Special issue

Vol 4 N°1

ISSN 2667-1204

Journal of Trial and Error

Consequences of the Scientific Reform Movement

Special Issue

Consequences of the Science Reform Movement

Volume 4
Issue 1
May 24, 2024

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

Contents

Reflections on the Unintended Consequences of the Science Reform Movement

Sarahanne Field[1], Noah van Dongen[2], Leo Tiokhin[3]

[1]University of Groningen
[2]University of Amsterdam
[3]IG&H: Utrecht

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
February 20, 2024
Accepted
February 27, 2024
Published
May 24, 2024
Issued
May 24, 2024

Correspondence
University of Groningen
field@trialanderror.org

The scientific community has entered a challenging era, as originally noted by Wagenmakers (**2012**). High-profile instances of fraud, failures to replicate foundational studies in psychology, and admissions of research misconduct (e.g., John et al., **2012**) cast a shadow over the field of psychology initially, and the broader enterprise of science over the decade that has passed. In response to these concerns, a movement aimed at reforming scientific practices has emerged (Field, **2022**; Munafò et al., **2017**; Spellman et al., **2018**). This movement has introduced various initiatives to enhance research methods, reduce misconduct, and increase transparency (see van Ravenzwaaij et al., **2023**, for an overview). Direct replications and articles reporting null results and errors (such as the *Journal of Trial and Error*; see Devine et al., **2020**) have become more accessible for publication. Concepts like preregistration and registered reports have gained significant popularity. Additionally, various aspects of science are now more "open," encompassing preprints, open-access publications, open peer review, and openly accessible data and code.

While these initiatives offer immediate benefits to both the scientific community and researchers, one might argue that the scientific reform movement is still in its early stages; the long-term effects of many interventions have yet to be assessed. Furthermore, the system of academic science is a complex one, and the downstream consequences of modifying complex systems are notoriously difficult to anticipate. Notably, Devezer et al. (**2020**, p. 2) have raised concerns regarding the "little evidentiary backing" and lack of a framework for assessing the validity and efficacy of reform policies. Ioannidis (**2014**) and Tiokhin et al. (**2021**) share similar apprehensions about the potential unintended consequences of well-intentioned reforms. Additionally, adopting reform practices might carry negative consequences for students and early career researchers, who are often beholden to supervisors and structures with their own dependence on traditional academia (Field, **2023**).

Personal Context

Field, van Dongen, and Tiokhin arrived at an articulation of these concerns individually, yet at a similar time. In the early 2020's, Tiokhin came to van Dongen about frustrations that people were talking about positive primary effects of the science reform movement, without considering potential second and even third order effects. They agreed that unintended consequences drive change just as much as their primary catalyst. Together, they decided they were going to write a book addressing unintended consequences in the reform movement, and that they would "make waves." Although the book has yet to materialize, their discussion produced a workshop for the Center for Unusual Collaborations in Amsterdam in 2021 and a session at the virtual Metascience conference of 2021.

Field's own concerns about the unintended consequences of science reform began in 2020, during her PhD which centred on the "science reform community" and its practices. While she was supportive of the changes the community had been ushering in, she was concerned about the lack of reflexivity that she saw in some of the community's members and the discomfort some members clearly experienced when questioned about how certain initiatives and practices would affect other aspects of the academic enterprise. Field is an editor at the Journal of Trial & Error and had pitched a special issue concept to the JOTE team in mid-2021, but was

unsure if she had the time and energy to bring it to fruition on her own. When she attended the 2021 Metascience session run by Tiokhin and van Dongen (among others), Field realized she could harness some of the fine brain power of her metascientific colleagues and reached out to Tiokhin and van Dongen in the hope that they would help her with the project as guest editors.

To address the challenges we identified, and to ensure that reform efforts remain credible and self-reflective, we argue that the scientific reform movement must continually interrogate its practices and proposed initiatives and remain vigilant to the unintended consequences that may arise. In this context, this special issue aims to inspire scholars to critically reflect on the trajectory of the scientific reform movement. We now outline the eight articles of the special issue, each of which deal with a different facet of the science reform movement.

The Contributions

Questionable Metