. *Endocannabinoids*, *231*, 33–259. https://doi.org/10.1007/978-3-319-20825-1_8 (see pp. 7, 14).

Fernández-Ruiz, J., Romero, J., Velasco, G., Tolón, R. M., Ramos, J., & Guzmán, M. (2006). Cannabinoid cb 2 receptor : A new target for controlling neural cell survival? *28*(1). https://doi.org/10.1016/j.tips.2006.11.001 (see p. 14).

Fogaça, M. V., Galve-Roperh, I., Guimarães, F. S., & Campos, A. C. (2013). Cannabinoids, neurogenesis and antidepressant drugs: Is there a link? *Current neuropharmacology*, *11*(3), 263–275. https://doi.org/10.2174/1570159X11311030003 (see p. 7).

Grau, G. E., Fajardo, L. F., Piguet, P. -F., Allet, B., Lambert, P. -H., & Vassalli, P. (1987). Tumor necrosis factor (cachetin) as an essential mediator in murine cerebral malaria. *Science*, *237*(4819), 1210–1213 (see p. 7).

Hunt, N. H., Golenser, J., Chan-Ling, T., Parekh, S., Rae, C., Potter, S., Medana, I. M., Miu, J., & Ball, H. J. (2006). Immunopathogenesis of cerebral malaria. *International Journal for Parasitology*, *36*(5), 569–582. https://doi.org/10.1016/J.IJPARA.2006.02.016 (see p. 6).

Hunt, N. H., & Grau, G. E. (2003). Cytokines: Accelerators and brakes in the pathogenesis of cerebral malaria. *Trends in Immunology*, *24*(9), 491–499. https://doi.org/10.1016/S1471-4906(03)00229-1 (see p. 11).

Linares, M., Marín-García, P., Pérez-Benavente, S., Sánchez-Nogueiro, J., Puyet, A., Bautista, J. M., & Diez, A. (2013). Brain-derived neurotrophic factor and the course of experimental cerebral malaria. *Brain Research*, *1490*, 210–224. https://doi.org/10.1016/j.brainres.2012.10.040 (see pp. 7, 10).

Lou, J., Lucas, R., & Grau, G. E. (2001). Pathogenesis of cerebral malaria. Recent *Experimental Data and Possible Applications for Humans*, *14*(4), 810–820. https://doi.org/10.1128/CMR.14.4.810 (see p. 11).

Marín-García, P., Sánchez-Nogueiro, J., Diez, A., León-Otegui, M., Linares, M., García-Palencia, P., Bautista, J. M., & Miras-Portugal, M. T. (2009). Altered nucleotide receptor expression in a murine model of cerebral malaria. *The Journal of Infectious Diseases*, *200*(8), 1279–1288. https://doi.org/10.1086/605896 (see p. 9)

Mariotti, R., & Bertini, G. (2011). Neuroinflammation and brain infections: Historical context and current perspectives. *Brain Research Reviews*, *66*(1-2), 152–173. https://doi.org/10.1016/J.BRAINRESREV.2010.09.008 (see p. 7).

Martinez, G., Linares, M., Marin-Garcia, P., Benavente, S. P., Puyet, A., Bautista, J., & Diez, A. (2013). *Anales de la real academia nacional defarmacia* (Vol. 79). (See p. 10).

Martínez-Pinilla, E., Varani, K., Reyes-resina, I., Angelats, E., Vincenzi, F., Ferreiro-vera, C., Oyarzabal, J., Canela, E. I., Lanciego, J. L., Nadal, X., Navarro, G., Borea, P. A., Franco, R., & Lane, J. R. D. (2017). Binding and signaling studies disclose a potential allosteric site for cannabidiol in cannabinoid cb 2 receptors. *Frontiers in Pharmacology*, *8*(October), 1–10. https://doi.org/10.3389/fphar.2017.00744 (see p. 8).

Medana, I. M., & Turner, G. D. H. (2006). Human cerebral malaria and the blood–brain barrier. *International Journal for Parasitology*, *36*(5), 555–568. https://doi.org/10.1016/J.IJPARA.2006.02.004 (see pp. 7, 11).

Morales, P., Gómez-Cañas, M., Navarro, G., Hurst, D. P., Carrillo-Salinas, F. J., Lagartera, L., Pazos, R., Goya, P., Reggio, P. H., Guaza, C., Franco, R., Fernández-Ruiz, J., & Jagerovic, N. (2016). Chromenopyrazole, a versatile cannabinoid scaffold with in vivo activity in a model of multiple sclerosis. *Journal of medicinal chemistry*, *59*(14), 6753–6771. https://doi.org/10.1021/acs.jmedchem.6b00397 (see p. 14).

Pacher, P., & Mechoulam, R. (2011). Is lipid signaling through cannabinoid 2 receptors part of a protective system? *Progress in lipid research*, *50*(2), 193–211. https://doi.org/10.1016/j.plipres.2011.01.001 (see p. 14).

Pazos, M. R., Cinquina, V., Gómez, A., Layunta, R., Santos, M., Fernández-Ruiz, J., & Martínez-Orgado, J. (2012). Cannabidiol administration after hypoxiaischemia to newborn rats reduces long-term brain injury and restores neurobehavioral function. *Neuropharmacology*, *63*(5), 776–783. https://doi.org/10.1016/j.neuropharm.2012.05.034 (see p. 7).

Picone, R. P., & Kendall, D. A. (2015). Minireview: From the bench, toward the clinic: Therapeutic opportunities for cannabinoid receptor modulation. *Molecular Endocrinology*, *29*(6), 801–813. https://doi.org/10.1210/me.2015-1062 (see p. 14).

Portier, M., Rinaldi-Carmona, M., Pecceu, F., Combes, T., Poinoit-Chazel, C., Calandra, B., Barth, F., Le Fur, G., & Casellas, P. (1999). Sr 144528, an antagonist for the peripheral cannabinoid receptor that behaves as an inverse agonist - pubmed. *Journal of Pharmacology and Experimental Therapeutics*, *288*(2), 582–589 (see p. 8).

Rhee, M. -H., & Sang-Keun, K. (2002). Sr144528 as inverse agonist of cb2 cannabinoid receptor. *Journal of Veterinary Science*, *3*(3), 179–184 (see p. 14).

Rinaldi-Carmona, M., Barth, F., Héaulme, M., Shire, D., Calandra, B., Congy, C., Martinez, S., Maruani, J., Néliat, G., Caput, D., Ferrara, P., Soubrié, P., Brelière, J. C., & Le Fur, G. (1994). Sr141716a, a potent and selective antagonist of the brain cannabinoid receptor. *FEBS letters*, *350*(2-3), 240–244. https://doi.org/10.1016/0014-5793(94)00773-X (see p. 8).

Rinaldi-Carmona, M., Barth, F., Millan, J., Derocq, J. M., Casellas, P., Congy, C., Oustric, D., Sarran, M., Bouaboula, M., Calandra, B., Portier, M., Shire, D., Brelière, J. C., & Le Fur, G. (1998). Sr 144528, the first potent and selective antagonist of the cb2 cannabinoid receptor. *Journal of Pharmacology and Experimental Therapeutics*, *284*(2), 644–650 (see p. 8).

Ross, M. H., Pawlina, W., & Negrete, J. H. (2007). Histología : Texto y atlas color con biología celular y molecular. *Médica Panamericana*. (See p. 13).

Turcotte, C., Blanchet, M. -R., Laviolette, M., & Flamand, N. (2016). The cb2 receptor and its role as a regulator of inflammation. *Cellular and molecular life sciences : CMLS*, *73*(23), 4449–4470. https://doi.org/10.1007/s00018-016-2300-4 (see p. 14).

The Impact of Incentivization on Recruitment, Retention, Data Quality, and Participant Characteristics in Ecological Momentary Assessments

Helge Giese [1,2], Laura M König [3,4]

[1] Department of Psychology, University of Konstanz

[2] Heisenberg Chair for Medical Risk Literacy and Evidence-based Decisions, Center of Anestesiology and Intensive Care Medicine, Charité - Universitätsmedizin Berlin

[3] Faculty of Life Sciences: Food, Nutrition and Health, University of Bayreuth

[4] Faculty of Psychology, University of Vienna

Received
November 13, 2023
Accepted
September 27, 2024
Published
November 16, 2024
Issued
December 23, 2024

Correspondence
Charité - Universitätsmedizin Berlin
helge.giese@charite.de

Ecological Momentary Assessment (EMA) study participation is usually incentivized using monetary (e.g., fixed or performance-contingent payment) or non-monetary (e.g., feedback) compensation. This study investigates the impact of this incentivization on recruitment, retention, data quality, and participant characteristics in a sample of 74 students. For this purpose, an EMA study (time-based sampling) was conducted in participants' daily life using a 2 Payment (fixed/ performance-contingent) x 2 Feedback (yes/ no) experimental between-subjects design. Offering feedback increased the likelihood of participation and reduced the likelihood of participants receiving fixed payment to drop out. Offering feedback additionally improved data quality. Furthermore, offering feedback attracted participants with higher interest in research and the study topic. Offering fixed vs performance-contingent payment had little effect on the outcomes of interest. Offering feedback as compensation in EMA studies may facilitate recruitment and increase data quality; however, it may also risk higher selection bias. Conclusions are drawn from a relatively small student sample; the results thus need to be replicated in larger and more diverse samples.

Keywords *ambulatory assessment, experience sampling, compensation, research participation effects, mood*

Ecological Momentary Assessment (EMA), i.e., the repeated assessment of study participants in real-time and real-life, may help to overcome limitations of traditional self-report measures such as recall bias in questionnaires. It furthermore allows to study phenomena as they naturally occur in different contexts in daily life, and repeatedly over time, so increasing ecological validity compared to questionnaire or experimental laboratory studies (Shiffman et al., 2008). EMA is therefore increasingly used in research to investigate behaviors and their determinants (Perski et al., 2023), thereby contributing to the formulation and testing of theory (Scholz, 2019) and interventions (Berli et al., 2021).

To draw meaningful conclusions from EMA data with sufficient certainty, large samples generating large numbers of responses are required (Bolger & Laurenceau, 2013). EMA usually requires participants to complete several assessments per day over the course of one to several weeks (König, Van Emmenis, et al., 2022; Wrzus & Neubauer, 2022), which may complicate recruitment and retention due to reduced willingness of taking part in or completing the study (Eisele et al., 2022). Furthermore, participants might miss prompts or events to record (Ziesemer et al., 2020), or respond carelessly (Eisele et al., 2022), which leads to reduced data quality.

To overcome this challenge, researchers commonly offer monetary incentivization for EMA study participation (Ottenstein & Werner,

Take-home Message

Incentivization methods in EMA studies need to be carefully selected not only based on feasibility considerations and availability of funding, but also based on the potential impact on attrition, data quality, and certain participant characteristics. However, the current results need to be replicated in larger samples to test generalizability to more diverse samples and other study topics.

2021; Wrzus & Neubauer, 2022), which has previously been shown to improve response rates in several study designs (Edwards et al., 2009; Gillies et al., 2021; Keusch, 2015; Voslinsky & Azar, 2021). Some EMA studies offer fixed payment, i.e., all participants receive the same amount of money irrespective of how many surveys or study days they completed; others offer performance-contingent payment, e.g., by offering a certain amount of money per answered prompt. Yet again other studies offer non-monetary incentives such as personalized feedback based on the collected data (Ottenstein & Werner, 2021; Wrzus & Neubauer, 2022).

A reason for non-monetary incentives beyond budget constraints is the hope of researchers that personalized feedback may improve recruitment and retention due to increased personal relevance and motivation (Pratap et al., 2020; van Gelder et al., 2018): If participants would like high quality feedback they have to participate conscientiously. However, providing personalized feedback might also increase recruitment bias, as it may especially appeal to people with a strong interest in the topic (Cheung et al., 2017). This may aggravate existing self-selection biases, as participants in behavioral research are healthier (Haynes & Robinson, 2019), younger, wealthier, and more highly educated (Henrich et al., 2010; Pratap et al., 2020) than the population average, limiting both the generalizability and the potential impact of this research.

Ottenstein and Werner (2021) compared the impact of incentivization on retention and data quality in EMA studies, indicating that performance-contingent monetary incentivization may increase compliance compared to fixed payment. However, they were unable to draw conclusions regarding differences between monetary and non-monetary incentives since the number of studies reporting to have offered non-monetary incentives was small. Even more importantly, incentivization schemes could only be compared across studies since direct experimental comparisons are rare (Ottenstein & Werner, 2021; Wrzus & Neubauer, 2022). Studies that include at least quasi-experimental incentive conditions merely indicate that not providing any rewards harms EMA participation rates (Ludwigs et al., 2020) or that course credit is less valued than monetary or prize rewards (Harari et al., 2017).

This study therefore aims to provide a comprehensive experimental test of fixed vs performance-contingent payment and additional provision of feedback as non-monetary incentivization on recruitment, retention, data quality (operationalized as internal consistency per participant), and participant characteristics in an EMA study. We expected monetary incentivization to impact recruitment, retention, and data quality through fixed payments. Specifically, we expected fixed payment to yield more participants initially recruited (compared to performance-contingent payment) since the outlook to receiving a fixed payment is more attractive (due to higher attractiveness of the reward; H1). At the same time, we expected fixed payment (vs. performance-contingent payment) to lead to more drop-outs (i.e. participants not recording data until the last day) since, in this condition, payment is certain independently of the performance (H2). Based on the same reasoning, we expected fixed payment to yield fewer prompts being answered compared to performance-contingent payment (H3). Finally, we expected fixed payment to yield better data quality compared to performance-contingent payment, as indicated by the internal consistency of the scales (H4). This is because participants might be more intrinsically motivated to provide a response if they receive the same amount of money independent of whether they respond to the prompt or not, while performance-contingent payment might induce speeding to make sure that the prompt, and thus the payment, will not be missed.

Furthermore, we expected feedback provision to impact recruitment, retention and

participant characteristics. Specifically, we expected feedback provision to decrease dropout (H5) and to increase data quality (H6) since the recorded data will become more meaningful to the individual with the promised feedback. We expected feedback provision to increase the number of people initially interested in the study (H7), but also to attract more participants with a stronger interest in the study topic (H8) since the feedback will allow them to learn about themselves. Finally, we also explored differences in demographic and psychological participant characteristics between groups.

Materials and methods

The study procedure and data analysis plan were preregistered prior to data collection (https://osf.io/cwb3h/). Materials and data are available in the Open Science Framework (https://osf.io/zbe5q/).

Sample

Participants were recruited from the online study pool of the University of Konstanz. All members of the pool (N = 707 at the start of the study) were randomly assigned to one of four groups and only saw the corresponding study advertisement. Eligible participants had to be able to read and write German and own an Android smartphone to be able to use the study app. Based on previous studies for which participants were recruited via the study pool, recruiting 200 participants was deemed feasible. However, recruitment had to be halted in November 2021 before reaching that threshold due to limited availability of funding.

Study design

The study used a between-subjects design and time-based sampling. Data was collected between February and November 2021. The study adhered to the Declaration of Helsinki; general ethical approval was obtained from the University of Konstanz ethics committee.

All members of the study pool were invited via email to take part in an EMA study about mood fluctuations. The study descriptions (see supplementary material) were identical for all conditions except the information about compensation. Participants were randomly assigned to one of the 2 *Payment* (fixed/performance-contingent) x 2 *Feedback* (yes/no) conditions. Participants in the fixed payment conditions received €15. Participants in the performance-contingent payment conditions received €1 for completing the online questionnaire and €0.20 per answered prompt (equaling to a maximum of €15 for perfect retention). All participants were able to convert money into course credit, with €5 amounting to 0.5 hours of course credit. Participants in the feedback condition additionally received printed feedback with visualizations depicting the fluctuations of their mood during the study period (see supplemental material for an example). Participants only received feedback after the study period to avoid influences on the mood data being collected (c.f. König, Allmeta, et al., 2022).

Procedure

Upon sign-up, participants filled in an online questionnaire. They provided informed consent by ticking a box on the first page. The questionnaire assessed basic demographic information and motivation for taking part in the study. At the end of the questionnaire, participants received instructions for how to install the study app (movisensXS, movisens GmbH, Karlsruhe) and a personalized QR code to download the study questionnaires on their phone. They were then asked to record their mood for 7 consecutive days by responding to 10 random prompts per day that were sent between 8:00 am and 9:00 pm, with at least 15 minutes between prompts. Each EMA prompt only contained the six items to assess mood, which took participants on average 20 seconds (median = 16 seconds) to complete. Afterwards, participants received instructions for how to uninstall the study app. To receive their payment and written feedback (if applicable), they met with a researcher at the university.

Recruitment was initially started in January 2021, but due to pandemic-related restrictions, recruitment was slow. We therefore halted recruitment in February and continued in May 2021. Since recruitment was still slow due to ongoing restrictions, we halted recruitment again during the semester break (July to October) and recruited again in November 2021.

Recruitment was then ended since the funding required for paying participants was no longer available.

Measures

Mood

Participants' mood was assessed using the Wilhelm and Schoebi (2007) scale, which assesses valence (content/discontent; unwell/well), calmness (agitated/calm; relaxed/tense) and energetic arousal (tired/awake; full of energy/without energy) with two items each. Responses were indicated on seven-point semantic differentials. Internal consistencies per group are reported as part of the results.

Questionnaire data

Trust in science was assessed with a three-item questionnaire from the German Wissenschaftsbarometer 2021 (Wissenschaft im Dialog/ Kantar, 2021). Responses were indicated on a 6-point scale from 1 - trust entirely to 5 - trust not at all (α = .42 in the entry questionnaire).

Motivations for signing up for the study were assessed with six items on a 5-point scale (1 - does not apply at all to 5 - fully applies). These items were evaluated separately, because they were meant to assess a specific reason each, namely: (1) personal interest in the topic; (2) to support research; (3) financial compensation; (4) to collect course credit; (5) participating in research studies is fun; (6) to learn more about myself.

Motivation to complete the study was assessed with one item (How motivated are you to complete the study?) on a 5-point scale (1 - not at all to 5 - very).

Affinity to technology (Franke et al., 2019) was assessed with the four-item short-form of the corresponding scale (Wessel et al., 2019). Responses were indicated on a 6-point scale from 1 - completely disagree to 6 - completely agree (α = .87 in the entry questionnaire).

Big Five. Basic personality traits according to the Big Five (Costa & McCrae, 1992) were assessed with the 10-item short form of the Big Five Inventory (Rammstedt & John, 2007; r_{ext} = .62; r_{agr} = .10; r_{con} = .46; r_{neu} = .49; r_{ope} = .57 in the entry questionnaire).

Demographic characteristics. Age ("Your age", followed by an open text field), gender ("You are ..." followed by the response options male, female, diverse), and monthly net household income (using the following six categories and "not specified": up to €500; €501 to €1,000; €1,001 to €1,500; €1,501 to €2,000; €2,001 to €2,500; €2,501 and more) were assessed.

Statistical analysis

Preprocessing of EMA data

Recruitment success was assessed by the proportion of participants enrolled per condition out of all invited members of the study pool at the start of the study. The speed of enrollment was determined by the rank of the day the participants completed the entry-questionnaire.

Retention was operationalized by a) the number of people completing the one-week EMA period (answering the 71st prompt) in relation to the number of invited and enrolled participants, and b) by the number of fully answered prompts. The drop-out speed was determined by the last prompt answered.

Data quality. Internal consistency of a six-item mood scale (Wilhelm & Schoebi, 2007) was computed per participant. That is, all entries of an individual participant across time were considered separate entries, and Cronbach's alpha was computed for each participant based on all EMA entries on the mood scale. In the context of this study, this measure is regarded as an indicator of high data quality of a participant and not used as a measure of scale reliability.

Statistical models

As pre-registered, we effect-coded the two between-factors and used logistic regressions for binary outcomes (recruitment success and drop-outs), cox regressions for analyses for speed of recruitment and drop-out, and ANOVAs for continuous outcomes. We decided against pre-registered MANOVAs for continuous outcomes, since it was not clear which outcomes should be summarized in them. Because the sample size was much lower than expected in the pre-registration, we decided not only to report significant results with $p < .05$, but also marginally significant results with

$p < .10$, for which we do not want to dismiss an effect in acknowledgement of the limited achieved power. Significant and marginally significant interactions were thus followed by simple-effect analyses.

The pre-registered strategy to determine drop-out speed was ambiguous as to whether the days of the last prompt should be used as a speed indicator and how participants not providing any EMA data should be handled. We opted for the last successful prompt analyses, also because with the day criterion, only seven people were considered dropping out early (1 after 4 days, and 3 on day 5 and 6). Regarding the handling of participants not providing EMA data, we opted to report test statistics excluding them, but also present data on statistics including them, where feasible.

Results

Sample description

A total of 88 participants started the entry questionnaire, 14 of which did not indicate an identifier or were double entries caused by linking difficulties and entry errors and were therefore discarded. The final sample for analyses regarding motivation and study participation thus consisted of 74 (86% female; M_{age} = 23.29, SD = 4.97 years; $n_{fixed\ and\ feedback}$ = 21, $n_{fixed\ and\ no\ feedback}$ = 14, $n_{performance\text{-}contingent\ and\ feedback}$ = 24, $n_{performance\text{-}contingent\ and\ no\ feedback}$ = 15) of the 707 eligible members of the study pool. Out of these individuals, 70 could successfully be linked to entries in the study app. The other four did not provide a matching identification code in the EMA questionnaire. From these 70 entries, one did not provide data in the app except the identification code, leaving 69 participants in the EMA dataset to be analyzed. On average, participants responded to 40.7 out of the 70 mood prompts, yielding a compliance rate of 58%.

Does the incentivization influence study participation and recruitment speed?

Overall, participants that were offered *feedback* were marginally more likely to participate (B = 0.48, OR = 1.62, 95% CI(0.99; 2.65), p = 0.056, 45/355 vs. 29/352) in the study. Similarly, these participants signed up for the study at a marginally higher rate across time by completing the entry questionnaire (B = –0.45, HR = 0.64 95% CI(0.40; 1.01), p = 0.057; see also Figure S1 in the supplementary material); this can be seen as initial evidence for H7. Both *payment* and the payment x feedback interaction played no discernible role in participation decisions (all $|B| \leq 0.11$; all $p \geq 0.816$); H1 was thus rejected.

Does the incentivization influence retention?

Considering completed EMA questionnaires, there was only a numerical, but no significant advantage for incentivization conditions including feedback (17/355 vs. 9/352, p = 0.103). When only considering all individuals that finished the entry questionnaire (N = 74), the recruitment strategies generally had no discernible effect on the number of successful EMA entries (all $F(1, 70) < 1.21$, $p \geq 0.275$, $\eta^2_p \leq 0.017$), or whether the final prompt was answered (all $p \geq 0.212$). Both H2 and H5 thus had to be rejected. However, when only considering the 69 individuals that actually provided EMA data, the rate of dropout determined by the final prompt answered was marginally affected by an unexpected *payment x feedback* interaction (B = 1.23, HR = 3.41 95%CI(1.00; 11.63), p = 0.050; see also Figure S2 in the supplementary material): In the fixed payment group, participants without feedback dropped out at a higher rate than with feedback (B = –0.92, HR = 0.40 95%CI(0.17; 0.95), p = 0.037), while no differences emerged for the performance-contingent payment groups based on feedback (B = 0.31, HR = 1.37 95%CI(0.57; 3.26), p = 0.484). Regarding the number of successfully answered prompts, participants with a performance-contingent payment scheme answered marginally more prompts (M = 46.19, SD = 14.35) compared to the fixed payment scheme (M = 40.85, SD = 14.52; $F(1, 65) = 3.00$, p = 0.088, η^2_p = 0.044). This can be seen as initial evidence for H3.

Does the incentivization influence EMA data quality?

While most of the 69 participants with EMA data missed the reverse-coding of

mood items in the EMA questionnaire and therefore had negative internal consistencies, both feedback (M = –1.79, SD = 1.27 vs M = –2.76, SD = 2.06; $F(1, 65) = 6.38$, $p = 0.014$, $\eta^2_p = 0.089$) and marginally also fixed payment (M = –1.80, SD = 1.36 vs M = –2.53, SD = 1.90, $F(1, 65) = 3.67$, $p = 0.060$, $\eta^2_p = 0.054$) led to higher values indicating potentially higher – but still completely unreliable – data quality. H6 was thus confirmed. There is also some initial evidence for H4, although the hypothesis had to be formally rejected since the analysis fell short of reaching the predetermined threshold for statistical significance.

Does the incentivization influence participation motivation and sample characteristics?

Results are summarized in Table 1 and Figure 1. Overall, participants indicated higher motivation to complete the study in the fixed compared to the performance-contingent payment condition. Furthermore, there was a marginally significant effect of payment with fixed payment yielding participants with somewhat higher epistemic interest, while feedback recruited people with both more epistemic interest and personal interest in the topic of the study; H8 was thus confirmed. The interest to support science was marginally lower in the performance-contingent payment groups when no additional feedback was provided. The marginal *payment x feedback* effect in trust in science additionally indicates that people with the no feedback, performance-contingent payment scheme were specifically more trusting in science than the other conditions.

Concerning general demographics, there was no difference by incentivization scheme, except that participants in the fixed payment scheme were more neurotic. Trait openness was specifically low for fixed payment incentivization with feedback.

Discussion

This study experimentally tested the impact of monetary and non-monetary incentivization on recruitment, retention, data quality, and participant characteristics in an EMA study. Monetary and non-monetary incentives differentially affected the outcomes of interest, indicating that careful considerations need to be made regarding incentivization to reach specific goals, e.g., regarding recruitment speed or diversity of the sample.

Implications for recruitment

It was assumed that providing fixed payment would yield more participants recruited (H1), but this hypothesis could not be confirmed. Results provide some evidence for the assumption that offering performance-contingent payment would lead to more prompts being answered (c.f., H3). This confirms the notion derived from comparing studies with different incentivization schemes in Ottenstein and Werner (2021) that performance-contingent payment might be advantageous over fixed payment, at least when it comes to the amount of data that can be collected per participant. At the same time, results provide a first indication that offering feedback may facilitate recruitment both regarding the total number of participants (c.f., H7) and the recruitment speed, with small to medium effect sizes. However, the effect fell short of reaching the threshold for statistical significance specified in the preregistration; replications are required to confirm the findings.

Implications for retention

Furthermore, it was expected that fixed payment would lead to more drop-out (H2) and that feedback provision would decrease dropout (H5). Indeed, fixed payment even led to higher drop-out rates when no additional incentive for answering prompts was provided by offering feedback. This points towards a potential buffering role of feedback, which may sustain interest in the study and willingness to record — a pattern similarly found in studies showing higher engagement with self-monitoring in interventions with feedback (Burke et al., 2011). Thus, if a provision of performance-contingent payment is not possible, e.g., because of a lacking ground truth as in event-contingent dietary EMA studies (König, Van Emmenis, et al., 2022; Ziesemer et al., 2020), offering feedback may be crucial in retaining participants. Yet, this suggestion is somewhat speculative since the present study did not include a "feedback only" condition, which should be explored in future research.

Implications for the sample composition

The implications for sample composition are especially important since potential benefits of providing feedback for participant recruitment may come at the cost of increasing selective sampling of participants as indicated by higher intrinsic motivation and interest of participants signing up for the study in feedback conditions (c.f., H8). Similar issues also occur when offering a fixed payment scheme irrespective of answered prompts. This selectivity may be particularly problematic in studies which aim to generalize effects to a population. Therefore, feedback and fixed payment schemes may impair efforts to recruit populations that are more difficult to reach. Since motivations to take part in the study were only assessed after and not prior to assignment to a condition, the present study cannot rule out that the condition might also have influenced motivation. Future studies should therefore tap into the underlying psychological mechanisms of study incentivization to shed further light on this issue.

Furthermore, one should also take into account that these two incentivization strategies may increase self-selection biases hampering study conclusions up to the point of misjudging e.g. predictor-outcome relationships or the effect of an intervention (Biele et al., 2019; Humphreys et al., 2014; see also Bethlehem, 2010). Thus, active self-selection needs to be carefully considered as a potentially negative consequence of offering feedback as a low-cost alternative incentivization. On the other hand, the potential prerequisite of higher trust when solely receiving performance-contingent payment (as a sign of trust in the scientist to fairly evaluate the performance) may also decrease generalizability when only relying on this incentivization method, especially because trust in science is an important predictor of belief in misinformation (e.g., Agley & Xiao, 2021) and even adherence to health-protective measures (Devine et al., 2021).

The present study did not find meaningful differences between groups regarding sociodemographic characteristics. This is potentially due to the homogeneous study pool used for recruitment, which mainly consisted of students who take part in studies in exchange for course credit. This also raises potential ethical concerns regarding incentivization of study participation; ethics committees are usually concerned with potential undue influence due to excessive payment, which may undermine consent and encourage participants to take risks that they would not accept without payment (see e.g., Gelinas et al., 2018 , for a discussion). In the context of the present study, we adhered to regulations of payment set by the study pool; in other contexts, researchers may need to carefully weigh payment thresholds and their potential influence on participants (Dickert et al., 2002). At the same time, some form of payment may only seem fair, given that participants volunteer their time which they could otherwise spend working a paid job (Bierer et al., 2021). This is especially true for participants with low socio-economic status which are often underrepresented in research (Krukowski et al., 2024). The research community thus should continue the debate on payment practices, taking not only feasibility and research outputs, but also ethical considerations into account.

Implications for data quality

As a potential positive consequence, the sampling of more highly motivated participants in feedback and fixed payment schemes may also increase data quality. A first, speculative indication of this may be seen in the effects the experimental condition had on internal consistency of individual responses for both fixed payment (c.f., H4) and feedback provision (c.f., H6). Thus, deciding against offering feedback or fixed payment to diversify the sample may impose potential costs in data quality. Because the internal consistencies were negative and responses of a single person violate the independence assumption, these results should be confirmed by other means such as the compliance to potential attention checks.

Limitations

This study experimentally tested claims previously based on between-study comparisons. Participants were recruited from one existing panel to avoid biases in recruitment by varying sources of study invitation or contents. As a trade-off for this high internal validity, this study pool was relatively homogenous. Aside from

motivational differences, we found that participants in the fixed payment conditions were more neurotic, which may reflect their desire for certainty (Hirsh & Inzlicht, 2008). Although student samples are common in EMA research (Ottenstein & Werner, 2021; Perski et al., 2023), a more diverse study population is necessary to rule out demographic effects. This is especially needed since a previous systematic review was unable to compare effects due to missing demographic information beyond age and gender (Wrzus & Neubauer, 2022). Furthermore, while it was technically assured by the panel that each participant could only participate in their assigned condition, we cannot rule out that study participants talked to each other before or while taking part in the study and noticed that the study was advertised in different ways to different people. However, since the study was mostly conducted while teaching was delivered online, it could be assumed that study participants were less likely to notice that their classmates were participating in the study under alternative conditions.

Compliance was relatively low for an EMA study. That may have been due to the high number of assessments, which participants may have perceived as burdensome (c.f., Wrzus & Neubauer, 2022). Furthermore, the overall quality of EMA responses was unsatisfactory. Apparently, most participants did not differentiate between normally coded and reverse-coded items in the EMA questionnaire resulting in negative internal consistencies. This may indicate that balancing scales or the assumption of unidimensionality may be problematic in highly intensive EMA studies (c.f., Eisele et al., 2021). Internal consistencies of unidimensional scales used in EMA studies are rarely reported, but studies reporting high reliability either use far less prompts per day (Wilhelm & Schoebi, 2007) or fewer items per event (Courvoisier et al., 2010; Mey et al., 2020). Furthermore, the quality of the same items in the entry questionnaire was satisfactory (α = .81). While the present study does not interpret the scale itself, a lack of internal consistency is more problematic for studies aiming to draw conclusions about mood and thus warrants further investigation.

The present study focused on mood assessments, which are commonly studied using EMA. It is important to note, however, that generalizability to other constructs studied with EMA may be limited. For instance, mood may not be seen as a topic highly relevant to one's self-concept, thus feedback may be of relatively little importance and not strongly encourage (or hinder) participation. Other constructs and be-

Original Purpose

Recent reviews stated that the method chosen to incentivize participation in Ecological Momentary Assessment (EMA) studies may impact compliance. Paying participants per completed prompt, for instance, was associated with higher compliance compared to paying participants a flat fee. However, these conclusions were drawn based on comparing studies with different incentivization schemes. Confounding factors related to study design, duration, or sample characteristics thus could not be appropriately controlled for. Furthermore, due to the limited number of studies, other relevant outcomes such as data quality or participants characteristics could not be studied.

To draw meaningful conclusions regarding the potentially far-reaching impact of choice of incentivization in EMA studies, experimental designs are urgently needed where participants are randomly assigned to one of several incentivization schemes. By using a university study pool which allowed for this random allocation, we thus aimed to experimentally compare the impact of providing fixed vs performance-contingent monetary incentives as well as feedback provision on compliance, data quality, and participant characteristics in an EMA study. We expected fixed payments to yield more participants initially recruited, due to higher attractiveness of the reward, but also expected more drop-outs, since the reward could be collected anyway. Moreover, we expected fixed payment to yield better data quality as indicated by the internal consistency of the scales, compared to performance-contingent payment due to its focus on the mere number of completed prompts. Furthermore, we expected feedback provision to decrease drop-out and to increase the number of people interested in the study as well as data quality, due to personal relevance of the collected data. At the same time, we expected feedback to yield a sample with stronger interest in the study topic.

haviors, such as diet or physical activity, which are often linked to normative recommendations and evaluations, might induce stronger (positive or negative) reactions, depending on the valence of the feedback and the individual's expectations about the feedback (Renner, 2004). Future research should therefore replicate the study in a wide range of constructs and behaviors studied with EMA to shed light on this issue.

Recruitment was impaired due to COVID-19-dependent closures of the university during data collection. Due to the university being partially closed and teaching being conducted mostly online, many students were not on campus or even in town, which complicated the in-person collection of the compensation and potentially made study participation less attractive. Future studies should consider paying participants via bank transfer, although this requires more sophisticated procedures to be set up to be compliant with data protection regulations. Recruitment might have further been impaired by the requirement of owning an Android device. This was necessary since the study app was only available for Android phones and no smartphones were available to be loaned. If possible, researchers thus should make sure that their data collection tool is compatible with all common smartphone operating systems to facilitate recruitment. The recruitment had to be halted before the targeted sample size of 200 could be reached because the funding was only available until the end of 2021. A-posteriori analyses revealed that a sufficient (80%) power was achieved to detect medium effects for both main effects and 2x2 interactions (Cohen's f = 0.33, Pearson's r = 0.31; Faul et al., 2009). As a consequence of the smaller sample size and lower power, we also opted to report and interpret marginally significant findings. These marginal effects specifically warrant replication and should be considered tentative. In addition, data was collected for 7 days, which is the median duration of EMA studies (Wrzus & Neubauer, 2022). However, one might argue that compliance may play an even bigger role in studies requiring longer and more intensive assessment periods. Replications are thus warranted to confirm and further differentiate the findings in larger samples and longer assessment periods.

Conclusion

The current research offers preliminary evidence that offering feedback as compensation in EMA studies may be an effective way to facilitate recruitment and increase data quality without affecting retention. The cost of feedback might be a higher selection bias. If this poses a problem for a study design, a performance-contingent payment scheme may be advisable to optimize the total amount of data obtained from a more diverse sample, but its relative advantage may be negligible.

References

Agley, J., & Xiao, Y. (2021). Misinformation about COVID-19: Evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, *21*, Article 89. https://doi.org/10.1186/s12889-020-10103-x

Berli, C., Inauen, J., Stadler, G., Scholz, U., & Shrout, P. E. (2021). Understanding between-person interventions with time-intensive longitudinal outcome data: Longitudinal mediation analyses. *Annals of Behavioral Medicine*, *55*(5), 476–488. http://doi.org/10.1093/abm/kaaa066

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161-188. https:/ / doi.org/ 10.1111/ j.1751-5823.2010.00112.x

Biele, G., Gustavson, K., Czajkowski, N. O., Nilsen, R. M., Reichborn-Kjennerud, T., Magnus, P. M., Stoltenberg, C., & Aase, H. (2019). Bias from self selection and loss to follow-up in prospective cohort studies. *European Journal of Epidemiology*, *34*(10), 927-938. http://doi.org/10.1007/s10654-019-00550-1

Bierer, B. E., White, S. A., Gelinas, L., & Strauss, D. H. (2021). Fair payment and just benefits to enhance diversity in clinical research. *Journal of Clinical and Translational Science*, *5*(1), Article e159. http://doi.org/10.1017/cts.2021.816

Bolger, N., & Laurenceau, J.-P. (2013). In *Intensive longitudinal methods: An introduction to diary and experience sampling research*. The Guilford Press.

Burke, L. E., Conroy, M. B., Sereika, S. M., Elci, O. U., Styn, M. A., Acharya, S. D., Sevick, M. A., Ewing, L. J., & Glanz, K. (2011). The effect of electronic self-monitoring on weight loss and

dietary intake: A randomized behavioral weight loss trial. *Obesity*, *19*(2), 338-344. http://doi.org/10.1038/oby.2010.208

Cheung, K. L., Peter, M., Smit, C., de Vries, H., & Pieterse, M. E. (2017). The impact of non-response bias due to sampling in public health studies: A comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health. *BMC Public Health*, *17*(1), Article 276. http://doi.org/10.1186/s12889-017-4189-8

Costa, P. T., & McCrae, R. R. (1992). Revised NEO personality inventory (NEO-PI-R) and NEO five-factor (NEO-FFI) inventory professional manual. *Odessa, Fl: PAR*.

Courvoisier, D. S., Eid, M., Lischetzke, T., & Schreiber, W. H. (2010). Psychometric properties of a computerized mobile phone method for assessing mood in daily life. *Emotion*, *10*(1), 115-124. http://doi.org/10.1037/a0017813

Devine, D., Gaskell, J., Jennings, W., & Stoker, G. (2021). Trust and the coronavirus pandemic: What are the consequences of and for trust? An early review of the literature. *Political Studies Review*, *19*(2), 274-285. https://doi.org/10.1177/1478929920948684

Dickert, N., Emanuel, E., & Grady, C. (2002). Paying research subjects: an analysis of current policies. *Annals of Internal Medicine*, *136*(5), 368-373. http://doi.org/10.7326/0003-4819-136-5-200203050-00009

Edwards, P. J., Roberts, I., Clarke, M. J., DiGuiseppi, C., Wentz, R., Kwan, I., Cooper, R., Felix, L. M., & Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, *2009*(3). http://doi.org/10.1002/14651858.MR000008.pub4

Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2021). Affective structure, measurement invariance, and reliability across different experience sampling protocols. *Journal of Research in Personality*, *92*, Article 104094. https://doi.org/10.1016/j.jrp.2021.104094

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136-151. http://doi.org/10.1177/1073191120957102

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456-467. https://doi.org/10.1080/10447318.2018.1456150

Gelinas, L., Largent, E. A., Cohen, I. G., Kornetsky, S., Bierer, B., & Fernandez Lynch, H. (2018). A framework for ethical payment to research participants. *The New England Journal of Medicine*, *378*(8), 766-771. http://doi.org/10.1056/NEJMsb1710591

Gillies, K., Kearney, A., Keenan, C., Treweek, S., Hudson, J., Brueton, V. C., Conway, T., Hunter, A., Murphy, L., & Carr, P. J. (2021). Strategies to improve retention in randomised trials. *Cochrane Database of Systematic Reviews*, *2021*(3). https://doi.org/10.1002/14651858.MR000032.pub3

Harari, G. M., Müller, S. R., Mishra, V., Wang, R., Campbell, A. T., Rentfrow, P. J., & Gosling, S. D. (2017). An evaluation of students' interest in and compliance with self-tracking methods: Recommendations for incentives based on three smartphone sensing studies. *Social Psychological and Personality Science*, *8*(5), 479-492. https://doi.org/10.1177/1948550617712033

Haynes, A., & Robinson, E. (2019). Who are we testing? Self-selection bias in laboratory-based eating behaviour studies. *Appetite*, *141*, Article 104330. http://doi.org/10.1016/j.appet.2019.104330

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29-29. https://doi.org/10.1038/466029a

Hirsh, J. B., & Inzlicht, M. (2008). The devil you know: Neuroticism predicts neural response to uncertainty. *Psychological Science*, *19*(10), 962-967. http://doi.org/10.1111/j.1467-9280.2008.02183.x

Humphreys, K., Blodgett, J. C., & Wagner, T. H. (2014). Estimating the efficacy of Alcoholics Anonymous without self-selection bias: An instrumental variables re-analysis of randomized clinical trials. *Alcoholism: Clinical*

and Experimental Research, *38*(11), 2688-2694. http://doi.org/10.1111/acer.12557

Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, *65*(3), 183-216. https://doi.org/10.1007/s11301-014-0111-y

König, L. M., Allmeta, A., Christlein, N., Van Emmenis, M., & Sutton, S. (2022). A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour. *Health Psychology Review*, *16*(4), 551-575. http://doi.org/10.1080/17437199.2022.2047096

König, L. M., Van Emmenis, M., Nurmi, J., Kassavou, K., & Sutton, S. (2022). Characteristics of smartphone-based dietary assessment: A systematic review. *Health Psychology Review*, *16*(4), 526-550. http://doi.org/10.1080/17437199.2021.2016066

Krukowski, R. A., Ross, K. M., Western, M. J., Cooper, R., Busse, H., Forbes, C., Kuntsche, E., Allmeta, A., Silva, A. M., & John-Akinola, Y. O. (2024). Digital health interventions for all? Examining inclusivity across all stages of the digital health intervention research process. *Trials*, *25*(1), Article 98. https://doi.org/10.1186/s13063-024-07937-w

Ludwigs, K., Lucas, R., Veenhoven, R., Richter, D., & Arends, L. (2020). Can happiness apps generate nationally representative datasets? A case study collecting data on people's happiness using the German socio-economic panel. *Applied Research in Quality of Life*, *15*, 1135-1149. https:/ / doi.org/ 10.1007/ s11482-019-09723-2

Mey, L. K., Chmitorz, A., Kurth, K., Wenzel, M., Kalisch, R., Tüscher, O., & Kubiak, T. (2020). Increases of negative affect following daily hassles are not moderated by neuroticism: An ecological momentary assessment study. *Stress and Health*, *36*(5), 615-628. http://doi.org/10.1002/smi.2964

Ottenstein, C., & Werner, L. (2021). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment, 29*(8), 1765-1776. http://doi.org/10.1177/10731911211032718

Perski, O., Keller, J., Kale, D., Asare, B. Y.-A., Schneider, V., Powell, D., Naughton, F., ten Hoor, G., Verboon, P., & Kwasnicka, D. (2023). Understanding health behaviours in context: A systematic review and meta-analysis of Ecological Momentary Assessment studies of five key health behaviours. *Health Psychology Review*, *16*(4), 576-601. http://doi.org/10.1080/17437199.2022.2112258

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., Mohebbi, M. H., Mooney, S., Suver, C., & Wilbanks, J. (2020). Indicators of retention in remote digital health studies: A cross-study evaluation of 100,000 participants. *NPJ Digital Medicine*, *3*, Article 21. https://doi.org/10.1038/s41746-020-0224-8

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203-212. https://doi.org/10.1016/j.jrp.2006.02.001

Renner, B. (2004). Biased reasoning: Adaptive responses to health risk feedback. *Personality and Social Psychology Bulletin*, *30*(3), 384-396. https://doi.org/10.1177/0146167203261296

Scholz, U. (2019). It's time to think about time in health psychology. *Applied Psychology: Health and Well-Being*, *11*(2), 173-186. http://doi.org/10.1111/aphw.12156

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

van Gelder, M. M. H. J., Vlenterie, R., IntHout, J., Engelen, L. J. L. P. G., Vrieling, A., & van de Belt, T. H. (2018). Most response-inducing strategies do not increase participation in observational studies: A systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *99*, 1-13. http://doi.org/10.1016/j.jclinepi.2018.02.019

Voslinsky, A., & Azar, O. H. (2021). Incentives in experimental economics. *Journal of Behavioral and Experimental Economics*, *93*, Article 101706. https://doi.org/10.1016/j.socec.2021.101706

Wessel, D., Attig, C., & Franke, T. (2019). ATI-S-An Ultra-Short Scale for Assessing Affinity for Technology Interaction in User Studies. In *Proceedings of Mensch und Computer 2019* (pp. 147-154). https://doi.org/10.1145/3340764.3340766

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale

to measure three basic dimensions of mood. *European Journal of Psychological Assessment*, *23*(4), 258-267. https://doi.org/10.1027/1015-5759.23.4.258

Wissenschaft im Dialog/ Kantar. (2021). *Wissenschaftsbarometer 2021*. https://www.wissenschaft-im-dialog.de/ projekte/ wissenschaftsbarometer/wissenschaftsbarometer-2021/

Wrzus, C., & Neubauer, A. B. (2022). Ecological Momentary Assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment, 30*(3), 825-846. http://doi.org/10.1177/10731911211067538

Ziesemer, K., König, L. M., Boushey, C. J., Villinger, K., Wahl, D., Butscher, S., Müller, J., Reiterer, H., Schupp, H. T., & Renner, B. (2020). Occurrence of and reasons for "missing events" in mobile dietary assessments: Results from three event-based EMA studies. *JMIR mHealth & uHealth*, *8*(10), Article e15430. https://doi.org/10.2196/15430

Figure 1 Means and standard deviations per condition for significant differences in participant motivation and sample characteristics in the incentivization scheme conditions.

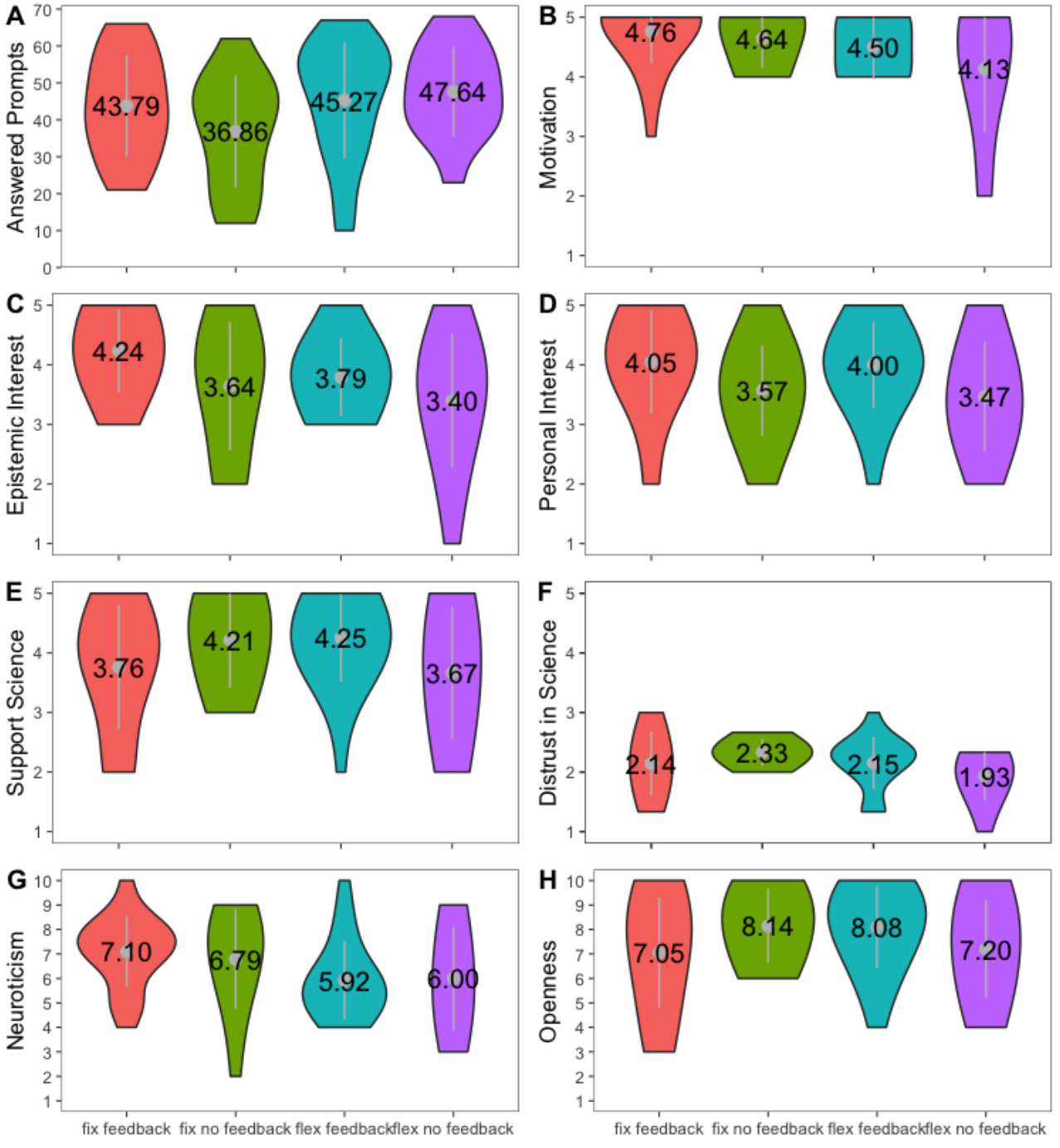


Table 1 ANOVA results for differences in participant motivation and sample characteristics in the incentivization scheme conditions.

Outcome	**Payment**			**Feedback**			**Payment x Feedback**		
	F	p	η^2_p	F	p	η^2_p	F	p	η^2_p
Study Motivation	5.96	.017	.078	2.36	.129	.033	0.61	.436	.009
Personal Interest	0.16	.695	.002	6.80	.011	.089	0.02	.883	.000
Support Science	0.02	.893	.000	0.09	.768	.001	5.50	.022	.073
Financial Reasons	0.00	.949	.000	0.05	.819	.001	2.50	.118	.034
Course Credit	0.04	.850	.001	1.39	.242	.019	2.41	.125	.033
Fun	1.47	.229	.021	0.89	.349	.013	2.47	.121	.034
Epistemic Interest	2.78	.100	.038	5.70	.020	.075	0.24	.624	.003
Trust in Science	3.53	.065	.048	0.02	.889	.000	3.90	.052	.053
Interest in Science	0.02	.893	.000	0.74	.394	.010	0.05	.822	.001
ATI	0.04	.838	.001	0.41	.524	.006	0.05	.823	.001
Extraversion	1.52	.221	.021	1.33	.253	.019	0.01	.911	.000
Agreeableness	2.56	.114	.035	1.06	.307	.015	0.18	.673	.003
Conscientiousness	0.83	.366	.012	0.74	.394	.010	0.93	.339	.013
Neuroticism	5.50	.022	.073	0.07	.788	.001	0.22	.641	.003
Openness	0.01	.919	.000	0.06	.816	.001	4.77	.032	.064
Age	0.02	.877	.000	.435	.512	.006	.345	.559	.005
Overall Mood	1.83	.181	.027	0.355	.554	.005	0.38	.540	.006
Mean Mood	0.00	.976	.000	0.589	.446	.009	0.14	.710	.002
SD Mood	1.42	.238	.021	0.835	.364	.013	1.02	.316	.015
	Wald	P	OR	Wald	P	OR	Wald	P	OR
Gender	.261	.609	1.43	.537	.464	1.67	.005	.905	0.91
Income	.339	.560	1.36	.611	.434	1.51	.056	.812	1.29

Three Persistent Myths about Open Science

Moin Syed[1]

Knowledge and implementation of open science principles and behaviors remains uneven across the sciences, despite over 10 years of intensive education and advocacy. One reason for the slow and uneven progress of the open science movement is a set of closely held myths about the implications of open science practices, bolstered by recurring objections and arguments that have long been addressed. This paper covers three of these major recurring myths: 1) that open science conflicts with prioritizing diversity, 2) that "open data" is a binary choice between fully open and accessible and completely closed off, and 3) that preregistration is only appropriate for certain types of research designs. Putting these myths to rest is necessary as we work towards improving our scientific practice.

Keywords *open science, diversity, data sharing, preregistration, meta-science*

I am hoping that we can at least all agree that we have some problems with how we conduct our science. *Serious problems*. In my home discipline of psychology, these problems include, but are not limited to, publication bias, low statistical power, p-hacking/questionable research practices/researcher degrees of freedom, HARKing, fraud, lack of diversity, weak measurement, weak theory, mistakes/sloppiness, jingle jangle fallacies, and problematic incentive structures (Chambers, 2017; Spellman, 2015; Syed, 2019).

It would be a mistake to believe that these problems only exist in psychology, or even more narrowly in social psychology. The problems are much larger and more widespread. Researchers across the natural sciences, social sciences, and humanities have all voiced concerns about the quality of research in their fields (Baker, 2016; Knöchelmann, 2019; Munafò et al., 2017). Similarly, efforts to improve research practices are evident in economics (Askarov et al., 2023), biomedicine (Errington et al., 2021), educational sciences (Fleming et al., 2021), ecology (Fraser et al., 2018), qualitative research (Humphreys et al., 2021), and many others.

Accordingly, none of us are immune from the problems that have been identified. For this reason, I use an all-inclusive "we" throughout this essay. Yes, that includes you. We know there are problems, and although we can debate how relevant and prevalent they are across different fields or sub-areas of individual fields, our time would be more productively used by admitting that there are problems and getting down to the work of improving our science, whatever that might mean for the area you work in.

Unfortunately, progress towards improving our science has been rather slow. This is, in part, because of the sheer scale of the problems and the fact that they infect all aspects of the scientific ecosystem (researchers, journals, institutions, funders, etc.). This makes change understandably hard. Another reason for the slow progress, however, is that not everyone is in agreement that change needs to happen. We don't all agree that there are actually problems in need of fixing. That we don't all agree is not what troubles me—that should be expected and even welcomed. No, what troubles me is the nature of some of the arguments.

The purpose of this essay is to take up a few recurring arguments to demonstrate that they

This article is based on a presentation given to the Purdue University Clinical Psychology Colloquium in November 2022. Slides and script of that talk, which serve as an earlier version of this paper, are available at https://osf.io/ahu84/is.

[1]Department of Psychology, University of Minnesota

Received
September 12, 2023
Accepted
January 25, 2024
Published
April 8, 2024
Issued
December 23, 2024

Correspondence
Department of Psychology, University of Minnesota
moin@umn.edu

are without merit. These arguments are 1) that diversity and open science conflict with one another, 2) that open data is a binary choice between being fully open or fully closed, and 3) that preregistration is only relevant for experimental designs. I refer to these arguments as "myths" because of their enduring nature and seeming resistance to counter-evidence. This is an opinion essay, which means that there could of course be some disagreements with my opinions, but the thrust of my arguments against the myths are based on the best available evidence or just plain facts. In making these arguments, I largely draw from psychology, but it would be a mistake to think that this essay is not relevant to you if you are from another field. Although many of the examples come from psychology, the terrain covered by the myths themselves are relevant to just about any discipline.

Brief background and rationale

Within psychology, most people will point to the year 2011 as the beginning of the "replication crisis." There were three notable events that year. First, was Bem's (2011) empirical evidence for the existence of extra-sensory perception published in the *Journal of Personality and Social Psychology*, an article that was met with widespread confusion and outrage. Second, was the False-Positive Psychology paper published in *Psychological Science* by Simmons et al. (2011), which unknowingly provided the "key" for how Bem was able to empirically demonstrate the impossible (i.e., researcher degrees of freedom). Third, was the discovery of rampant fraud by the respected social psychologist Diederik Stapel (Wicherts, 2011). Although problems in the field had been known long before 2011, they had also been mostly ignored by the mainstream of the field (see Syed, 2019, for several examples). As of 2011, that was no longer the case, and the three focal events begot a cascade of events that would eventually be known as the replication crisis (even though the three focal events were not about replication, per se, subsequent events were focused squarely on replication), as well as a groundswell of reform efforts that would come to be referred to as the open science movement.

The current year is 2024 which means it has been 13 years since our collective consciousnesses have been raised. It has been 13 years of learning more and more of the details and specifics of what is wrong with how we do our science, and it has been 13 years of proposed and implemented reforms. Thirteen years is a long time–what have we gained in 13 years?

On my optimistic days, I would say quite a lot. Terms like publication bias, p-hacking, questionable research practices, HARKing, and preregistration are no longer on the fringes of the field. An increasing number of scientific journals are supporting—and in some cases requiring—sharing of study data and materials (Kidwell et al., 2016; Nosek et al., 2018). Over 300 journals across a wide range of disciplines have adopted Registered Reports (Chambers & Tzavella, 2021) as a submission option, which to me is the single most effective intervention to fix our science that we currently have at our disposal. And, importantly, early career researchers are acutely aware of the problems and the need to address them (Farnham et al., 2017). We have made a lot of advances, and in many ways, the future of our science looks bright.

But the pessimism also lurks. As I think about the 300 journals *across the sciences* that have adopted Registered Reports, I also think about the denominator. How many journals are there? I find it amusing that nobody seems to know, but I have seen several references to 30,000 (e.g., Brembs, 2018). From personal experience, and the stories of others, there is a lot of resistance among journal editors towards adopting Registered Reports (see Chambers & Tzavella, 2021). It is not just editors' resistance to Registered Reports, but the general reluctance among scientists to change the way they go about their business. Beyond resistance, which could be for very good reasons, one thing that really gets me down is seeing the *same arguments* circulating again and again across these 13 years. Arguments that have been addressed and shown to be without merit. Recognizing those moments makes me think we have not made as much progress as I would like to think.

The current essay is about some of these recurring arguments against open science that seemingly *just will not die*: 1) that diversity and open science conflict with one another, 2) that open data is a binary choice between being

fully open or fully closed, and 3) that preregistration is only relevant for experimental designs. I have chosen to focus on these three arguments based on my own extensive experience being part of open science conversations, giving open science presentations, and engaging in advocacy around scientific reform, particularly with respect to journal operations. I am not claiming that these are the only recurring arguments or even that they are the most important. Someone could just as easily have selected different ones. They are, however, the ones I tend to come across most often and felt were worthy of taking head-on. For each of the three topics, I describe the basic background, followed by my argument for why the expressed concerns are misplaced, or in some cases, simply wrong.

Beyond these specific arguments, the purpose of this paper is to make a broader point about progress and criticism in the field. If you are going to have strong opinions about some topic, if you are to use your platform to argue against a particular initiative, then you best be informed about that topic. Too often, arguments are rooted in misunderstandings, faulty assumptions, and ignorance. This is unacceptable, and arguments made from this stance should no longer be treated as reasonable objections that must be addressed.

To be clear, I am *not* arguing that people should avoid criticizing open science practices. They should. And I am *not* arguing that reasonable criticisms should not be taken seriously. They should. I am arguing that we should differentiate between informed and uninformed criticisms. Additionally, too often concerns are framed as a way to end the conversation, rather than as a way to begin a conversation about how the concerns can be addressed.

Myth #1: Diversity and Open Science are in conflict

The lack of diversity in psychological science, both in terms of global diversity and within-country racial/ethnic diversity, has been a persistent problem in the field. Indeed, the problem has been long-recognized, but also long-ignored. Many folks in psychology did not seem to think too much about the issue until Henrich et al.'s (2010) paper on the overreliance on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) samples and then again with Roberts et al.'s (2020) article on racial inequality in publishing and editing. But the first major work to highlight the lack of racial/ethnic diversity in the field was Guthrie's (1976) book, *Even the Rat was White*, and since then there has been a consistent stream of papers raising the issue (e.g., Arnett, 2008; Draper et al., 2022; Graham, 1992; Green et al., 2022; Hall & Maramba, 2001; Hartmann et al., 2013; Lin & Li, 2022; Moriguchi, 2022; Nielsen et al., 2017; Ponterotto, 1988; Thalmayer et al., 2021). I think at this point we have a pretty good sense of the problem, and maybe—*maybe*—we are now serious about our efforts to actually make some changes (see Syed, 2023, for a discussion of the complexities of the issue).

The timing of the Henrich et al. (2010) WEIRD article is notable: 2010—awfully close to the 2011 "ground zero" year for the replication crisis and ensuing open science movement. Thus, interest and energy around diversifying the field was occurring alongside heightened awareness of the many problematic research practices widespread in the field.

These dual concerns, not surprisingly, largely existed in parallel with one another, and I don't think it is controversial to say that diversity was not a primary concern in the early days of the replication crisis (e.g., Beer et al., 2023; Lewis, 2017). Furthermore, because diversity was not a core component of the replication crisis, it did not figure heavily into proposed reforms of the time. The result was a familiar dynamic, where folks were advocating for field-wide shifts in how we go about our business without a strong consideration of the implications for diversity in the field. In other words, those with interests and concerns about diversity in open science were not part of the conversation.

Indeed, as has been discussed elsewhere in detail (Syed & Kathawalla, 2022), the open science movement can be understood as a *structural* movement, seeking to change the governing structure from an oligarchy to a democracy. However, it is not a *social-structural* movement, in that social power dynamics (e.g., racial and gender) were not part of the movement. Accordingly, the open science movement runs the real risk of reproducing, rather than disrupting, existing power imbalances in the field, despite its democratic focus (Grzanka & Cole, 2021).

Does the fact that diversity was not central to the open science movement mean that the two are inherently in conflict? No, and the distinction between diversity being included in the movement and diversity being in conflict with open science is one that is critical to maintain considering this important issue. Here, I elaborate on two primary reasons why the claims that diversity and open science are in conflict are without merit and, therefore, a myth.

First, the arguments that have been advanced have lacked evidence or compelling argumentation. This statement may be taken as dismissive or even disrespectful. It is not at all intended that way—I greatly appreciate that researchers who are concerned about diversity and representation are engaging with issues around open science. But at the end of the day, the central arguments made are without merit and do not square with the actual open science practices that they criticize (e.g., Bahlai et al., 2019; Fox Tree et al., 2022; Fuentes et al., 2022; Grzanka & Cole, 2021). Some of these concerns include worries about sharing sensitive or identifiable data, sharing data that were resource-intensive to collect, that exploratory studies are devalued, that qualitative research will be further marginalized, and that open science practices could lead to researchers' ideas being "scooped." All of these concerns have been addressed in the literature and either have clear solutions, have been demonstrated to be unfounded, or are based on incomplete understanding of the core issue. That is to say, it's not that the aforementioned concerns are not important—they very much are. Rather, the issue is that they are all concerns that have clear solutions and are, in a sense, resolved. What I find interesting about these stated concerns is that they are not at all unique to diversity-related research, and indeed are the same concerns expressed by folks generally skeptical of open science reforms (see Tackett et al., 2017, for a discussion in the context of clinical psychology). This is, in fact, why they have since been addressed.

One specific example that is frequently stated across the sciences is concern about article-processing charges (APCs) associated with open access publishing (e.g., Bahlai et al., 2019). Concern about APCs is framed as a diversity issue because researchers at under-represented institutions and/or who do work on under-represented populations and topics may have less access to the resources needed to cover the APC, which can indeed be ridiculously large. Such discussions, however, are often framed as though there is a mandate to pay APCs to make articles freely available. This so called "gold open access" is but one route to making research products openly available. An option that is possible for nearly all journals in psychology is to publish a "green open access" version of an article, which usually consists of an author-formatted version of the article that is made publicly available by posting it to an institutional or organizational repository (e.g., OSF, bioRxiv, PsyArXiv) or on the author's personal website (see Moshontz et al., 2021, for details). Moreover, there is an increasing number of "diamond open access" journals that do not charge any APCs at all (for example, journals published by PsychOpen or listed with the Free Journal Network). APCs and their role in scientific publishing are a major issue that needs be addressed, but there is no mandate within the open science movement for researchers to pay them to make their work open. That is just false and is an erroneous argument that is used to support the claim that diversity and open science are in conflict.

The second reason why the claim that diversity and open science are in conflict is a myth is that it overlooks the substantial work that has successfully integrated them. Indeed, we need to be careful of cautionary narratives about the relation between diversity and open science because they can inadvertently erase the efforts of the people who actually are working in this area. When Lee (2017) was the Editor of *Cultural Diversity and Ethnic Minority Psychology* he instituted a whole array of open science practices—and it was the first journal published by the American Psychological Association to do so! The Psychological Science Accelerator, which is a collaborative network of research teams from 84 countries around the world, has contributed some of the most substantial and rigorous work in the field, while elevating the contributions of scholars from under-represented regions (Moshontz et al., 2018). In a similar vein, the Framework for Open and Reproducible Research Training (FORRT; Azevedo et al., 2019) is a large global network of researchers and educators working together to produce educational resources

and original research, with diversity central to their mission and their work (Elsherif et al., 2022). Syed and Kathawalla (2022) outlined how various open science practices could contribute to cultural psychology, but also how cultural psychology could help further push open science. Many others have contributed empirical, conceptual, and structural analyses about open science and diversity (Humphreys et al., 2021; Ledgerwood et al., 2022; Lui et al., 2022). Taking these examples together—which is not an exhaustive list—it is important to recognize that there are those who do not view diversity and open science as in conflict, and indeed that there are researchers showing, with their actions, just how the two can work together quite well.

Diversity and open science are clearly not in conflict with one another. Remember the crucial distinction though, between the potential conflict and the fact that diversity has never been a central concern to the movement at large. The latter continues to be an issue, but concerns about lack of inclusion in decision-making, mistrust, and resource constraints are persistent structural problems in the field, and not new problems that have popped up in the context of open science. Because the open science movement never took issues of diversity as central to the mission, it is understandable that folks for whom that is a prime concern will be skeptical of the effort and perhaps something they should be wary of getting involved in. Adopting this perspective would be a huge mistake. Open science is a structural reform, and should take diversity seriously, but that can only happen if people with such expertise participate. If diversity-focused researchers opt out of open science, they run the risk of even greater marginalization within the field (see Causadias et al., 2021, for a similar argument). Moreover, the mainstream of the field would greatly benefit from their expertise.

To repeat myself from earlier, rather than being the end of the conversation, concerns that are raised should be the beginning of a conversation on how we can successfully move forward. In the next sections I take on the specific practices of open data and preregistration.

Myth #2: Open data is a binary: fully open or fully closed

The myth about diversity and open science is clearly complex, and I admittedly have extreme views on the matter that others will certainly not share. What I am trying to do is clearly lay out my arguments, and most critically, not base them on lack of information, misunderstanding, or unfounded fears. That is not to say that any of the preceding is absent from my arguments—I do not claim to be all knowing and always reserve doubt—but that is what I am *attempting* to do, at least.

The myth around open data is an entirely different matter. To me, this is a clear-cut issue, and one that commentators get wrong repeatedly. This is a recurring concern raised in the context of diversity and open science (e.g., Fox Tree et al., 2022; Grzanka & Cole, 2021), as well as clinical psychology (see Tackett et al., 2017) and researchers working with qualitative data (see Field et al., 2021), to name just a few. I just stated that I always reserve doubt—and that is true here, too—but not all reservations are of the same magnitude, and here it is quite tiny.

Open data refers to making available the data that are reported in a research product (e.g., published article, preprint, technical report). The central issue is that open data is viewed as a binary at the extremes: either the data are fully open and publicly available for anyone to access, or they are fully closed, never to be seen by anyone except for members of the research team (or, in reality, the person responsible for collecting the data and/or the data analyst). The first clue that this construction is false is that it is a binary, which is rarely an accurate representation of beliefs or practices. The same can be said about open science in general, which is often seen as an "all or nothing" enterprise, rather than a wide variety of policies and practices that can be implemented (see Bergmann, 2023; Kathawalla et al., 2021; Silverstein et al., 2024).

If your understanding of what "open data" means is the extreme end of openness, where the full data are publicly available for anyone to access if they so choose, then I certainly understand why you would have concerns. Concerns about sensitive details in the data, re-identification risk of participants, further plans for using the data, betraying what the partic-

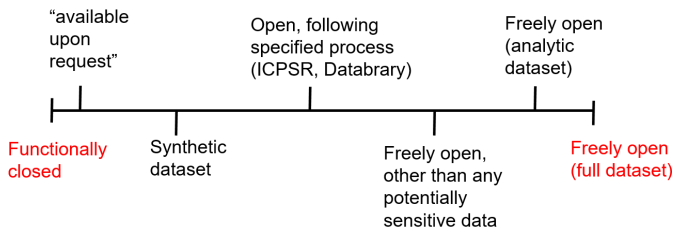


Figure 1 A continuum of data sharing

ipants consented to be a part of, and so on, are all serious and important concerns when we are talking about fully open and publicly accessible data.

Such concerns, however, are divorced from the reality of the practice. The reality of the practice is not a binary, but a continuum (*Figure 1*). There are many meanings of open data, with gradations of practice that fill the wide space between radically open and radically closed. Readers are directed to Meyer's (2018) excellent paper on the topic, as well as the discussion in Syed and Kathawalla (2022) with respect to work with marginalized populations. Here, I just briefly describe some key options along the continuum and why one might choose each option:

- **Freely open (full dataset)**
  This is the situation I described previously: the full dataset is publicly posted and available for download and reuse. I have no data to support this claim, but I believe strongly that this is what people have in mind when they argue against data sharing.
- **Freely open (analytic dataset)**
  An argument that I hear frequently from people is that they would be ok with making their data available, but they are not yet done with it. That is, they have plans for further analyses that could lead to publication, so they do not want someone else to publish their ideas. This is a totally reasonable concern and comes up frequently in the context of large datasets. That said, I suspect that we all vastly overestimate the degree to which we are "not yet done" with our datasets. Speaking for myself, I have had many plans for analyzing data that never came to fruition. More to the point, this concern seems to stem from the belief that the full dataset needs to be shared. Alternatively, authors could share only the data used for the analyses reported in the paper. It is extremely unlikely that a research team would have further plans to publish with the exact same data (at least, I hope so), but making the analytic data available would be useful for error detection or other researchers interested in fitting alternative models to the same data. The latter can even generate an additional publication, with the advent of the Verification Report format (Chambers, 2020). If there are no privacy or ethical concerns, then there is little reason for authors to *not* do this.
- **Freely open, other than sensitive data**
  Of course, oftentimes there *are* privacy and ethical concerns with making data available. In this case, an option that can be paired with making the full dataset or analytic dataset available is to not share any of the sensitive variables but make all other data available. For example, if researchers are concerned that including demographics could lead to participant re-identification—which is a real risk for minority participants—then those data could be held back, with a specified process outlined for how they could be obtained and under what conditions (see next).
- **Open following a specified process**
  Indeed, sometimes the data simply cannot be made openly available for a variety of reasons. We can now make an important distinction between freely and publicly open data, and data that are openly available only following a specified process. This process could involve an application, which might require researchers to specify how they will store the data and for what purpose they will use it. Researchers can specify, in advance, the conditions under which the data will be shared. This is a suitable option for when the data are potentially identifiable, include sensitive information, or where the consent form may not permit public sharing or certain types of analyses. This option also addresses a concern that arises in discussions of diversity and open science, that sharing data could lead to exploitation of marginalized research participants, as the terms of use of the data can be set by the research team to prevent such a thing. The

good news is that researchers do not need to develop this process themselves. The Inter-university Consortium for Political and Social Research (ICPSR) has long provided this service. There are also discipline-specific protected repositories, such as Databrary (Gilmore et al., 2016) for developmental research and LDbase (Hart et al., 2020) for educational and developmental research. This option will also typically be a strong fit for qualitative data which, more often than not, will contain identifying information. In addition to the previously named repositories, the Qualitative Data Repository was developed for precisely this purpose.

- **Synthetic dataset**
  Creating a specified process for data sharing is a great way to share data, but it is also time-consuming to set up and manage requests (although I would say a worthy investment of resources). Quintana (2020) developed an alternative option, developing the R package *synthpop*, which generates simulated data that reproduces various statistical properties of the original dataset while preserving confidentiality of the data. This simulated file is shared in lieu of the actual data.
- **Data available upon request**
  It is now well-documented that statements of "data available upon request," which are commonly found in published articles, effectively mean that the data are not available at all (Gabelica et al., 2022; Miyakawa, 2020; Wicherts et al., 2006). Accordingly, researchers should not use these statements, and journals should not allow them, as markers of data sharing.
- **Functionally closed**
  Unfortunately, this is the state of much of the research data in psychology, and there is little justification for it. I say "little" justification because there are of course some situations in which the data are so sensitive that they need to be heavily restricted (e.g., some genetic data) or an institutional ethics board has taken a hard line, but these situations are the exception rather than the rule. Sometimes data are functionally closed even *within* a research team. If you are part of a project, and the data are not available for your inspection, absent *compelling* rationale, you should be very concerned.

All of the above, with the exception of "data available upon request" and functionally closed, are perfectly acceptable forms of data sharing for different data situations. Readers are encouraged to embrace the maxim, "as open as possible, as closed as necessary," and seek to make their data FAIR (Findable, Accessible, Interoperable, and Reusable; Wilkinson et al., 2016). Data are too complex and varied to have a one-size-fits-all approach, and I tried to highlight some of the options along the continuum of fully open to fully closed. I urge you to be wary of criticisms about data sharing that do not acknowledge these complexities—or even better, I urge you to actively push back on them, because if they present a binary of extremes, they are simply presenting false information. This is not acceptable.

Myth #3: Preregistration is only for experimental designs

I often think about a parallel universe in which the replication crisis did not start in social psychology, but in a different area like developmental or clinical psychology. In many ways, it is unfortunate that social psychology was the origin, because it is one of the most methodologically narrow subfields in psychology. Now, some of you will take issue with that statement and ask for empirical evidence, but plenty of social psychologists have discussed the fact that simple lab experiments dominate the field (Baumeister et al., 2007; Cialdini, 2009; Rozin, 2001). Accordingly, the focus of the initial reforms was on the kinds of issues that come up in lab experiments. That meant that we thought about reforms starting with the simplest case, and then had to layer the complexity on top of it. That has certainly been the case with preregistration.

Preregistration involves specifying the research questions, hypotheses, methods, and analyses *before* conducting a study, via a time-stamped and unalterable repository (Nosek et al., 2018). There are quite a lot of claims about why we should be skeptical about preregistration (MacEachern & Van Zandt, 2019; McDermott, 2022; Pham & Oh, 2021; Szollosi et al., 2020). These include that it will stifle creativity, that it devalues exploratory work, that it does not allow you to make mistakes or change your mind, that it is unnecessary if you have strong

theory, that is it redundant with the ethics application, that it is too bureaucratic, and that it is not appropriate for _____ work (fill in the blank with qualitative, longitudinal, secondary data, etc.). These are all false claims based on limited understanding of what preregistration is and why we should do it[1], but my focus here is on the last one, that it is not appropriate for certain types of work.

One of the challenges of understanding preregistration—and the criticisms of it—are that there are different rationales for why researchers should do it. These include clearly distinguishing between what decisions were made prior to seeing the data ("confirmatory" analyses) from what decisions were made after seeing the data ("exploratory" analyses), and preventing the latter as being framed as the former in a research report (Wagenmakers et al., 2012); reducing the prevalence of undisclosed data-dependent decision-making (i.e., p-hacking, questionable research practices, researcher degrees of freedom; Srivastava, 2018); evaluating the severity of a test (Lakens, 2019); and serving as formal documentation of the study design and analysis plan (Haven et al., 2019).

Of course, these rationales are not mutually exclusive, and can all work together. Indeed, for me it is the final rationale—to serve as formal documentation of the design and analysis plan—that is the most functional way to think about preregistration and subsumes all of the other rationales. Thinking of it in this way has rather massive implications for the practice. If you think of preregistration as formal specification of the study design, then *it is clearly applicable to any type of research*. There is no form of research in which absolutely no plans or intentions exist prior to data collection or analysis. That is just not possible. Even a radically exploratory qualitative study involves the specification of a research question, a plan for recruiting participants into the study, some idea of what questions will be asked or how observations will be made, and ideally some plans for how the responses will be analyzed. All of this can be specified ahead of time, and to great benefit to the project.

Some might counter this argument by stating that preregistration is not necessary to achieve the aforementioned goals. Rather, researchers can maintain a rigorous and detailed "lab notebook" approach in which they include all details of a project prior to implementation (see Crüwell et al., 2019, for a discussion of this approach in relation to preregistration). I think that this is technically true, but also there is something quite different about posting the plans to a permanent, unalterable, and public repository. Doing so makes the plans feel somehow more real, and it is also a clear step in the project development phase, marking the transition to project execution. After working with many students on preregistrations over many years, I can assure you that they hold a very different meaning from if they were only kept internally.

Indeed, a major reason why I am so keen on preregistration is that I have witnessed the benefits time and again when working with students. I assure you that I am quite averse to anything that increases the bureaucratic nature of our work. What preregistration does, more than anything, is *make you think* about what you are doing. If you have to specify your plans in advance, then you have to think about *why* you are doing what you plan to do and *how* you plan to do it. These are questions that are relevant for *all* types of research. Since we started preregistering the work in my research group, I have had much more challenging and generative discussions with students about conceptualization, study design, and data analysis.

That is not to say that preregistration can be implemented with the same ease across study designs. I have experience with many different types of designs, and so have developed a continuum (once again) of difficulty with the practice (*Figure 2*). Difficulty here corresponds to the quantity of details and degrees of uncertainty involved in the preregistration process. At the far end of most difficult, we have longitudinal designs, meta-analyses, and most qualitative designs. At the far end of least difficult, we have simple experiments and ex-

[1] The one possible exception is the claim that preregistration is unnecessary if you have strong theory (Szollosi et al., 2020). This could be the case for some areas of psychology, but the vast majority of "theory" in psychology is quite weak, and certainly does not constrain decision-making. Rather, in most cases it makes it worse (LeBel & Peters, 2011). Moreover, if you follow my subsequent arguments, even if a study does have a strong formal theory that guides it, there are benefits to specifying the study plans in advance.

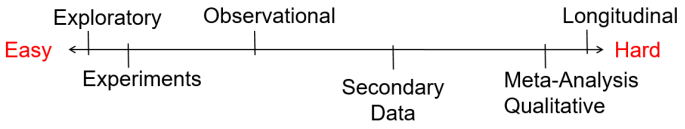


Figure 2 Another continuum, this time of preregistration difficulty

ploratory studies. Somewhere in the middle falls observational work and projects using secondary data. This continuum is not meant to be taken as a formal pronouncement of the relative difficulty of preregistering these specific designs that is applicable to all situations. Rather, it is based on my own experience using all of these designs, and more specifically my experience preregistering these designs. All study designs can be preregistered, but the details and difficulty will vary. The value, as far as I am concerned, is high throughout all situations.

Thus, any statements about, "preregistration is not appropriate for _____" are simply wrong. Preregistration can be used for any study design, although of course it will look different for different designs. Once again, we should actively reject any arguments—those that are both for and against open science practices—that take a one-size-fits-all approach. This may have been the case in the early days of advocacy for the practice, but our collective thinking has evolved on the issue, and the associated criticisms must be aware of that progress. There are some very smart and industrious people working on how to handle preregistration across the variability in study designs. For example, there are now templates and guides for how to preregister qualitative studies (Haven et al., 2019), secondary data (Weston et al., 2019), and meta-analyses (Moreau & Gamble, 2022), among many others. These folks have embraced the attitude I have repeated several times, that criticisms should not end a conversation, but serve as a launching point to continue the conversation and generate solutions.

Importantly, preregistration is *not* about getting everything right or perfect. We are fallible humans, and mostly mediocre scientists, so we *will* make mistakes. Deviations from a preregistration plan are perfectly acceptable so long as they are transparently reported (Willroth & Atherton, 2024). At that point, it is up to the reader to determine how the deviation impacts the credibility of the study. On that note, I want to stress that just because a study is preregistered, that does not mean that it is a quality study. I have seen plenty of bad preregistered studies, including those that have undisclosed deviations (see Claesen et al., 2021). It does, however, allow readers greater information about how the study was planned relative to how it was reported, and thus facilitates stronger interpretations of the stated claims.

Conclusion

Thirteen years since the start of the replication crisis in psychology, along with the many efforts to identify problems in other fields, tells me that we should all know and admit that we have problems with how we do our science. As I stated from the outset, I very intentionally use "we" because it is true for all of us. None of us are immune, within psychology for sure, but also across the sciences. To paraphrase Simine Vazire, a prominent leader in the scientific reform movement, if you don't think your field has a problem, that's probably because you haven't looked.[2]

At this point, given all that we know, it is irresponsible to be an active psychological scientist and to not be informed about the problems that have been unearthed via the replication crisis, and the solutions that have been proposed as part of the open science movement. The keyword is *informed*. Opinions are cheap and plentiful, and it is easy to have a negative reaction about a new practice that is quite different from what you have been accustomed to. Such reactions, however, are often based on uninformed, surface understandings of the issues, rather than careful study and consideration of the details. To reiterate, criticism is good and needed, and it is far from clear that certain open science practices are an unqualified good or benefit to the quality of research. But uninformed criticism and criticism that retreads previously resolved issues is not helpful. Here,

[2](Simine is not sure that she ever said this but agrees that it sounds like something she would have said; personal communication Dec 14, 2021)

I highlighted three recurring debates—about diversity and open science, data sharing, and preregistration—that I argue have been based on insufficient information. As active scientists in the field, it is our responsibility to reject and correct such false claims if we are to have any hope for progress.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614. https:/ / doi.org/ 10.1037/ 0003-066X.63.7.602 (see p. 34).

Askarov, Z., Doucouliagos, A., Doucouliagos, H., & Stanley, T. D. (2023). Selective and (mis)leading economics journals: Meta-research evidence. *Journal of Economic Surveys*, 1–26. https://doi.org/10.1111/joes.12598 (see p. 32).

Azevedo, F., Parsons, S., Micheli, L., Strand, J., Rinke, E., Guay, S., Elsherif, M., Quinn, K., Wagge, J. R., Steltenpohl, C., Kalandadze, T., Vasilev, M., de Oliveira, C. F., Aczel, B., Miranda, J., Galang, C. M., Baker, B. J., Pennington, C. R., Marques, T., & FORRT. (2019). Introducing a framework for open and reproducible research training (FORRT). *OSF Preprints*. https://doi.org/10.31219/osf.io/bnh7p (see p. 35).

Bahlai, C., Bartlett, L., Burgio, K., Fournier, A., Keiser, C., Poisot, T., & Whitney, K. (2019). Open science isn't always open to all scientists. *American Scientist*, *107*(2), 78. https://doi.org/10.1511/2019.107.2.78 (see p. 35).

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, Article 452-454. https://doi.org/10.1038/533452a (see p. 32).

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x (see p. 38).

Beer, J., Eastwick, P., & Goh, J. X. (2023). Hits and misses in the last decade of open science: Researchers from different subfields and career stages offer personal reflections and suggestions. *Social Psychological Bulletin*, *18*, 1–23. https://doi.org/10.32872/spb.9681 (see p. 34).

Bem, D. J. (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 (see p. 33).

Bergmann, C. (2023). The buffet approach to open science. *CogTales*. https:/ / cogtales.wordpress.com/ 2023/ 04/ 16/ the-buffet-approach-to-open-science/ (see p. 36).

Brembs, B. (2018). Prestigious science journals struggle to reach even average reliability. *Frontiers in Human Neuroscience*, *12*, 37. https:/ / doi.org/ 10.3389/ fnhum.2018.00037 (see p. 33).

Causadias, J. M., Korous, K. M., Cahill, K. M., & Rea-Sandin, G. (2021). The importance of research about research on culture: A call for meta-research on culture. *Cultural Diversity and Ethnic Minority Psychology*, *29*(1), 85–95. https:/ / doi.org/ 10.1037/ cdp0000516 (see p. 36).

Chambers, C. D. (2017). The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. *Princeton University Press*. (see p. 32).

Chambers, C. D. (2020). Verification reports: A new article type at Cortex. *Cortex*, *129*, 1–3. https://doi.org/10.1016/j.cortex.2020.04.020 (see p. 37).

Chambers, C. D., & Tzavella, L. (2021). The past, present and future of registered reports. *Nature Human Behaviour*, *6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7 (see p. 33).

Cialdini, R. B. (2009). We have to break up. *Perspectives on Psychological Science*, *4*(1), 5–6. https://doi.org/10.1111/j.1745-6924.2009.01091.x (see p. 38).

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https: / / doi.org/ 10.1098/ rsos.211037 (see p. 40).

Crüwell, S., Stefan, A. M., & Evans, N. J. (2019). Robust standards in cognitive science. *Computational Brain & Behavior*, *2*(3–4), 255–265. https://doi.org/10.1007/s42113-019-00049-8 (see p. 39).

Draper, C. E., Barnett, L. M., Cook, C. J., Cuartas, J. A., Howard, S. J., McCoy, D. C., Merkley, R., Molano, A., Maldonado-Carreño, C., Obradović, J., Scerif, G., Valentini, N. C., Venetsanou, F., & Yousafzai, A. K. (2022). Publishing child development research from around the world:

An unfair playing field resulting in most of the world's child population under-represented in research. *Infant and Child Development*, *32*(6), Article e2375. https://doi.org/10.1002/icd.2375 (see p. 34).

Elsherif, M. M., Middleton, S. L., Phan, J. M., Azevedo, F., Iley, B. J., Grose-Hodge, M., Tyler, S. L., Kapp, S. K., Gourdon-Kanhukamwe, A., Grafton-Clarke, D., Yeung, S. K., Shaw, J. J., Hartmann, H., & Dokovova, M. (2022). Bridging neurodiversity and open scholarship: How shared values can guide best practices for research integrity. https://doi.org/10.31222/osf.io/k7a9p (see p. 36).

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & A, N. B. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, Article e67995. https://doi.org/10.7554/eLife.67995 (see p. 32).

Farnham, A., Kurz, C., Öztürk, M. A., Solbiati, M., Myllyntaus, O., Meekes, J., Pham, T. M., Paz, C., Langiewicz, M., Andrews, S., Kanninen, L., Agbemabiese, C., Guler, A. T., Durieux, J., Jasim, S., Viessmann, O., Frattini, S., Yembergenova, D., Benito, C. M., & Hettne, K. (2017). Early career researchers want Open Science. *Genome Biology*, *18*(1), Article 221. https://doi.org/10.1186/s13059-017-1351-7 (see p. 33).

Field, S. M., van Ravenzwaaij, D., Pittelkow, M. M., Hoek, J. M., & Derksen, M. (2021). Qualitative open science–pain points and perspectives. *OSF Preprints*. https://osf.io/e3cq4/ (see p. 36).

Fleming, J. I., Wilson, S. E., Hart, S. A., Therrien, W. J., & Cook, B. G. (2021). Open accessibility in education research: Enhancing the credibility, equity, impact, and efficiency of research. *Educational Psychologist*, *56*(2), 110–121. https://doi.org/10.1080/00461520.2021.1897593 (see p. 32).

Fox Tree, J., Lleras, A., Thomas, A., & Watson, D. (2022). The inequitable burden of open science. https://featuredcontent.psychonomic.org/the-inequitable-burden-of-open-science/ (see pp. 35, 36).

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), Article e0200303. https://doi.org/10.1371/journal.pone.0200303 (see p. 32).

Fuentes, M. A., Zelaya, D. G., Delgado-Romero, E. A., & Butt, M. (2022). Open science: Friend, foe, or both to an antiracist psychology? *Psychological Review*, *130*(5), 1351–1359. https://doi.org/10.1037/rev0000386 (see p. 35).

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 (see p. 38).

Gilmore, R. O., Adolph, K. E., & Millman, D. S. (2016, August). Curating identifiable data for sharing: The Databrary project. In *2016 New York Scientific Data Summit (NYSDS)* (pp. 1–6). IEEE. https://doi.org/10.1109/NYSDS.2016.7747817 (see p. 38).

Graham, S. (1992). "most of the subjects were White and middle class": Trends in published research on African Americans in selected APA journals, 1970–1989. *American Psychologist*, *47*(5), 629–639. https://doi.org/10.1037/0003-066X.47.5.629 (see p. 34).

Green, K. H., Van De Groep, I. H., Te Brinke, L. W., van der Cruijsen, R., van Rossenberg, F., & El Marroun, H. (2022). A perspective on enhancing representative samples in developmental human neuroscience: Connecting science to society. *Frontiers in Integrative Neuroscience*, *16*, Article 981657. https://www.frontiersin.org/articles/10.3389/fnint.2022.981657 (see p. 34).

Grzanka, P. R., & Cole, E. R. (2021). An argument for bad psychology: Disciplinary disruption, public engagement, and social transformation. *American Psychologist*, *76*(8), 1334–1345. https://doi.org/10.1037/amp0000853 (see pp. 35, 36).

Guthrie, R. V. (1976). Even the rat was white: A historical view of psychology. Allyn & Bacon. (see p. 34). Hall, G. C. N., & Maramba, G. G. (2001). In search of cultural diversity: Recent literature in cross-cultural and ethnic minority psychology. *Cultural Diversity and Ethnic Minority Psychology*, *7*(1), 12–26. https://doi.org/10.1037/1099-9809.7.1.12 (see p. 34).

Hart, S. A., Schatschneider, C., Reynolds, T. R., Calvo, F. E., Brown, B. J., Arsenault, B., Hall, M. R. K., van Dijk, W., Edwards, A. A., Shero, J. A., Smart, R., & Phillips, J. S. (2020). LDbase. https://doi.org/10.33009/ldbase (see p. 38).

Hartmann, W. E., Kim, E. S., Kim, J. H. J., Nguyen, T. U., Wendt, D. C., Nagata, D. K., & Gone, J. P. (2013). In search of cultural di-

versity, revisited: Recent publication trends in cross-cultural and ethnic minority psychology. *Review of General Psychology*, *17*(3), 243–254. https://doi.org/10.1037/a0032260 (see p. 34).

Haven, T., Grootel, V., & L, D. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244. https:/ / doi.org/ 10.1080/08989621.2019.1580147 (see pp. 39, 40).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X (see p. 34).

Humphreys, L., Lewis, N. A., Sender, K., & Won, A. S. (2021). Integrating qualitative methods and open science: Five principles for more trustworthy research. *Journal of Communication*, *71*(5), 855–874. https://doi.org/10.1093/joc/jqab026 (see pp. 32, 36).

Kathawalla, U. -K., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, *7*(1), 18684. https://doi.org/10.1525/collabra.18684 (see p. 36).

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., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), Article 1002456. https: / / doi.org/ 10.1371/ journal.pbio.1002456 (see p. 33).

Knöchelmann, M. (2019). Open Science in the humanities, or: Open humanities? *Publications*, *7*(4), 65. https://doi.org/10.3390/publications7040065 (see p. 32).

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/jbh4w (see p. 39).

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172 (see p. 39).

Ledgerwood, A., Hudson, S. K. T. J., Lewis Jr, N. A., Maddox, K. B., Pickett, C. L., Remedios, J. D., Cheryan, S., Diekman, A. B., Dutra, N. B., Goh, J. X., Goodwin, S. A., Munakata, Y., Navarro, D. J., Onyeador, I. N., Srivastava, S., & Wilkins, C. L. (2022). The pandemic as a portal: Reimagining psychological science as truly open and inclusive. *Perspectives on Psychological Science*, *17*(4), 937–959. https://doi.org/10.1177/17456916211036654 (see p. 36).

Lee, R. M. (2017). Editorial. *Cultural Diversity and Ethnic Minority Psychology*, *23*(3), 311. https://doi.org/10.1037/cdp0000172 (see p. 35).

Lewis, N. (2017). Reflections on SIPS (guest post by Neil Lewis, Jr.) *The Hardest Science*. https:/ / thehardestscience.com/ 2017/ 08/ 11/ reflections-on-sips-guest-post-by-neil-lewis-jr/ (see p. 34).

Lin, Z., & Li, N. (2022). Global diversity of authors, editors, and journal ownership across subdisciplines of psychology: Current state and policy implications. *Perspectives on Psychological Science*, *18*(2), 358–377. https:/ / doi.org/ 10.1177/ 17456916221091831 (see p. 34).

Lui, P. P., Gobrial, S., Pham, S., Giadolor, W., Adams, N., & Rollock, D. (2022). Open science and multicultural research: Some data, considerations, and recommendations. *Cultural Diversity and Ethnic Minority Psychology*, *28*(4), 567–586. https://doi.org/10.1037/cdp0000541 (see p. 36).

MacEachern, S. N., & Van Zandt, T. (2019). Preregistration of modeling exercises may not be useful. *Computational Brain & Behavior*, *2*(3), 179–182. https:/ / doi.org/ 10.1007/ s42113-019-00038-x (see p. 38).

McDermott, R. (2022). Breaking free: How preregistration hurts scholars and science. *Politics and the Life Sciences*, *41*(1), 55–59. https://doi.org/10.1017/pls.2022.4 (see p. 38).

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. https://doi.org/10.1177/2515245917747656 (see p. 37).

Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, *13*(1), 1–6. https://doi.org/10.1186/s13041-020-0552-2 (see p. 38).

Moreau, D., & Gamble, B. (2022). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*, *27*(3), 426–432. https:/ / doi.org/ 10.1037/ met0000351 (see p. 40).

Moriguchi, Y. (2022). Beyond bias to Western participants, authors, and editors in developmental science. *Infant and Child Development*, *31*, Article e2256. https://doi.org/10.1002/icd.2256 (see p. 34).

Moshontz, H., Binion, G. E., Walton, H., Brown, B. T., & Syed, M. (2021). A guide to posting and managing preprints. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–11 (see p. 35).

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., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., & Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607 (see p. 35).

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), Article 0021. https://doi.org/10.1038/s41562-016-0021 (see p. 32).

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017 (see p. 34).

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114 (see pp. 33, 38).

Pham, M. T., & Oh, T. T. (2021). Preregistration is neither sufficient nor necessary for good science. *Journal of Consumer Psychology*, *31*(1), 163–176. https://doi.org/10.1002/jcpy.1209 (see p. 38).

Ponterotto, J. G. (1988). Racial/ethnic minority research in the Journal of Counseling Psychology: A content analysis and methodological critique. *Journal of Counseling Psychology*, *35*(4), 410–418. https://doi.org/10.1037/0022-0167.35.4.410 (see p. 34).

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, *9*, Article e53275. https://doi.org/10.7554/eLife.53275 (see p. 38).

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https:/ / doi.org/ 10.1177/1745691620927709 (see p. 34).

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review*, *5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1 (see p. 38).

Silverstein, P., Elman, C., Montoya, A., McGillivray, B., Pennington, C. R., Harrison, C. H., Steltenpohl, C. N., Röer, J. P., Corker, K. S., Charron, L. M., Elsherif, M., Malicki, M., Hayes-Harb, R., Grinschgl, S., Neal, T., Evans, T. R., Karhulahti, V. -M., Krenzer, W. L. D., Belaus, A., … Syed, M. (2024). A guide for social science journal editors on easing into open science. *Research Integrity and Peer Review*, *9*(1). https://doi.org/10.1186/s41073-023-00141-5 (see p. 36).

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 (see p. 33).

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. Perspectives on *Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918 (see p. 32).

Srivastava, S. (2018). Sound inference in complicated research: A multi-strategy approach. *PsyArXiv*. https:/ / doi.org/ 10.31234/osf.io/bwr48 (see p. 39).

Syed, M. (2023). The slow progress towards diversification in psychological research. *PsyArXiv*. https://psyarxiv.com/bqzs5/ (see p. 34).

Syed, M., & Kathawalla, U. K. (2022). Cultural psychology, diversity, and representation in open science. In K. McLean (Ed.), *Cultural methods in psychology: Describing and transforming cultures* (pp. 427–454). Oxford University Press. https:/ / psyarxiv.com/ t7hp2/ (see pp. 34, 36, 37).

Syed, M. (2019). The open science movement is for all of us. *PsyArXiv*. https://doi.org/10.31234/osf.io/cteyb (see pp. 32, 33).

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/ 10.1016/ j.tics.2019.11.009 (see pp. 38, 39).

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*(5), 742–756. https://doi.org/10.1177/1745691617690042 (see pp. 35, 36).

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*, *76*(1), 116-129. https:/ / doi.org/ 10.1037/ amp0000622 (see p. 34).

Wagenmakers, E. - J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078 (see p. 39).

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of pre-existing data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684 (see p. 40).

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*, 7. https://doi.org/10.1038/480007a (see p. 33).

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. https://doi.org/10.1037/0003-066X.61.7.726 (see p. 38).

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1), 1–14. https://doi.org/10.31234/osf.io/dwx69 (see p. 40).

Type I Error Rates are Not Usually Inflated

Mark Rubin [1]

The inflation of Type I error rates is thought to be one of the causes of the replication crisis. Questionable research practices such as *p*-hacking are thought to inflate Type I error rates above their nominal level, leading to unexpectedly high levels of false positives in the literature and, consequently, unexpectedly low replication rates. In this article, I offer an alternative view. I argue that questionable and other research practices do not usually inflate relevant Type I error rates. I begin by introducing the concept of Type I error rates and distinguishing between statistical errors and theoretical errors. I then illustrate my argument with respect to model misspecification, multiple testing, selective inference, forking paths, exploratory analyses, *p*-hacking, optional stopping, double dipping, and HARKing. In each case, I demonstrate that relevant Type I error rates are not usually inflated above their nominal level, and in the rare cases that they are, the inflation is easily identified and resolved. I conclude that the replication crisis may be explained, at least in part, by researchers' misinterpretation of statistical errors and their underestimation of theoretical errors.

Keywords *false positives, questionable research practices, replication crisis, significance testing, Type I error rate inflation*

During significance testing, a Type I error occurs when a researcher decides to reject a true null hypothesis (Neyman & Pearson, 1928, p. 177; Neyman & Pearson, 1933, p. 296). Type I errors are thought to play an important role in explaining the replication crisis in science. When the results of a study fail to replicate, the replication failure may be attributed to a Type I error in the original study. In other words, one of several reasons for a failed replication is that the null hypothesis is true and the original study's significant result was a false positive (Nosek et al., 2022, p. 726).

The replication "crisis" occurred because replication rates were lower than "expected or desired" (Nosek et al., 2022, p. 724; see also Munafò et al., 2017, p. 1; Open Science Collaboration, 2015, p. 7). Unexpectedly low replication rates have been attributed to larger than expected Type I error rates which, in turn, have been attributed to the use of questionable research practices. For example, Simmons et al. (2011) argued that "despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .050), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates" (p. 1359). Hence, questionable research practices are thought to inflate actual Type I error rates above the nominal conventional level of .050, leading to an unexpectedly high level of false positives in the literature and, consequently, unexpectedly low replication rates.

In this article, I offer the alternative view that relevant Type I error rates are not usually inflated by questionable research practices (e.g., *p*-hacking) or other research practices (e.g., model misspecification, exploratory analyses). I agree that Type I errors can be responsible for some replication failures. However, I argue that Type I error rate *inflation* is relatively rare, and that when it does occur it is easily identified and resolved. I conclude that theoretical errors provide a better explanation of the replication crisis than Type I error rate inflation.

My argument is consistent with philosophies of science and statistics that consider the logical relations between a hypothesis and its test result in the context of justification independent from their psychological origins, includ-

[1] Durham University

Received
December 5, 2023
Accepted
May 11, 2024
Published
November 16, 2024
Issued
December 23, 2024

Correspondence
Durham University
Mark-Rubin@outlook.com

ing researcher biases in the context of discovery (Popper, 1962, p. 140; Popper, 2002 p. 7; Reichenbach, 1938, p. 5). Based on this "logico-historical" view (Musgrave, 1974, p. 3), I argue that "actual" Type I error rates can be logically deduced from reported statistical inferences without referring to planned testing procedures that include the way in which hypotheses and evidence are constructed and selected. Consequently, my logical inference-based approach diverges from Mayo's (1996, 2018) error statistical approach, and it does not require preregistration to assess Type I error rates (Rubin, 2020a).

I begin with an introduction to Type I error rates that distinguishes them from theoretical errors. I then consider a range of questionable and other research practices that are thought to inflate Type I error rates. In particular, I consider model misspecification, multiple testing, selective inference, forking paths, exploratory analyses, *p*-hacking, optional stopping, double dipping, and HARKing. I demonstrate that relevant Type I error rates are not usually inflated by these practices. I conclude by summarizing my arguments, discussing the evidence for Type I error rate inflation, and considering some implications for our understanding of the replication crisis.

Before proceeding further, it is worth addressing a few potential misconceptions about my position. I am not arguing that questionable research practices do not exist. There is clear evidence that researchers sometimes engage in questionable research practices (e.g., John et al., 2012). My point is that these practices do not usually inflate Type I error rates. I am also not arguing that researchers are always honest or that they do not exploit researcher degrees of freedom. For example, I accept that researchers may falsely portray a post hoc hypothesis as an a priori hypothesis. However, I argue that this and other questionable research practices do not usually inflate Type I error rates. Furthermore, none of the points that I make condone dishonesty and the nondisclosure of potentially relevant information. Again, they only support the argument that questionable and other research practices do not usually inflate relevant Type I error rates. Finally, I am not denying that some of these research practices can be problematic. However, my argument is that they may increase *theoretical errors,* rather than *statistical errors*. In other words, they may lead to theoretical misinterpretation rather than Type I error rate inflation.

Introduction to Type I Error Rates

What Is a Type I Error Rate?

A Type I error rate is the frequency with which a researcher would decide to reject a true null hypothesis when they base their decision on the results of a significance test that is performed on a long run of random samples that are drawn from a null population in an imaginary situation in which random sampling error is the only source of error. There are four points to note about this definition.

First, the word "population" refers to not only a population of research participants, but also a population of study-specific research methods and conditions. This population of participants, methods, and conditions is randomly sampled each time the significance test is conducted (Fisher, 1922, p. 313).

Second, in scientific contexts, the population is not fully known. By definition, a "scientist" does not fully understand the relevant and irrelevant aspects of the populations that they are studying. Indeed, it is for this reason that they are studying those populations! Consequently, scientists face a reference class problem when attempting to specify the population to which their Type I error rate applies (Venn, 1876). They handle this problem by making theoretically informed guesses about the relevant and irrelevant aspects of the populations that are the subject of their statistical inferences. However, these guesses can be wrong! As Fisher (1956) explained, during significance testing "the population in question is hypothetical, … it could be defined in many ways, and … the first to come to mind may be quite misleading" (p. 78). Hence, scientists must continually ask themselves: "Of what population is this a random sample?" (Fisher, 1922, p. 313).

Third, Type I error rates are based on an imaginary situation in which random sampling error is the *only* source of error that can affect a researcher's decision to reject the null hypothesis. Of course, in the real world, many other sources of error can influence a researcher's decision (e.g., errors in data collec-

tion and entry, errors in research methodology, and/or errors in theoretical interpretation). However, Type I error rates do not refer to any of these other sources of error. They only refer to errors based on random sampling error (Berk et al., 1995, p. 423; Fisher, 1956, p. 44; Neyman & Pearson, 1928, pp. 177, 232). Hence, we must imagine that *if* a null hypothesis was true (i.e., if all samples were drawn from the null population), and *if* random sampling error was the *only* source of decision-making error, *then* the Type I error rate would indicate the frequency with which a researcher would reject the null hypothesis in a long run of repeated random sampling.

Finally, the idea of a frequency of decision-making errors during a long run of repeated random sampling from a null population is consistent with the Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1928, 1933). However, a Type I error can also be conceptualized within the alternative Fisherian approach (Fisher, 1956, 1971). In this case, a Type I error *probability* (not rate) represents the *epistemic* probability (not aleatory probability) of making a Type I error in relation to a *single* decision (not a long run of decisions) to reject a null hypothesis based on an observed test statistic from the *current sample* (not a long run of samples; Rubin, 2020b, 2021b).

What Is Type I Error Rate Inflation?

Type I error rate inflation occurs when the *actual* Type I error rate is higher than the *nominal* Type I error rate. The nominal Type I error rate is the rate that is set by the researcher, and it is used to determine whether an observed *p* value is "significant" or "nonsignificant." Hence, the nominal Type I error rate is also referred to as a *significance threshold* or *alpha level*. It is used to control the frequency of making a Type I error during a long run of repeated sampling.

The actual Type I error rate can be higher than the nominal Type I error rate in the context of multiple testing. For example, imagine that a researcher aims to make a decision about a null hypothesis based on three tests of that hypothesis, each with an alpha level of .050. Further imagine that the researcher is prepared to accept a significant result on *any* of the three tests as sufficient grounds for rejecting the null hypothesis. In this case, the researcher's actual (familywise) Type I error rate for their decision will be .143. Consequently, if they set the alpha level for their decision at .050, then their actual Type I error rate (.143) will be inflated above their nominal Type I error rate (.050). As I discuss later, this multiple testing problem underlies several of the research practices that are thought to inflate Type I error rates (selective inference, forking paths, exploratory analyses, *p*-hacking, optional stopping).

Importantly, the word "actual" in the phrase "actual Type I error rate" does not imply that we are able to identify "real" false positive results in any given study. The probability of a "real" false positive result would refer to the conditional posterior probability that a null hypothesis is true given its rejection (i.e., Pr[H_0 is true | reject H_0]). However, it is not possible to quantify this probability in scientific contexts (Meehl, 1997, p. 397; Neyman & Pearson, 1928, p. 176; Pollard & Richardson, 1987, p. 162). Instead, we must consider the hypothetical probability of rejecting a null hypothesis when it is true (i.e., Pr[reject H_0 ; H_0 is true]).[1] It is this hypothetical probability, rather than the conditional posterior probability, that represents the actual Type I error rate.

People sometimes confuse the actual Type I error rate with the conditional posterior probability (Mayo & Morey, 2017; Pollard & Richardson, 1987). For example, they might argue that, if you reject 200 null hypotheses using an alpha level of .050, and only 100 of those hypotheses are true, then you will end up with five actual Type I errors (i.e., 100 × .050). However, this scenario refers to the probability that a null hypothesis is true when it is rejected, Pr(H_0 is true | reject H_0), rather than the probability of rejecting a null hypothesis when it is true, Pr(reject H_0; H_0 is true). Confusing these two types of probability can be described as a Bayesian inversion fallacy (Gigerenzer, 2018; Greenland et al., 2016). During significance testing, the actual Type I error rate does not refer to the probability or prevalence of true null hypotheses; it simply assumes that each null hypothesis is true and then represents the frequency with which each hypothesis would be rejected given random sampling error per se (Fisher, 1971, p. 17). Hence, to return to the previous example, a person who rejects 200 null hypotheses using an alpha level of .050 should

expect to make 10 *actual* Type I errors (i.e., 200 × .050), not 5 (100 × .050), because they should assume (imagine) that all 200 hypotheses are true.[2]

Finally, it should be noted that nominal Type I error rates tend to be based on a research field's conventional alpha level (e.g., $p \leq .050$). Nonetheless, individual researchers can choose alpha levels that are more or less stringent than the conventional level (Benjamin et al., 2018; Maier & Lakens, 2022). Hence, Type I error rate inflation should be distinguished from unconventional alpha levels. In particular, Type I error rate inflation should be judged by comparing the actual Type I error rate for a statistical inference with the nominal Type I error rate for that inference (i.e., the alpha level) on the understanding that the nominal level may be set at a conventional or unconventional level. For example, although an actual Type I error rate of .100 may be higher than a field's conventional alpha level of .050, it cannot be said to be "inflated" if the researcher has explicitly set their alpha level at the unconventional level of .100. If Type I error rate inflation was judged relative to the conventional alpha level, rather than the nominal alpha level, then any researcher who set their alpha level higher than the conventional level could be said to have an inflated Type I error rate!

Type I Error Rates Do Not Refer to Theoretical Errors

It is important to distinguish between *statistical inferences* and *theoretical inferences* because Type I error rates only refer to the former (e.g., Meehl, 1978, p. 824; Meehl, 1997, p. 401; see also Bolles, 1962; Chow, 1998; Cox, 1958, p. 357; Hager, 2013, p. 259; Mayo & Spanos, 2006, p. 341; Neyman, 1950, p. 290). During significance testing, a statistical inference refers to a statistical null hypothesis which states that samples are drawn from a study-specific null population in the context of random sampling error per se. Statistical inferences are always supported by inferential statistics, and they usually describe test results as being either "significant" or "nonsignificant." For example, the following statement is a statistical inference: "Compared to male participants, female participants reported significantly more positive attitudes towards ice cream, $t(326) = 2.62$, $p = .009$." In this example, the researcher has provisionally rejected the statistical null hypothesis that female participants *do not* report more positive attitudes towards ice cream than male participants. Their alpha level (e.g., .050) indicates the frequency with which they would make an error in rejecting this statistical null hypothesis in a long run of random sampling from the statistical null population.

In contrast to statistical inferences, theoretical inferences refer to substantive hypotheses that generalize beyond the specifics of the current study. Consequently, they are not directly associated with study-specific Type I error rates. For example, a substantive inference might be that, "compared to men, women have more positive attitudes towards ice cream."

The distinction between statistical and theoretical inferences leads to a parallel distinction between statistical and theoretical errors. Statistical errors refer to Type I and Type II errors. In contrast, theoretical errors refer to a wide range of misinterpretations about (a) theory (e.g., misinterpreted theoretical rationales, hypotheses, and predictions), (b) methodology (e.g., misspecified participant populations, sampling procedures, testing conditions, stimuli, manipulations, measures, controls, etc.), (c) data (e.g., misspecified procedures for data selection, entry, coding, cleaning, aggregation, etc.), and (d) analyses (e.g., misspecified statistical models and assumptions, misinterpreted statistical results).

Theoretical errors may occur in the absence of statistical errors. In other words, researchers may make theoretical misinterpretations before and after correctly rejecting statistical null hypotheses. Meehl's (1990b, 1997) concept of *crud* provides a good example. Crud is a real but theoretically trivial effect (e.g., a methodological artefact). As Meehl (1990b) explained, crud consists of "*real* differences, *real* correlations, *real* trends and patterns" (pp. 207-208, emphasis in original). Hence, crud "does not refer to statistical error, whether of the first or the second kind" (i.e., Type I or II errors; Meehl, 1997, p. 402). In particular, "we are not dealing here with some source of statistical error (the occurrence of random sampling fluctuations). That source of error is limited by the significance level we choose" (Meehl, 1990b, p. 207). Nonetheless, researchers may

make theoretical errors about crud by misinterpreting it as theoretically important effects. Such errors may be conceptualized as *theoretical* false positives (i.e., incorrectly accepting crud as being supportive of a substantive alternative hypothesis) rather than *statistical* false positives (i.e., incorrectly rejecting a statistical null hypothesis).

Theoretical errors may also have a larger impact than statistical errors (Bolles, 1962, p. 645; Cox, 1958, p. 357; Fisher, 1926, pp. 504-505; Greenland, 2017b, p. 640). Hence, a researcher's probability of incorrectly rejecting a substantive null hypothesis and incorrectly accepting a substantive alternative hypothesis may be greater than their alpha level because their decisions are influenced by numerous theoretical errors in addition to Type I errors.

Researchers may sometimes confuse statistical errors with theoretical errors and assume that their Type I error rate indicates the probability of incorrectly rejecting a substantive null hypothesis in the real world rather than a statistical null hypothesis in an imaginary long run of repeated sampling. Greenland (2017b, 2023) described this confusion as "statistical reification." He argued that researchers sometimes forget that their "statistical analyses are merely thought experiments" based on idealized assumptions that are unlikely to be true in the real world. The outcome of this confusion is "overconfident inference" (Greenland, 2017a, p. 4; Greenland, 2017b, p. 640; see also Brower, 1949, p. 327; Gigerenzer, 1993, p. 329; Meehl, 1990b, p. 225). In particular, researchers may have unwarranted credulity in a substantive conclusion based on their incorrect belief that the Type I error rate covers one or more theoretical errors.

Finally, unlike statistical errors, theoretical errors cannot be quantified. As Meehl (1997) explained, "it is tempting to conflate the inference relation between statistics and parameters with the relation between accepted parameter values and the substantive theory; and because the former is numerified (e.g., a Bayesian posterior, a confidence belt, a significance level), one tends to think the latter is numerified also, or (somehow) *should* be" (p. 397; emphasis in original). However, as Neyman and Pearson (1928) explained, "the sum total of the reasons which will weigh with the investigator in accepting or rejecting the hypothesis can very rarely be expressed in numerical terms. All that is possible for him is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon a priori or a posteriori considerations" (p. 176). Note that rigorous pretesting, validation, and preregistration of a priori considerations and testing assumptions do not make them infallible and do not exclude their contribution to the overall probability of making a substantive error in scientific research (Uygun-Tunç & Tunç, 2023, p. 6). Even in the most rigorous research, the probability of a substantive error refers to not only statistical errors but also the unquantifiable and unknowable probability that the researcher has made a theoretical error.

The Impact of Various Research Practices on Type I Error Rates

Model Misspecification

Model misspecification does not inflate Type I error rates because a Type I error rate assumes that the associated null model is "true" or at least "adequate," which means that it is correctly (adequately) specified, and that the only source of influential error is random sampling error. To argue that modelling error inflates Type I error rates is to commit the Bayesian inversion fallacy and believe that Type I error rates are influenced by the probability of the correctness of the model to which they refer (Gigerenzer, 2018; Greenland et al., 2016; Pollard & Richardson, 1987).

This is not to say that null models are always correctly specified. The point here is only that the frequentist concept of a Type I error rate assumes that they are. Of course, in reality, null models may be misspecified. In particular, (a) the statistical null model may not adequately represent the experimental null model, and/or (b) the experimental null model may not adequately represent the theoretical null model (Devezer & Buzbas, 2023; Spanos, 2006). These model misspecifications may then lead to serious inferential errors. However, these errors are theoretical rather than statistical. Hence, it is more appropriate to conceive model misspecification as inflating *Type III* errors, rather than Type I errors. As Dennis et al. (2019) explained, Type III errors occur

when "neither the null nor the alternative hypothesis model adequately describes the data (Mosteller, 1948)" (p. 2).

Multiple Testing

The term *multiple testing* covers several types of testing situation. Here, I distinguish between (a) single tests of multiple individual hypotheses and (b) multiple tests of a single joint hypothesis. I argue that Type I error rate inflation never occurs in the first situation, and that it is not problematic in the second situation because it is easily identified and resolved.

Single Tests of Multiple Individual Null Hypotheses

Imagine that a researcher conducts a study in which they test for gender differences on 20 different dependent variables that measure a variety of different attitudes (e.g., attitudes towards abortion, environmentalism, the death penalty, pizza, ice cream, and so on). In this case, the researcher is testing 20 different null hypotheses (i.e., $H_{0,1,}$ $H_{0,2,}$ $H_{0,3,}$ $_{,\ldots}$ $H_{0,20}$). Further imagine that the researcher sets their alpha level at the conventional level of .050 for each statistical inference that they make about each null hypothesis. We can call this alpha level the *individual* alpha level (i.e., $\alpha_{Individual}$ = .050) because it refers to the frequency with which the researcher would incorrectly reject each individual null hypothesis. This type of testing situation represents single tests of multiple individual null hypotheses. In this case, there is no more than one opportunity to make a Type I error in relation to each individual null hypothesis because none of the hypotheses undergo more than one test.

In a review of 109 research articles published in the journals *Behavior Research Methods* and *Psychological Science* between 2021 and June 2022, García-Pérez (2023) found that *all* cases of multiple testing represented single tests of multiple individual hypotheses. Extrapolating from this evidence, it is likely that single tests of multiple individual hypotheses represent the most common type of multiple testing.

It has also been repeatedly shown that single tests of multiple individual hypotheses do not result in Type I error rate inflation, regardless of how many single tests are undertaken (Armstrong, 2014, p. 505; Cook & Farewell, 1996, pp. 96–97; Fisher, 1971, p. 206; García-Pérez, 2023, p. 15; Greenland, 2021, p. 5; Hewes, 2003, p. 450; Hitchcock & Sober, 2004, pp. 24-25; Hurlbert & Lombardi, 2012, p. 30; Matsunaga, 2007, p. 255; Molloy et al., 2022, p. 2; Parker & Weir, 2020, p. 564; Parker & Weir, 2022, p. 2; Rothman, 1990, p. 45; Rubin, 2017b, pp. 271–272; Rubin, 2020a, p. 380; Rubin, 2021a, 2021c, pp. 10978-10983; Rubin, 2024, p. 3; Savitz & Olshan, 1995, p. 906; Senn, 2007, pp. 150-151; Sinclair et al., 2013, p. 19; Tukey, 1953, p. 82; Turkheimer et al., 2004, p. 727; Veazie, 2006, p. 809; Wilson, 1962, p. 299). Consequently, it can be concluded the most common form of multiple testing does not inflate Type I error rates.

People sometimes doubt the absence of Type I error rate inflation during single tests of multiple individual null hypotheses, but it is easy to demonstrate: In general, the actual Type I error rate is computed using the formula $1 - (1 - \alpha)^k$, where k is the number of tests that are used to make a decision about a specific null hypothesis. During single tests of multiple individual null hypotheses, $k = 1$ because only one test is used to make a decision about each null hypothesis. Hence, the actual Type I error rate for each statistical inference is equal to $1 - (1 - \alpha_{Individual})^1$, which is equal to the nominal $\alpha_{Individual}$ (e.g., $1 - [1 - .050]^1 = .050$; see Rubin, 2024, Confusion II).

Multiple Tests of a Single Joint Null Hypothesis

Now imagine that the researcher groups some of the 20 dependent variables together for some reason. For example, they might consider attitudes about abortion, environmentalism, and the death penalty to be theoretically exchangeable in the context of a broader joint hypothesis about political orientation. In this case, the researcher might be prepared to accept a significant gender difference in relation to *at least one* of these three hypotheses in order to make a statistical inference that there is a significant gender difference in political orientation. Here, the three null hypotheses ($H_{0,1}$, $H_{0,2}$, & $H_{0,3}$) are treated as *constituent* null hypotheses that form part of a broader *joint* intersection null hypothesis about political orientation: "$H_{0,1}$ and $H_{0,2}$ and $H_{0,3}$." In this case, the rejection of *any one* of the three constituent null hypotheses is sufficient to reject the entire intersection null hypothesis and make the

statistical inference that, for example, "compared to male participants, female participants reported significantly more left-wing attitudes: abortion $t(326) = 2.54$, $p = .011$; environmentalism $t(326) = .030$, $p = .979$; death penalty $t(326) = 1.44$, $p = .150$." In this example, the test statistics refer to each of the tests of the three constituent hypotheses and, because at least one of the p values is significant at the conventional level ($p = .011$), the researcher can reject the entire joint null hypothesis about "left-wing attitudes."

During this *union-intersection testing* (e.g., Hochberg & Tamrane, 1987, p. 28; Kim et al., 2004), the actual familywise Type I error rate for the joint null hypothesis is always larger than the nominal alpha level for each of the constituent hypotheses: $\alpha_{Constituent}$. For example, if $\alpha_{Constituent}$ is set at .050, then the familywise error rate will be $1 - (1 - \alpha_{Constituent})^k$, where k is the number of constituent null hypotheses that are included in the joint null hypothesis. Hence, in the present example, the familywise error rate will be $1.00 - (1 - .050)^3 = .143$.

One concern here is that the Type I error rate for decisions about each of the *constituent* null hypotheses becomes inflated. However, this concern is unwarranted. $\alpha_{Constituent}$ is the nominal alpha level for the per comparison Type I error rate, and this error rate does not become inflated during multiple tests of a single joint null hypothesis for the same reason that $\alpha_{Individual}$ does not become inflated during single tests of multiple individual null hypotheses (i.e., $1 - [1 - \alpha_{Constituent}]^1 = \alpha_{Constituent}$; Rubin, 2021c, p. 10979; Rubin, 2024; Tukey, 1953).

Another concern is that the Type I error rate for the decision about the *joint* null hypothesis can become inflated. This concern is legitimate. In order to identify this Type I error rate inflation, we need to check whether the *actual* familywise Type I error rate for the *joint* null hypothesis is higher than the *nominal* familywise Type I error rate for the *joint* null hypothesis: α_{Joint}. If both $\alpha_{Constituent}$ and α_{Joint} are set at the same level (e.g., the conventional level of .050), then the actual Type I error rate for the joint null hypothesis will be inflated above α_{Joint}. However, researchers can avoid this inflation by adjusting $\alpha_{Constituent}$ downwards until the familywise Type I error rate is at α_{Joint}. For example, if $k = 3$ and both $\alpha_{Constituent}$ and α_{Joint} are originally set at .050, then a Bonferroni adjustment (i.e., $\alpha \div k$) can be used to reduce $\alpha_{Constituent}$ from .050 to .017 in order to maintain the familywise Type I error rate for the joint null hypothesis at the nominal α_{Joint} of .050.

What happens if researchers do not adjust $\alpha_{Constituent}$? In this case, there will be Type I error rate inflation. However, this inflation can be easily identified and resolved by readers. For example, reconsider the previous statistical inference: "Compared to male participants, female participants reported significantly more left-wing attitudes: abortion $t(326) = 2.54$, $p = .011$; environmentalism $t(326) = .030$, $p = .979$; death penalty $t(326) = 1.44$, $p = .150$." Here, it is clear that three test results are used to make a single statistical inference about a joint hypothesis that is broader than any of the three constituent hypotheses (i.e., "significantly more left-wing attitudes"). In the absence of any other information, we can assume that the tests use a conventional unadjusted $\alpha_{Constituent}$ of .050. It is also clear that not all of the tests need to be significant to make the statistical inference (i.e., ps = .011, .979, & .150). Hence, the statistical inference is based on union-intersection testing. Finally, we can assume that α_{Joint} has also been set at the conventional level of .050. Consequently, we can conclude that the actual familywise Type I error rate for this statistical inference is greater than a conventional α_{Joint} of .050. In other words, there is Type I error rate inflation. However, the inflation is transparent, and so it can be the target of criticism by reviewers and readers. The inflation is also easily computed (.143 relative to a conventional α_{Joint} of .050). Finally, the inflation is easily resolved. For example, any reader can implement a Bonferroni adjustment to remove the Type I error rate inflation and conclude that, using a conventional α_{Joint} of .050 and an adjusted $\alpha_{Constituent}$ of .017, the researcher's overall statistical inference would remain valid due to the $p = .011$ result.

Is transparent reporting necessary to identify and resolve Type I error rate inflation in this situation? In particular, to determine k, do readers need to be aware of all other tests that a researcher conducted? No, they do not, because the researcher is obviously not using any other test results to make their statistical inference. More generally, k is the number of tests that are formally associated with a reported statistical inference. k is not necessarily the

number of tests that a researcher conducted, or planned to conduct, in their study. This position is consistent with philosophies of science and statistics that focus on the logical relations between a hypothesis and its evidence without considering their psychological origins (Fisher, 1956; Popper, 1962, 2002). Following this philosophy, a familywise error rate can be logically deduced from a potentially unplanned reported statistical inference rather than needing to be predetermined based on planned inferences that may never eventuate (cf. Mayo, 1996, p. 296). Furthermore, undisclosed alterations to planned inferences do not impact the validity of a familywise error rate that is logically deduced from an unplanned statistical inference.

Certainly, a familywise error rate may be logically inconsistent with a reported statistical inference (Rubin, 2024). For example, a reported statistical inference may be formally associated with *three* significance tests when the familywise error rate is computed on the basis of only *two* of these tests. However, as previously discussed, this logical error is readily identifiable by checking (a) the number of tests that are reported as being associated with a statistical inference and (b) the number of tests that are reported as being used to compute the associated familywise error rate. Again, it is not necessary to consult a preregistered plan to identify and resolve this logical inconsistency in the reported information.

Finally, it is worth noting that flexible theorizing and flexible theories allow the post hoc construction of multiple different joint null hypotheses and associated familywise error rates. As I discuss later in the section on *p*-hacking, researchers' personal biases (including hindsight and confirmation biases) may influence which constructions are tested and reported. However, scientific hypotheses are usually judged on the basis of theoretical virtues (e.g., plausibility, universality, precision, depth, breadth, coherence, parsimony, etc.; Kuhn, 1977, p. 103; Mackonis, 2013; Popper, 1962, p. 56, p. 232; Popper, 2002, p. 438) rather than a researcher's personal motives and biases (e.g., to try to report a significant result and publish a paper; Reichenbach, 1938, p. 5; Popper, 1962, p. 140; Popper, 2002 p. 7; Rubin, 2022, pp. 541-542; Rubin & Donkin, 2022, p. 19). Consequently, the lack of transparency about these personal biases does not prevent a rigorous appraisal of the theoretical rationales in question (Rubin, 2017, p. 314; Rubin, 2022, p. 539).

Studywise Type I Error Rates

Researchers are sometimes concerned about the probability of making at least one Type I error in their study or experiment. This concern about a *studywise* or *experimentwise* Type I error rate implies that they are making a statistical inference about a joint studywise null hypothesis that can be rejected by *any* single significant result in their study. In practice, however, it is not common for researchers to make this type of inference because joint studywise null hypotheses do not usually have a useful theoretical basis. Consequently, most researchers should not be concerned about their studywise Type I error rate because it relates to a joint null hypothesis that they are not testing. Instead, researchers should be concerned about the error rates for the individual and/or joint hypotheses about which they actually make statistical inferences (Rubin, 2021c, p. 10991; Rubin, 2024).

To illustrate, reconsider the previous example in which a researcher tested for gender differences on 20 different dependent variables that measured a wide range of attitudes. Recall that three of these attitudes could be grouped into a theoretically meaningful joint null hypothesis about political orientation (i.e., attitudes towards abortion, environmentalism, and the death penalty). However, the other attitudes had nothing to do with political orientation and so had no theoretical basis for being included in this joint null hypothesis (e.g., attitudes about pizza and ice cream). The same issue of theoretical relevance applies to the consideration of joint studywise null hypotheses. Hence, in the present example, the researcher may not have a good theoretical basis for considering all 20 null hypotheses as constituents of a single joint null hypothesis that can be rejected following at least one significant gender difference. This being the case, they should not make a statistical inference about this studywise null hypothesis, and they should not be concerned about its associated studywise Type I error rate.

To be clear, I am not claiming that studywise error rates are *always* irrelevant. They are rele-

vant whenever researchers make statistical inferences about associated joint studywise null hypotheses on the basis of union-intersection testing and, in this case, researchers should adjust their $\alpha_{Constituent}$ in order to control their studywise error rate at α_{Joint}. My point is that this situation is likely to be relatively rare because, typically, a single study will include a collection of disparate hypotheses that do not form a theoretically meaningful joint studywise null hypothesis in aggregate (for similar views, see Bender & Lange, 2001, p. 343; Hancock & Klockars, 1996, p. 270; Hewes, 2003, p. 450; Hochberg & Tamrane, 1987, p. 7; Morgan, 2007, p. 34; Oberauer & Lewandowsky, 2019, p. 1609; Parker & Weir, 2020, p. 2; Perneger, 1998, p. 1236; Rothman et al., 2008, pp. 236-237; Rubin, 2017b, p. 271; Rubin, 2020a, p. 382; Schulz & Grimes, 2005, p. 1592). In such cases, a statistical inference about a joint studywise null hypothesis would need to be relatively vague and atheoretical, along the lines of: "The study's effect was significant." It is not common for researchers to make this sort of abstract atheoretical statistical inference. Instead, researchers usually make more specific, theory-based statistical inferences that relate to substantive theoretical claims. Hence, as in the gender differences example, the researcher might infer that, "compared to male participants, female participants reported significantly more left-wing attitudes: abortion $t(326) = 2.54$, $p = .011$; environmentalism $t(326) = .030$, $p = .979$; death penalty $t(326) = 1.44$, $p = .150$."

Summary

Type I error rate inflation is neither common nor problematic during multiple testing. The most common form of multiple testing is represented by single tests of multiple individual null hypotheses (García-Pérez, 2023). Most experts agree that this type of multiple testing does not inflate relevant Type I error rates because only one test is used to make a decision about each null hypothesis. Consequently, the most common form of multiple testing does not inflate Type I error rates. Indeed, the inappropriate (illogical) use of alpha adjustments in this type of testing situation may result in Type I error rate *deflation*, rather than *inflation* (Rubin, 2024).

A less common form of multiple testing is represented by multiple tests of a single joint null hypothesis (union-intersection testing). Type I error rate inflation is possible in this case because, if both $\alpha_{Constituent}$ and α_{Joint} are set at the same level, then the actual Type I error rate for the joint null hypothesis will be inflated above α_{Joint}. However, this inflation can be easily identified by comparing (a) the number of statistical tests that are formally associated with a reported statistical inference and (b) the number of statistical tests that are used to compute the associated familywise error rate. Type I error rate inflation will be readily apparent if (a) is larger than (b), and it can be easily resolved by recomputing the familywise error rate using the correct number of tests as represented by (a).

Finally, multiple testing increases the studywise Type I error rate above $\alpha_{Constituent}$. However, researchers do not usually make statistical inferences about the associated joint studywise null hypotheses, and so this increase is usually irrelevant. Nonetheless, if it does become relevant, then it can be easily identified and resolved by adjusting $\alpha_{Constituent}$ as described above.

Selective Inference

Imagine that a researcher checks the effect sizes of 100 correlations between 200 different variables (e.g., x_1–y_1, x_2–y_2, x_3–y_3, ... x_{100}–y_{100}) and then decides to perform a significance test on the correlation between variables x_{57} and y_{57} because it had the largest effect size (for a similar example, see Taylor & Tibshirani, 2015, p. 7629). If the researcher uses the significant x_{57}–y_{57} result to reject a joint null hypothesis that could be rejected by a significant result on *any* of the 100 constituent correlations, then they would need to adjust their $\alpha_{Constituent}$ downwards to prevent their familywise error rate from exceeding their α_{Joint}. However, it is uncommon for researchers to make such a *selective inference*. Instead, researchers tend to make statistical inferences that are *limited to* their selected test and data rather than extended to unselected tests and data that they could have used (Birnbaum, 1962, pp. 278-279; Cox, 1958, p. 359-361; Cox & Mayo, 2010, p. 296; Lehmann, 1993, pp. 1245-1246; Mayo, 2014, p. 232; Reid & Cox, 2015, p. 300). Hence, the researcher in

the present example would be more likely to use the significant x_{57}–y_{57} result to reject an individual null hypothesis about the relationship between x_{57} and y_{57} rather than a joint intersection null hypothesis about the relationships between x_1–y_1, x_2–y_2, x_3–y_3, ... x_{100}–y_{100}. In the individual case, $\alpha_{Constituent}$ and α_{Joint} are irrelevant to the researcher's inference, and an unadjusted $\alpha_{Individual}$ can be used without any concern about Type I error rate inflation. Note that, as discussed previously, this point stands even if the researcher does not disclose that they considered the other 99 correlations because those correlations are not formally associated with their reported statistical inference, which is about the relationship between x_{57} and y_{57}.

Alternatively, the researcher may decide to make a statistical inference about a joint null hypothesis based on a family of correlations, such as "x_{56}–y_{56} and x_{57}–y_{57} and x_{58}–y_{58}." For example, in an fMRI study, these correlations might occur within the same brain region and, again, they may be selected because they "look promising." In this case, the familywise error rate can be logically deduced from the statistical inference (i.e., $k = 3$), and the formal theoretical rationale for the joint hypothesis can be critically evaluated on the basis of its theoretical virtues. Again, the familywise error rate for the statistical inference is not inflated by the existence of alternative planned or unplanned statistical inferences that could have been based on other joint null hypotheses from the set of correlations.

People sometimes confuse individual and constituent null hypotheses in this situation. For example, it is true that $\alpha_{Constituent}$ would need to be adjusted when using the significant x_{57}–y_{57} correlation to reject a joint null hypothesis that could be rejected by at least one significant result from among 100 potential correlations, because $k = 100$ for this selective inference. Similarly, as described above, $\alpha_{Constituent}$ would need to be adjusted when using the significant x_{57}–y_{57} correlation to reject a joint null hypothesis that could be rejected by at least one significant result from among three potential correlations (e.g., "x_{56}–y_{56} and x_{57}–y_{57} and x_{58}–y_{58}"), because $k = 3$ for this selective inference. However, this point does not imply that $\alpha_{Individual}$ needs to be adjusted when using the significant x_{57}–y_{57} result to reject the x_{57}–y_{57} null hypothesis, because $k = 1$ for this individual inference.

Part of the confusion here is that people assume that selective *analyses* always lead to selective *inferences*. On the contrary, selective analyses usually lead to inferences that are *limited to the selection*. For example, a researcher may select data and/or variables for analysis from a broader set. However, this act does not then obligate them to make a selective inference about a joint null hypothesis that refers to other data or variables from that set. Instead, it would be more usual for them to limit their inference to the data or variables that they have selected (i.e., an *unconditional* inference about the selected data rather than a *conditional* inference that is conditioned on the selection procedure).

More generally, it is important to avoid confusion about the reference sets to which statistical inferences refer. As Neyman (1950) explained, "many errors in computing probabilities are committed because of losing sight of the set of objects to which a given probability is meant to refer" (p. 15). A Type I error rate is meant to refer to a decision about a specified statistical (individual or joint) null population and not to any broader population from which that null population may have been selected. Hence, the selection of a particular subset of data for testing from a more inclusive set because it "looks promising" will not inflate the Type I error rate for a statistical inference as long as that inference refers to a population based on *that particular subset of data* and not to a population based on the more inclusive set of data (for related discussions, see Fisher, 1956, p. 88-89; Kotzen, 2013, p. 167; Rubin, 2021c, p. 10983). For example, conducting multiple tests of a gender difference in the preference for ice cream in different European countries and then selectively reporting the results for France may bias claims about a gender difference in Europe but not claims about a gender difference in France. Of course, the result in France may represent a Type I error. However, this possibility is covered by the alpha level of the test that was conducted on the French data.

Forking Paths

A forking path occurs during data analysis when a result from one sample of data inspires a researcher to conduct a specific test in a situation in which they would have conducted a different test if they had observed a different result using a different sample (Gelman & Loken, 2014). For example, a researcher might report that "this variable was included as a covariate in the analysis because it was significantly correlated with the outcome variable." The implication here is that the variable would *not* have been included as a covariate if it had *not* been significantly correlated with the outcome variable in a different sample of data. Consequently, if the researcher makes a statistical inference about a joint null hypothesis that can be rejected following a significant result on at least one of the two tests (i.e., the test that includes the covariate and the test that does not), then their familywise Type I error rate will be greater than the $\alpha_{Constituent}$ for each test (Rubin, 2020a, p. 380). Hence, the forking paths problem resolves to a case of multiple testing in which the "invisible multiplicity" is only apparent in a long run of repeated sampling (Gelman & Loken, 2014, p. 460).

The forking paths problem assumes that a researcher will make a statistical inference about a joint null hypothesis that comprises the two forking paths in their analysis (e.g., a test that includes a covariate and a test that does not). *If* the researcher makes this statistical inference, then they can adjust their $\alpha_{Constituent}$ to retain their α_{Joint} at the actual familywise error rate (i.e., $\alpha_{Constituent} \div 2$; Rubin, 2017a, p. 324). However, it is more likely that the researcher would make a more limited statistical inference based on only *one* test in *one* of the two forking paths (Birnbaum, 1962, pp. 278-279; Cox, 1958, p. 359-361; Cox & Mayo, 2010, p. 296; Lehmann, 1993, pp. 1245-1246; Mayo, 2014, p. 232; Reid & Cox, 2015, p. 300). In this case, the researcher's inference would refer to an imaginary long run of repeated sampling that, for example, always includes the variable as a covariate and never excludes it. An unadjusted $\alpha_{Individual}$ would then be appropriate. Note that the Type I error rate is not inflated in either of these two situations.

Exploratory Analyses

An exploratory data analysis is one in which a study's analytical approach is guided by idiosyncratic results in one sample of data that may not occur in other samples. Consequently, the tests that are undertaken in one instance of an exploratory analysis may be quite different to those undertaken in repetitions of that analysis. This issue leads to multiple tests of a joint studywise null hypothesis both *within each repetition* of the exploratory study and *across the long run of its repetitions*. In theory, the associated studywise error rate would then need to account for every null hypothesis that could possibly be tested during the exploratory analysis and its repetitions. This situation has led several people to conclude that the exploratory studywise error rate cannot be computed or controlled (e.g., Hochberg & Tamrane, 1987, p. 6; Nosek & Lakens, 2014, p. 138; Wagenmakers, 2016). As Wagenmakers (2016) explained, "the problem is one of multiple comparisons with the number of comparisons unknown (De Groot, 1956/2014)."

There are three problems with this line of reasoning. First, in practice, a researcher is *least* likely to make a statistical inference about a joint studywise null hypothesis during an *exploratory* analysis because, in this situation, they are least likely to have a satisfactory theoretical rationale for aggregating a potentially infinite number of result-contingent null hypotheses into a single joint studywise null hypothesis.

Second, if a researcher does make a statistical inference about an exploratory joint studywise null hypothesis, then it would need to be relatively abstract and atheoretical (e.g., "the study's effect was significant"). Again, it is not common for researchers to make this type of statistical inference. Instead, they are likely to make more specific theory-based statistical inferences that relate to substantive theoretical inferences.

Finally, if a researcher was to make a statistical inference about an exploratory joint studywise null hypothesis, then they would need to tie it to specific statistical results. They could not simply report that, "based on an unknown number of unspecified tests that could have been conducted, the study's effect was significant." They would need to specify the tests (ac-

tual and potential) and results (*p* values) that form the basis for their statistical inference. Of course, they would only be able to specify a finite number of tests and, consequently, it would be possible for them to specify *k* and adjust $\alpha_{Constituent}$ in order to prevent Type I error rate inflation. In other words, the act of specifying a statistical inference includes making known the statistical tests upon which it rests. It is worth noting that recent work on multiverse analyses and specification curve analyses demonstrates the feasibility of making known large numbers of diverse statistical tests during exploratory data analyses (Del Giudice & Gangestad, 2021; Simonsohn et al., 2020; Steegen et al., 2016).

In summary, in most cases, there is no need for researchers to be concerned about the inflation of an exploratory studywise Type I error rate because this error rate is irrelevant to the more limited and theoretically defined statistical inferences that they usually make. However, if researchers do proceed to make vague atheoretical statistical inferences about exploratory joint studywise hypotheses, then they will need to specify the tests involved, and so they will be able to specify *k*, adjust $\alpha_{Constituent}$, and control the associated studywise error rate at α_{Joint}.

P-Hacking

P-hacking is a questionable research practice that is intended to find and selectively report significant results. In their seminal article on "false positive psychology," Simmons et al. (2011) proposed that *p*-hacking inflates Type I error rates due to multiple testing. As they explained (p. 1359),

> *it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.*

On this basis, Simmons et al. (2011, p. 1359) argued that "undisclosed flexibility in data collection and analysis allows presenting anything as significant."

A more formal analysis of Simmons et al.'s (2011) argument is as follows: If both $\alpha_{Constituent}$ and α_{Joint} are set at .050 during the union-intersection testing of an exploratory joint studywise null hypothesis, then the actual exploratory studywise error rate will exceed α_{Joint}, and it will be easier to reject the exploratory joint studywise null hypothesis than it would be if the studywise error rate matched α_{Joint}. Importantly, however, this situation does not allow researchers to present "anything as significant." It only allows researchers to present the *exploratory joint studywise alternative hypothesis* as significant and, given its atheoretical rationale, this hypothesis will be abstract and scientifically useless, akin to: "The study's effect was significant." Again, researchers tend to make more specific and theoretically informative statistical inferences based on (a) single tests of theory-based individual hypotheses and (b) multiple (union-intersection) tests of theory-based joint null hypotheses. Hence, I consider the implications of *p*-hacking in each of these two contexts below.

P-Hacking During Single Tests of Multiple Individual Null Hypotheses

As explained previously, the actual Type I error rate for a single test of an individual null hypothesis remains at $\alpha_{Individual}$ even if (a) multiple such tests are conducted side-by-side within the same study, and (b) some of the tests are conducted but not reported (for related discussions, see Rubin, 2017a, 2020a, 2021c, 2024). Simmons et al. (2011) are correct that, when $k \geq 2$, familywise Type I error rates will be greater than $\alpha_{Constituent}$ and greater than .050 when $\alpha_{Constituent}$ is set at .050. However, neither of these points imply the inflation of Type I error rates for statistical inferences based on single tests of multiple individual null hypotheses. The argument that *p*-hacking causes Type I error rate inflation during single tests of multiple individual null hypotheses confuses statistical inferences about *individual* null hypotheses with statistical inferences about *joint* null hypotheses.

To illustrate, consider Simmons et al.'s (2011) demonstration of the impact of *p*-hacking on Type I error rates. Simmons et al. performed multiple tests on a real data set until they found a significant result that supported the outlandish theoretical inference that listening to

the Beatles' song "When I'm Sixty-Four" makes people chronologically younger relative to listening to a control song (so-called "chronological rejuvenation"). The researchers then performed a series of simulations to compute the percentage of random samples in which there was at least one significant result in a family of, for example, "three *t* tests, one on each of two dependent variables and a third on the average of these two variables" (i.e., Situation A in Table 1 of Simmons et al., 2011, p. 1361). Using an $\alpha_{Constituent}$ of .050, an actual familywise Type I error rate of .095 was calculated. Note that, because the dependent variables were correlated with one another, this error rate is lower than the expected rate of .150 (Simmons et al., 2011, p. 1365, Note 3). Nonetheless, it is higher than the $\alpha_{Constituent}$ of .050. Consequently, the researchers concluded that "flexibility in analyzing two dependent variables (correlated at r = .50) nearly doubles the probability of obtaining a false-positive finding" (p. 1361). This conclusion is correct. However, the "false-positive finding" in question relates to the incorrect rejection of a *joint* null hypothesis (i.e., that song condition has no effect on *any* of the three dependent variables), not an *individual* null hypothesis (e.g., that song condition has no effect on the first dependent variable), and the statistical inferences that are made in Simmons et al.'s demonstration are about *individual* null hypotheses, not *joint* null hypotheses. Hence, although Simmons et al.'s conclusion is correct, it is also irrelevant to the type of statistical inference that they consider.

To illustrate further, consider this part of a larger example that Simmons et al. (2011, p. 1364) used to demonstrate a fictitious researcher's selective reporting. In the following extract, the italicized text refers to a statistical inference that the researcher decided to report because it referred to a significant result, and the nonitalicized text refers to a statistical inference that the researcher decided not to report because it referred to a nonsignificant result:

> *An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted* M *= 20.1 years) rather than to "Kalimba" (adjusted* M *= 21.5 years),* F*(1, 17) = 4.92,* p *= .040.* Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

Importantly, in the above extract, both of the researcher's statistical inferences are based on single tests of individual null hypotheses, each underwritten by a separate statistical test and *p* value (i.e., k = 1 for each decision about each null hypothesis). Consequently, the actual Type I error rates for each statistical inference are consistent with a conventional $\alpha_{Individual}$ of .050. Furthermore, the fact that the second test is not reported has no impact on the actual Type I error rate of the first test, because the Type I error rate for the first statistical inference refers to an imaginary long run of random sampling in which father's age is *always* included as a covariate in the test.

Simmons et al.'s (2011) concern about Type I error rate inflation would only be warranted if their fictitious researcher made a statistical inference about a *joint* null hypothesis that referred to *both* the italicized and the nonitalicized tests. Obviously, in the current example, the researcher does not report the nonitalicized test, and so they do not make a statistical inference about an associated joint null hypothesis. However, even if the researcher reported both tests as indicated above, they would still not be making a statistical inference about a joint null hypothesis. They would be following the more common approach of making two statistical inferences about two individual null hypotheses. They would only make a statistical inference about a joint null hypothesis if they used two or more significance tests to make a single decision about a single (joint) null hypothesis (Rubin, 2024). To illustrate, the following example represents a case of Type I error rate inflation during a statistical inference about a joint null hypothesis (assuming that $\alpha_{Constituent}$ and α_{Joint} are both set at .050):

> An ANCOVA found that type of song had a significant effect on birth date (father's age included as a covariate, *F*(1, 17) = 4.92, *p* = .040; father's age excluded as a covariate, *F*(1, 18) = 1.01, *p* = .330.

Contrary to the present view, Nosek et al.

(2018) argued that nontransparent selective reporting inflates the Type I error rate. Consequently, as they explained, "transparent reporting that 1 in 20 experiments or 1 in 20 analyses yielded a positive result will help researchers identify the one as a likely false positive" (p. 2603). However, like Simmons et al. (2011), this explanation confuses statistical inferences about individual hypotheses with statistical inferences about joint hypotheses. If researchers make a statistical inference based on a single test of an individual hypothesis, then their actual Type I error rate for this inference will be $\alpha_{Individual}$ regardless of whether they make 1, 20, or a million other statistical inferences and even if their statistical result for that inference is the only significant result that they obtain or report. On the other hand, if researchers make a statistical inference based on the union-intersection testing of a joint null hypothesis that includes 20 constituent experiments or analyses, then their actual familywise Type I error rate for that inference will be $1 - (1 - \alpha_{Constituent})^{20}$. However, contrary to Nosek et al., this familywise error rate does not refer to a "likely false positive" in relation to any "one" experiment or analysis (i.e., a constituent hypothesis; Rubin, 2024). It refers to an incorrect decision about a *joint* null hypothesis that is based on the entire *family* of 20 experiments or analyses. Again, (a) if researchers do not make a statistical inference about this joint null hypothesis, then there is no need for them to be concerned about its familywise error rate, (b) researchers do not usually make statistical inferences about joint null hypotheses unless they have some theoretical rationale for doing so, and (c) if researchers do make statistical inferences about joint null hypotheses, then they can adjust their $\alpha_{Constituent}$ in order to control their α_{Joint} at some specified level.

P-Hacking During Union-Intersection Testing of a Joint Null Hypothesis

Does *p*-hacking inflate the actual familywise Type I error rate when a researcher makes a statistical inference about a joint null hypothesis based on union-intersection testing? No, it does not, because the familywise error rate refers to the researcher's specified reported statistical inference and not to any other statistical inference that they made and then failed to report during their *p*-hacking.

To illustrate, imagine that a researcher undertakes 20 union-intersection tests of a joint null hypothesis. They set $\alpha_{Constituent}$ at .0025 in order to maintain α_{Joint} at the conventional level of .050 (i.e., .0025 × 20). Using this $\alpha_{Constituent}$, they find no significant result. However, they notice that the smallest of their 20 *p* values is .004. They decide not to report their first set of union-intersection tests because it failed to reject the joint null hypothesis (i.e., they engage in selective reporting). Instead, they conduct a second set of union-intersection tests that includes 10 of the previous 20 null hypotheses, and they deliberately include the hypothesis that yielded the $p = .004$ result in this family of constituent hypotheses. This time, they set $\alpha_{Constituent}$ at .005, which continues to maintain α_{Joint} at the conventional level of .050 (i.e., .005 × 10). Note that the researcher now knows that the $p = .004$ result will be significant using an $\alpha_{Constituent}$ of .005, and so they know in advance that they are able to reject their new joint null hypothesis. Does this *p*-hacking and selective reporting inflate the actual familywise Type I error rate above α_{Joint}? No, it does not, because the nominal α_{Joint} of .050 matches the actual familywise Type I error rate for the researcher's specified statistical inference, which has a *k* of 10 and an $\alpha_{Constituent}$ of .005.

The fact that *k* was 20 for an unreported statistical inference does not affect the Type I error rate for the reported statistical inference, for which *k* is 10. More generally, the Type I error rate for a statistical inference about an individual or joint null hypothesis is not impacted by other statistical inferences that could, would, or should have been made about other individual or joint null hypotheses. Additionally, it is not impacted by other statistical inferences that were planned or actually made and then either reported or not reported. Arguing that Type I error rates should be adjusted when making multiple statistical inferences confuses $\alpha_{Constituent}$ with $\alpha_{Individual}$ and α_{Joint}. It is necessary to adjust $\alpha_{Constituent}$ when using multiple tests to make a single statistical inference about a single joint null hypothesis. However, it is not necessary to adjust either $\alpha_{Individual}$ or α_{Joint} when making multiple statistical inferences about multiple individual or joint null hypotheses (see Rubin, 2024, Confusion III).

Also, note that the researcher deliberately

selected the reported set of union-intersection tests *because* they yielded a significant result. Again, however, this point does not alter the actual Type I error rate for their statistical inference, which matches their α_{Joint} of .050. Consequently, they are entitled to claim that their test provides a conventionally low rate of erroneous rejection in a hypothetical long run of repeated random sampling.

Finally, in reporting their test, the researcher is likely to provide a theoretical rationale for the inclusion and exclusion of the specific constituent hypotheses in their joint null hypothesis, and they may omit the fact the $p = .004$ result inspired their construction of this hypothesis. Again, however, this situation is not necessarily problematic because a researcher's personal motives and informal inspirations are not usually taken into account during the formal evaluation of scientific hypotheses (Popper, 1962, p. 140; Popper, 2002 p. 7; Reichenbach, 1938, p. 5; Rubin, 2022, pp. 541-542; Rubin & Donkin, 2022, p. 19). Instead, hypotheses tend to be judged on the basis of theoretical virtues (e.g., plausibility, universality, precision, depth, breadth, coherence, parsimony, etc.; Kuhn, 1977, p. 103; Mackonis, 2013; Popper, 1962, p. 56, p. 232; Popper, 2002, p. 438). Hence, in the present example, reviewers and other readers would be able to evaluate the quality of the researcher's theoretical rationale for their joint null hypothesis, even if they are unaware of the researcher's *p*-hacking (Rubin, 2017, p. 314; Rubin, 2022, p. 539). If the theoretical rationale for including and excluding the various constituent hypotheses in the joint hypothesis is cogently deduced from a well-established, coherent theory that explains a broad range of other effects in a relatively deep and efficient manner, then it should be given serious consideration regardless of the timing of its deduction.

Summary

In summary, it is true that *p*-hacking makes it easier to reject an exploratory joint studywise null hypothesis comprised of every null hypothesis that could possibly be tested in a study and its repetitions. However, researchers do not usually make statistical inferences about such hypotheses because they are not usually theoretically informative. For example, researchers do not usually claim that "the study's effect was significant," independent of any theoretical explanation. Instead, they tend to make more specific statistical inferences about (a) single tests of theory-based individual hypotheses and (b) multiple (union-intersection) tests of theory-based joint null hypotheses. *P*-hacking does not usually inflate the relevant Type I error rates in either of these cases.

P-Hacking May Increase Theoretical Errors, Not Statistical Errors

To be clear, I am not denying that *p*-hacking occurs, and I am not arguing that it is always harmless. I am only arguing that it does not usually inflate relevant Type I error rates. In some cases, *p*-hacking may be problematic because it increases *theoretical errors,* rather than *statistical errors*, and it does so as a result of *biased selective reporting* (Rubin, 2020a, p. 383).

For example, in Simmons et al.'s (2011) scenario, the inference that listening to the song "When I'm Sixty-Four" makes people younger may represent a theoretical error rather than a statistical error. In particular, the *p*-hacked result may be a statistical *true positive* that has been misinterpreted as representing "chronological rejuvenation" when it actually represents Meehlian crud or some other statistically real but theoretically misleading effect (Meehl, 1990b, pp. 207-208). In this case, the problem is theoretical misinterpretation rather than Type I error rate inflation.

Theoretical errors are more likely to occur when the overall pattern of theoretically relevant evidence is obscured due to biased selective reporting. Hence, it is important to try to identify and reduce biased reporting through the use of open science practices such as open data, open research materials, and robustness, multiverse, and specification curve analyses. For example, reporting all of the evidence for and against chronological rejuvenation would be likely to show that the *p*-hacked result is part of a tiny minority of confirmatory evidence compared to a vast majority of disconfirmatory evidence, including other significant results that can be misinterpreted as showing the opposite of chronological rejuvenation (i.e., accelerated ageing!).

Importantly, theoretical errors can also be reduced through a rigorous critical evaluation of relevant theory. For example, in Simmons et al.'s (2011) scenario, we might ask: What

larger theoretical framework explains "chronological rejuvenation"? What is the quality of that theory relative to other explanations for the results? How well does that theory justify the specific methodological and analytical decisions that the researcher made (e.g., including father's age as a covariate)? And what other evidence is there for and against the theory in the current study aside from a single ANCOVA result? These theoretical issues were not considered in Simmons et al.'s scenario. However, in practice, they would operate as an important (not infallible) line of defence against theoretical errors by helping to (a) screen out low-quality theories and (b) motivate and guide efforts to detect biased selective reporting (see also Simmons et al., 2011, p. 1363, Point 3; Rubin, 2017, p. 314; Syrjänen, 2023, p. 16).

Neyman and Pearson (1928, p. 232) cautioned that significance "tests should only be regarded as tools which must be used with discretion and understanding, and not as instruments which in themselves give the final verdict" (see also Bolles, 1962, p. 645; Boring, 1919, pp. 337-338; Chow, 1998, p. 169; Cox, 1958, p. 357; Hager, 2013, p. 261; Haig, 2018, p. 199; Lykken, 1968, p. 158; McShane et al., 2023; Meehl, 1978, p. 824; Meehl, 1997, p. 401; Szollosi & Donkin, 2021, p. 5). *P*-hacking is most problematic for those who ignore this advice and rely on *p* values as the sole arbiters of scientific decisions rather than as mere steppingstones on the way to making substantive theoretical inferences during a fallible process of inference to the best explanation (Haig, 2009; Mackonis, 2013).

Optional Stopping

In the case of undisclosed optional stopping or data peeking, a researcher tests a hypothesis using a certain sample size and then collect more data and retest that same hypothesis using a larger sample size if their first test does not yield a significant result. They then continue this process until they obtain a significant result, at which point they report their significant result and hide their nonsignificant results.

Undisclosed optional stopping represents result-dependent multiple testing across a series of tests that have different sample sizes. A key concern here is that the $\alpha_{Constituent}$ for each test needs to be adjusted to account for the union-intersection testing of a joint null hypothesis that will be rejected when one of the tests yields a significant result. Failure to adjust $\alpha_{Constituent}$ will lead to inflation of the familywise error rate above α_{Joint}. A further a concern is that, if the number of "stop-and-tests" is not specified in advance, then the actual familywise Type I error rate will be incalculable in repetitions of an exploratory optional stopping procedure. However, neither of these concerns is warranted.

First, a researcher who engages in undisclosed optional stopping has no choice but to limit their statistical inference to their final reported sample size because, by definition, they do not refer to any of their previous tests that used different sample sizes. In this case, it is appropriate for the researcher to use an unadjusted $\alpha_{Individual}$. Here, $\alpha_{Individual}$ refers to the frequency with which they would make an incorrect decision to reject the specified statistical null hypothesis during an imaginary long run of repeated random sampling in which samples are the same size as that used in the final reported test (e.g., N = 300; Fraser, 2019, p. 140; Reid, 1995, p. 138). This long run would not include any of the other unreported tests that yielded nonsignificant results (e.g., Ns = 270, 280, & 290) or any of the tests that might have occurred had the current test not yielded a significant result (e.g., Ns = 310, 320, 330, etc.). Certainly, it is possible to make an inference about a joint null hypothesis that refers to other tests in the series (see below). However, this would represent a different statistical inference that is warranted by a different (familywise) Type I error rate. The current statistical inference is restricted to a reference set that excludes the unreported tests. As in the case of *p*-hacking, the fact that this inference is reported *because* it refers to a significant result does not alter the individual hypothetical probability of that result. To illustrate, imagine that you throw a 20-sided dice, hoping to get an "8," and you only throw the dice again if you fail to get an "8" on your previous throw. You finally get an "8" on your 20th throw. In this case, it would be correct to report that you had a .050 probability of getting an "8" *on that particular throw* even if you did not report your first 19 unsuccessful throws. This individual (marginal) probability is not invalidated by the

fact that the familywise (union) probability of getting an "8" in *at least one of the 20 throws* is .642. Furthermore, given that your probability statement refers to a single throw, a repetition of the associated procedure would only entail a single throw, and not 20 throws.

Second, if a researcher wanted to adjust their $\alpha_{Constituent}$ to maintain their familywise Type I error rate at α_{Joint} during the process of optional stopping, then they could do so without planning the number of stop-and-tests in advance. Again, relevant Type I error rates refer to reported statistical inferences and not to planned but unreported statistical inferences. Hence, for example, if a researcher planned to adjust their familywise Type I error rate for a series of five stop-and-tests but ended up deviating from that plan and making a statistical inference about a series of only three stop-and-tests, then they should adjust their $\alpha_{Constituent}$ based on $k = 3$, not $k = 5$. In this case, their actual familywise error rate for their specified ($k = 3$) statistical inference would match their α_{Joint} for that inference.

There may be a concern that the researcher may not report their "actual" number of stop-and-tests in the previous example. However, we should not be concerned about the overall number of tests that a researcher has happened to perform. Instead, we should be concerned about the "actual" number of tests that are formally associated with a specific statistical inference (i.e., k) and, in the current case, that number is three, not five.

Finally, as with *p*-hacking, there may also be a concern that a researcher's selective reporting is hiding relevant disconfirming evidence (i.e., biased selective reporting). It is debatable whether the null results from prior stop-and-tests represent "disconfirming evidence" given that null results represent the absence of evidence rather than evidence of absence (Altman & Bland, 1995). Nonetheless, even if null results are accepted as disconfirming evidence, the presence of this undisclosed evidence will not inflate the Type I error rate because, as discussed previously, hiding disconfirming evidence biases theoretical inferences, not statistical inferences. Furthermore, a significant result obtained via undisclosed optional stopping may either confirm or disconfirm a directional hypothesis. Hence, if a researcher stops data collection when they obtain a significant result, regardless of whether that result confirms or disconfirms their hypothesis, then their optional stopping will not bias their theoretical inference about their directional hypothesis (Rubin, 2020a, p. 381).

Double Dipping

It has been proposed that the same data cannot be used to both generate and then test the same hypothesis (e.g., Nosek et al., 2018, p. 2600; Wagenmakers et al., 2012, p. 633). Engaging in this double dipping strategy is thought to inflate Type I error rates. For example, Wagenmakers et al. (2012, p. 633) argued that, "if you carry out a hypothesis test on the very data that inspired that test in the first place then the statistics are invalid… Whenever a researcher uses double-dipping strategies, Type I error rates will be inflated and *p* values can no longer be trusted."

Contrary to this argument, carrying out a hypothesis test on the same data that inspired the test does not necessarily invalidate the statistics. For example, it is perfectly acceptable to use the result from one statistical test to create a statistical null hypothesis for a second test which is then tested using the same data as long as the second test's result is independent from the first test's result. The logical problem of circularity only occurs when the same result is used to both (a) support the theoretical rationale for a hypothesis and (b) claim additional support for that hypothesis (Devezer et al., 2021; Kriegeskorte et al., 2009, p. 535; Rubin & Donkin, 2022, pp. 5-6; Spanos, 2010, p. 216; Worrall, 2010, p. 131). Furthermore, this problem of circularity represents a theoretical error, rather than a statistical error. Consequently, although double dipping may sometimes invalidate theoretical inferences, it does not inflate Type I error rates.

HARKing

Hypothesizing after the results are known, or HARKing, refers to the questionable research practice of "presenting post hoc hypotheses in a research report as if they were, in fact, a priori hypotheses" (Kerr, 1998, p. 197). HARKing is thought to inflate Type I error rates (e.g., Bergkvist, 2020; Stefan & Schönbrodt, 2023, p. 4). However, HARKing represents post hoc *the-*

orizing, and so it affects theoretical inferences rather than statistical inferences. Indeed, in his seminal article on the subject, Kerr (1998, p. 205) did not argue that HARKing inflates Type I error rates. Instead, his concern was that, when "a Type I error is followed by HARKing, then 'theory' is constructed to account for what is, in fact, an illusory effect" (p. 205). In other words, he was not concerned that HARKing inflates Type I error rates, but rather that it may be used to "translate Type I errors into theory" (Kerr, 1998, p. 205).

Kerr (1998) conceded that Type I errors can also be translated into theory following explicit, transparent, post hoc theorizing, rather than undisclosed HARKing. However, he believed that the translation is more problematic following HARKing because "an explicitly post hoc hypothesis implicitly acknowledges its dependence upon the result in hand as a cornerstone (or perhaps, the entirety) of its foundation, and thereby sensitizes the reader to the vulnerability of the hypothesis to the risks of an immediate Type I error" (Kerr, 1998, p. 205). Contrary to this reasoning, "the risks of an immediate Type I error" do not vary as a function of either the origin or quality of a hypothesis. To believe that they do is to commit the Bayesian inversion fallacy. Hence, there is no reason to believe that HARKing either inflates Type I error rates or that it exacerbates the costs of Type I errors.

Evidence for Type I Error Rate Inflation?

What about the evidence of Type I error rate inflation? There are two problems with this evidence that threaten its validity.

First, similar to Simmons et al.'s (2011) demonstration, evidence of Type I error rate inflation tends to confound statistical inferences about *individual* null hypotheses with statistical inferences about *joint* null hypotheses. For example, simulations of actual Type I error rates compute the familywise error rate for a joint null hypothesis and then apply that error rate to individual null hypotheses, claiming that, because the familywise error rate is, for example, .143, there is a .143 chance of incorrectly rejecting *each individual* null hypothesis. Again, this reasoning is widely acknowledged to be incorrect (Armstrong, 2014, p. 505; Cook & Farewell, 1996, pp. 96–97; Fisher, 1971, p. 206; García-Pérez, 2023, p. 15; Greenland, 2021, p. 5; Hewes, 2003, p. 450; Hurlbert & Lombardi, 2012, p. 30; Matsunaga, 2007, p. 255; Molloy et al., 2022, p. 2; Parker & Weir, 2020, p. 564; Parker & Weir, 2022, p. 2; Rothman, 1990, p. 45; Rubin, 2017b, pp. 271–272; Rubin, 2020a, p. 380; Rubin, 2021a, 2021c, pp. 10978-10983; Rubin, 2024, p. 3; Savitz & Olshan, 1995, p. 906; Senn, 2007, pp. 150-151; Sinclair et al., 2013, p. 19; Tukey, 1953, p. 82; Turkheimer et al., 2004, p. 727; Veazie, 2006, p. 809; Wilson, 1962, p. 299).

Second, evidence of Type I error rate inflation may also depend on a fallacious comparison between (a) the probability of rejecting a null hypothesis when it is true and (b) the probability of a null hypothesis being true when it is rejected (Pollard & Richardson, 1987). As discussed previously, the first probability is equivalent to a frequentist Type I error rate: Pr(reject H_0 ; H_0 is true). However, the second probability does not provide an appropriate benchmark against which to judge Type I error rate inflation because it represents a conditional posterior probability about the truth of the null hypothesis: Pr(H_0 is true | reject H_0). Hence, showing that Pr(H_0 is true | reject H_0) > Pr(reject H_0 ; H_0 is true) does not provide a valid demonstration of Type I error rate inflation. Instead, it demonstrates the Bayesian inversion fallacy because it confuses the unconditional probability of rejecting a true null hypothesis with the conditional probability that a null hypothesis is true given that it has been rejected (Gigerenzer, 2018; Greenland et al., 2016; Mayo & Morey, 2017; Pollard & Richardson, 1987).

Implications for the Replication Crisis

Type I error rate inflation may not be a major contributor to the replication crisis. Certainly, some failed replications may be due to Type I errors in original studies. However, actual Type I error rates are rarely inflated above their nominal levels, and so the level of Type I errors in a field is liable to be around that field's conventional nominal level (see also Neyman, 1977, p. 108). Hence, Type I error rate inflation cannot explain unexpectedly low replication rates.

In contrast, theoretical errors may be higher than expected. In particular, unacknowledged misinterpretations of theory, methodology, data, and analyses may all inflate theoretical

errors above their "nominal" expected level, resulting in incorrect theoretical inferences and unexpectedly low replication rates. For example, researchers may assume a higher degree of theoretical equivalence between an original study and a "direct" replication than is warranted. A failed replication may then represent the influence of an unrecognized "hidden moderator" that produces a true positive result in the original study and a true negative result in the replication study. Of course, scientists should attempt to specify and investigate such hidden moderators in future studies (Klein et al., 2018, p. 482). Nonetheless, ignoring hidden moderators does not mitigate their deleterious impact on replicability!

From this perspective, the replication crisis may be explained, at least in part, by researchers' underestimation of theoretical errors and their misinterpretation of statistical errors (i.e., statistical reification; Greenland, 2017b, 2023; see also Brower, 1949; Gigerenzer, 1993). Specifically, researchers may overestimate (a) their theoretical understanding of effects and (b) the extent to which a Type I error rate implies replicability. These two issues may combine to produce overconfident researchers who have unrealistically high expectations about replication rates during "direct" replications (Rubin, 2021, pp. 5828-5829). Accordingly, an appropriate response to the replication crisis is for researchers to adopt a more modest perspective that recognizes (a) the important role of scientific ignorance during theoretical inferences (e.g., Feynman, 1955; Firestein, 2012; Merton, 1987) and (b) the limited scope of Type I error rates during statistical inferences (e.g., Bolles, 1962; Cox, 1958; Fisher, 1926; Greenland, 2017a, 2017b). This more modest perspective may help to provide more realistic expectations about replication rates and a better appreciation of replication failures as a vital aspect of scientific progress (Barrett, 2015; Redish et al., 2018; Rubin, 2021b).

Summary and Conclusion

The replication crisis has been partly explained in terms of Type I error rate inflation. In particular, it has been argued that questionable research practices inflate actual Type I error rates above their nominal levels, leading to an unexpectedly high level of false positives in the literature and, consequently, unexpectedly low replication rates. In this article, I have offered the alternative view that questionable and other research practices do not usually inflate relevant Type I error rates.

During significance testing, each statistical inference is assigned a nominal Type I error rate or alpha level. Type I error rate inflation occurs if the *actual* Type I error rate for that inference is higher than the *nominal* error rate. The actual Type I error rate can be calculated using the formula $1 - (1 - \alpha)^k$, in which k is the number of significance tests that are used to make the statistical inference. I have argued that the actual Type I error rate is not usually inflated above the nominal rate and that, when it is, the inflation is transparent and easily resolved because k is known by readers. Indeed, k must be known by readers because the researcher must formally associate their statistical inference with one or more significance tests, and k is the number of those tests. A key point here is that k is not the number of tests that a researcher conducted, including those that they conducted and did not report. Instead, k is the number of tests that the researcher formally associates with a statistical inference about a specified null hypothesis.

It is true that some actual Type I error rates may be above a field's conventional alpha level. However, this issue does not necessarily represent Type I error rate inflation. Type I error rate inflation only occurs when the actual Type I error rate for a specified statistical inference is higher than the nominal Type I error rate for that inference, regardless of whether that nominal rate is higher or lower than the conventional level.

It is true that the Type I error rate for a researcher's specified statistical inference about a particular individual or joint null hypothesis may be different to the Type I error rates for other statistical inferences that they could, would, or should have made about other individual or joint null hypotheses or for other statistical inferences that they planned to make or actually made and then either reported or failed to report. It is also true that different researchers may disagree about which are the most appropriate or theoretically relevant statistical inferences or alpha levels in any given research situation. However, none of these points imply that the actual Type I error rate

for a researcher's reported statistical inference is inflated above the alpha level that they have set for that particular inference.

It is true that the actual familywise Type I error rate is always above $\alpha_{Constituent}$. However, researchers can adjust $\alpha_{Constituent}$ to avoid Type I error rate inflation with respect to their statistical inferences about associated joint null and, if they do not adjust $\alpha_{Constituent}$, then the extent of the inflation will be transparent to others and easily resolved. Hence, this potential issue is not problematic.

It is also true that the *studywise* Type I error rate is always above $\alpha_{Constituent}$. However, researchers do not usually make statistical inferences about joint studywise null hypotheses, and so this point is usually irrelevant. Nonetheless, if a studywise error rate does become relevant, then it can be identified and controlled by adjusting $\alpha_{Constituent}$. This adjustment is even possible if researchers take the unusual step of making relatively vague and atheoretical statistical inferences about exploratory joint studywise null hypotheses (e.g., "the study's effect was significant") because, even in this case, they will need to specify the relevant statistical tests that they are using to make this inference.

Finally, it is true that a researcher's probability of incorrectly rejecting a substantive null hypothesis and incorrectly accepting a substantive alternative hypothesis may be greater than their alpha level because this probability may be influenced by theoretical errors as well as Type I errors. However, these theoretical errors cannot be said to inflate Type I error rates because Type I error rates refer to random sampling error per se. They do not account for theoretical errors.

Based on these points, I have argued that the following research practices do not usually inflate relevant Type I error rates: model misspecification, multiple testing, selective inference, forking paths, exploratory analyses, *p*-hacking, optional stopping, double dipping, and HARKing. My view is consistent with a logico-historical approach to hypothesis testing that considers the logical relations between a reported hypothesis and its test results independent from the psychological origin of the hypothesis and results (e.g., researcher bias; Musgrave, 1974, p. 17; Reichenbach, 1938, p. 5; Popper, 1962, p. 140; Popper, 2002 p. 7). Hence, contrary to Mayo's (1996, 2018) error statistical approach, my logical inference-based approach denies that "biasing selection effects" in the testing context inflate Type I error rates. Instead, it offers a more restricted conceptualization of Type I error rates as indicators of the frequency of incorrect decisions in hypothetical and potentially unplanned testing situations in which random sampling error is the only source of decision-making error. This more modest conceptualization may help to provide more realistic expectations about replication rates and, consequently, less concern about replication failures.

Endnotes

1. The semicolon in "Pr(reject H_0 ; H_0 is true)" is used to indicate that "H_0 is true" is a fixed assumption, rather than a random variable that can be true or false. Hence, Pr(reject H_0 ; H_0 is true) is an unconditional probability rather than a conditional probability. In contrast, the vertical bar in "Pr(H_0 is true | reject H_0)" is used to indicate a conditional probability (Mayo & Morey, 2017, Footnote 2; Mayo & Spanos, 2006, p. 331; Wasserman, 2013).
2. Fisher (1930, p. 530) explained that the Bayesian approach of "inverse probability" is applicable when "we know that the population from which our observations were drawn had itself been drawn at random from a super-population of known specification" (e.g., a superpopulation of 200 null populations of which 100 are known to be true and 100 are known to be false). Hence, as Cox (1958) explained, "if the population sampled has itself been selected by a random procedure with known prior probabilities, it seems to be generally agreed that inference should be made using Bayes's theorem" (pp. 357-358).

Acknowledgement

I am grateful to members of the academic community on social media for providing feedback on a previous version of this article. I am also grateful to Sander Greenland for his suggestions.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311*(7003), 485–485. https://doi.org/10.1136/bmj.311.7003.485

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34*(5), 502-508. https://doi.org/10.1111/opo.12131

Barrett, L. F. (2015). Psychology is not in crisis. *The New York Times*. https://www.nytimes.com/ 2015/ 09/ 01/ opinion/ psychology-is-not-in-crisis.html

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology, 54*(4)*,* 343-349. https://doi.org/10.1016/S0895-4356(00)00314-0

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Bergkvist, L. (2020). Preregistration as a way to limit questionable research practice in advertising research. *International Journal of Advertising*, *39*(7), 1172-1180. https://doi.org/10.1080/02650487.2020.1753441

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological Methodology, 25,* 421-458. https://doi.org/10.2307/271073

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, *57*(298), 269-306. https://doi.org/10.1080/01621459.1962.10480660

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, *11*(3), 639-645. https://doi.org/10.2466/pr0.1962.11.3.639

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin, 16*(10), 335-338. https://doi.org/10.1037/h0074554

Brower, D. (1949). The problem of quantification in psychological science. *Psychological Review, 56*(6), 325–333. https://doi.org/10.1037/h0061802

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21*(2), 169-194. https://doi.org/10.1017/S0140525X98001162

Cook, R. J., & Farewell, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159*(1)*,* 93-110. http://doi.org/10.2307/2983471

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Mathematical Statistics, 29*(2), 357-372. http://doi.org/10.1214/aoms/1177706618

Cox, D. R., & Mayo, D. G. (2010). Objectivity and conditionality in frequentist inference. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 276-304). Cambridge University Press. http://doi.org/10.1017/CBO9780511657528

Del Giudice, M., & Gangestad, S. W. (2021). 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, 4*(1). https://doi.org/10.1177/2515245920954925

Dennis, B., Ponciano, J. M., Taper, M. L., & Lele, S. R. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution, 7,* Article 372. https://doi.org/10.3389/fevo.2019.00372

Devezer, B., & Buzbas, E. O. (2023). Rigorous exploration in a model-centric science via epistemic iteration. *Journal of Applied Research in Memory and Cognition, 12*(2), 189–194. https://doi.org/10.1037/mac0000121

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*(3), Article 200805. https://doi.org/10.1098/rsos.200805

Feynman, R. P. (1955). The value of science. *Engineering and Science, 19*(3), 13-15. https://calteches.library.caltech.edu/1575/1/Science.pdf

Firestein, S. (2012). *Ignorance: How it drives science*. Oxford University Press.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222,* 309-368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33,* 503-515. https://doi.org/10.23637/rothamsted.8v61q

Fisher, R. A. (1930). Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 26*(4), 528-535. https://doi.org/10.1017/S0305004100016297

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd.

Fisher, R. A. (1971). *The design of experiments* (9th ed.). Hafner Press.

Fraser, D. A. S. (2019). The *p*-value function and statistical inference. *The American Statistician*, *73*(sup1), 135-147. https://doi.org/10.1080/00031305.2018.1556735

García-Pérez, M. A. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology*, *8*, Article 100120. https://doi.org/10.1016/j.metip.2023.100120

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102,* Article 460. http://doi.org/10.1511/2014.111.460

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates.

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. https://doi.org/10.1177/2515245918771329

Greenland, S. (2017a). For and against methodologies: Some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology, 32,* 3-20. https://doi.org/10.1007/s10654-017-0230-6

Greenland, S. (2017b). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology*, *186*(6), 639-645. https://doi.org/10.1093/aje/kwx259

Greenland, S. (2021). Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. *Paediatric and Perinatal Epidemiology, 35*(1), 8-23. https://doi.org/10.1111/ppe.12711

Greenland, S. (2023). Connecting simple and precise *p*-values to complex and ambiguous realities. *Scandinavian Journal of Statistics, 50*(3), 899-914. https://doi.org/10.1111/sjos.12645

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337-350. https://doi.org/10.1007/s10654-016-0149-3

Hager, W. (2013). The statistical theories of Fisher and of Neyman and Pearson: A methodological perspective. *Theory & Psychology, 23*(2), 251-270. https://doi.org/10.1177/0959354312465483

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *The American Journal of Psychology*, *122*(2), 219–234. https://doi.org/10.2307/27784393

Haig, B. D. (2018). *Method matters in psychology: Essays in applied philosophy of science*. Springer. https://doi.org/10.1007/978-3-030-01051-5

Hancock, G. R., & Klockars, A. J. (1996). The quest for α: Developments in multiple comparison procedures in the quarter century since. *Review of Educational Research*, *66*(3), 269-306. https://doi.org/10.3102/00346543066003269

Hewes, D. E. (2003). Methods as tools: A response to O'Keefe. *Human Communication Research, 29*(3), 448-454. https://doi.org/10.1111/j.1468-2958.2003.tb00847.x

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science, 55*(1), 1-34. https://doi.org/10.1093/bjps/55.1.1

Hochberg, Y., & Tamrane, A. C. (1987). *Multiple comparison procedures.* Wiley. http://doi.org/10.1002/9780470316672

Hurlbert, S. H., & Lombardi, C. M. (2012). Lopsided reasoning on lopsided tests and multiple comparisons. *Australian & New Zealand Journal of Statistics, 54*(1), 23-42. https://doi.org/10.1111/j.1467-842X.2012.00652.x

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. http://doi.org/10.1207/s15327957pspr0203_4

Kim, K., Zakharkin, S. O., Loraine, A., & Allison, D. B. (2004). Picking the most likely candidates for further development: Novel intersection-union tests for addressing multi-component hypotheses in comparative genomics. *Proceedings of the American Statistical Association, ASA Section on ENAR Spring Meeting* (pp. 1396-1402). http:/ / www.uab.edu/ cngi/ pdf/ 2004/JSM%202004%20-IUTs%20Kim%20et%20al.pdf

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, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … & Sowden, W. (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

Kotzen, M. (2013). Multiple studies and evidential defeat. *Noûs, 47*(1), 154-180. http://www.jstor.org/stable/43828821

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12*(5), 535-540. https://doi.org/10.1038/nn.2303

Kuhn, T. S. (1977). *The essential tension: Selected studies in the scientific tradition and change.* The University of Chicago Press.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association, 88*(424), 1242-1249. https://doi.org/10.1080/01621459.1993.10476404

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3), 151-159. https://doi.org/10.1037/h0026141

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese, 190*(6), 975-995. https://doi.org/10.1007/s11229-011-0054-y

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science, 5*(2). https://doi.org/10.1177/25152459221080396

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. *Communication Methods and Measures, 1*(4), 243-265. https://doi.org/10.1080/19312450701641409

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. The University of Chicago Press.

Mayo, D. G. (2014). On the Birnbaum argument for the strong likelihood principle. *Statistical Science, 29,* 227-239. http://doi.org/10.1214/13-STS457

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. http://doi.org/10.1017/9781107286184

Mayo, D. G., & Morey, R. D. (2017). *A poor prognosis for the diagnostic screening critique of statistical tests.* OSFPreprints. https://doi.org/10.31219/osf.io/ps38b

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323-357. https://doi.org/10.1093/bjps/axl003

McShane, B. B., Bradlow, E. T., Lynch, J. G. Jr., & Meyer, R. J. (2023). "Statistical significance" and statistical reporting: Moving beyond binary. *Journal of Marketing, 88*(3), 1-19. http://doi.org/10.1177/00222429231216910

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141. http://www.jstor.org/stable/1448768

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Lawrence Erlbaum Associates.

Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phe-

nomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology, 13*(1), 1-29. https://doi.org/10.1146/annurev.so.13.080187.000245

Molloy, S. F., White, I. R., Nunn, A. J., Hayes, R., Wang, D., & Harrison, T. S. (2022). Multiplicity adjustments in parallel-group multi-arm trials sharing a control group: Clear guidance is needed. *Contemporary Clinical Trials*, *113*, Article 106656. https://doi.org/10.1016/j.cct.2021.106656

Morgan, J. F. (2007). *P* value fetishism and use of the Bonferroni adjustment. *Evidence-Based Mental Health*, *10*, 34-35. http://doi.org/10.1136/ebmh.10.2.34

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, Article 0021. https://doi.org/10.1038/s41562-016-0021

Musgrave, A. (1974). Logical versus historical theories of confirmation. *The British Journal for the Philosophy of Science, 25*(1), 1-23. https://doi.org/10.1093/bjps/25.1.1

Neyman, J. (1950). *First course in probability and statistics.* Henry Holt.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36,* 97–131. https://doi.org/10.1007/BF00485695

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika 20A*(1/2)*,* 175–240. http:/ / doi.org/10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231,* 289-337. https://doi.org/10.1098/rsta.1933.0009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11)*,* 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

Nosek, B. A., & Lakens, D. (2014). Registered reports*. Social Psychology, 45*(3), 137-141. https://doi.org/10.1027/1864-9335/a000192

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596-1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Parker, R. A., & Weir, C. J. (2020). Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification. *Clinical Trials, 17*(5), 562-566. https://doi.org/10.1177/1740774520941419

Parker, R. A., & Weir, C. J. (2022). Multiple secondary outcome analyses: Precise interpretation is important. *Trials*, *23*, Article 27. https://doi.org/10.1186/s13063-021-05975-2

Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., Kelly, C. D., Gurevitch, J., & Nakagawa, S. (2016). Transparency in ecology and evolution: Real problems, real solutions. *Trends in Ecology & Evolution, 31*(9), 711-719. https://doi.org/10.1016/j.tree.2016.07.002

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal, 316,* 1236-1238. https://doi.org/10.1136/bmj.316.7139.1236

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*(1), 159-163. https://doi.org/10.1037/0033-2909.102.1.159

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge.* Basic Books.

Popper, K. R. (2002). *The logic of scientific discovery.* Routledge.

Redish, D. A., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences, 115*(20), 5042–5046. https://doi.org/10.1073/pnas.1806370115

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge.* The University of Chicago Press. https://philarchive.org/archive/REIEAP-2

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science*, *10*(2), 138-157. https://doi.org/10.1214/ss/1177010027

Reid, N., & Cox, D. R. (2015). On some principles of statistical inference. *International Statistical Review, 83*(2)*,* 293-308. http://doi.org/10.1111/insr.12067

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1)*,* 43-46. https://www.jstor.org/stable/20065622

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Lippincott Williams & Wilkins.

Rubin, M. (2017a). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology, 21*(4)*,* 321-329. https://doi.org/10.1037/gpr0000135

Rubin, M. (2017b). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*(3)*,* 269-275. https://doi.org/10.1037/gpr0000123

Rubin, M. (2020a). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology, 16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376

Rubin, M. (2020b). "Repeated sampling from the same population?" A critique of Neyman and Pearson's responses to Fisher. *European Journal for Philosophy of Science, 10,* Article 42, 1-15. https:/ / doi.org/ 10.1007/ s13194-020-00309-6

Rubin, M. (2021a). There's no need to lower the significance threshold when conducting single tests of multiple individual hypotheses. *Academia Letters,* Article 610. https://doi.org/10.20935/AL610

Rubin, M. (2021b). What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese, 198,* 5809–5834. https://doi.org/10.1007/s11229-019-02433-0

Rubin, M. (2021c). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese, 199,* 10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Rubin, M. (2022). The costs of HARKing. *British Journal for the Philosophy of Science, 73*(2), 535-560. https:/ / doi.org/ 10.1093/bjps/axz050

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology, 10,* Article 100140. https://doi.org/10.1016/j.metip.2024.100140

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology, 37*(8), 2019-2047. https://doi.org/10.1080/09515089.2022.2113771

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology, 142*(9), 904-908. https://doi.org/10.1093/oxfordjournals.aje.a117737

Schulz, K. F., & Grimes, D. A. (2005). Multiplicity in randomised trials I: Endpoints and treatments. *The Lancet, 365*(9470)*,* 1591-1595. https:/ / doi.org/ 10.1016/ S0140-6736(05)66461-6

Senn, S. (2007). *Statistical issues in drug development* (2nd ed.)*.* Wiley.

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*(11), 1208-1214. https://doi.org/10.1038/s41562-020-0912-z

Sinclair, J., Taylor, P. J., & Hobbs, S. J. (2013). Alpha level adjustments for multiple dependent variable analyses and their applicability—A review. *International Journal of Sports Science Engineering, 7*(1)*,* 17-20.

Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. *Optimality, 49,* 98-119. https://doi.org/10.1214/074921706000000419

Spanos, A. (2010). Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, *158*(2), 204-220. https://doi.org/10.1016/j.jeconom.2010.01.011

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on*

Psychological Science, 11(5), 702-712. https://doi.org/10.1177/1745691616658637

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), Article 220346. https://doi.org/10.1098/rsos.220346

Syrjänen, P. (2023). Novel prediction and the problem of low-quality accommodation. *Synthese*, *202*, Article 182. https://doi.org/10.1007/s11229-023-04400-2

Szollosi, A., & Donkin, C. (2021). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science, 16*(4), 717-724. https://doi.org/10.1177/1745691620966796

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences, 112*(25), 7629-7634. https://doi.org/10.1073/pnas.1507583112

Tukey, J. W. (1953). *The problem of multiple comparisons.* Princeton University.

Turkheimer, F. E., Aston, J. A., & Cunningham, V. J. (2004). On the logic of hypothesis testing in functional imaging. *European Journal of Nuclear Medicine and Molecular Imaging, 31,* 725-732. https://doi.org/10.1007/s00259-003-1387-7

Uygun-Tunç, D., & Tunç, M. N. (2023). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. *Meta-Psychology, 7*. https://doi.org/10.15626/MP.2021.2756

Veazie, P. J. (2006). When to combine hypotheses and adjust for multiple tests. *Health Services Research, 41*(3 pt 1), 804-818. http://doi.org/10.1111/j.1475-6773.2006.00512.x

Venn, J. (1876). *The logic of chance* (2nd ed.). Macmillan and Co.

Wagenmakers, E. J. (2016, September 1). Statistical tools such as p-values and confidence intervals are meaningful only for strictly confirmatory analyses. In turn, preregistration is one. [Comment on the blog post "Why preregistration makes me nervous"]. *Psychological Science.* https://www.psychologicalscience.org/ observer/ why-preregistration-makes-me-nervous#comment-7860633

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. https://doi.org/10.1177/1745691612463078

Wasserman, L. (2013, March 14). Double misunderstandings about *p*-values. *Normal Deviate.* https:/ / normaldeviate.wordpress.com/2013/03/14/double-misunderstandings-about-p-values/

Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin, 59*(4), 296-300. https://doi.org/10.1037/h0040447

Worrall, J. (2010). Theory confirmation and novel evidence. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 125-169). Cambridge University Press. http://doi.org/10.1017/CBO9780511657528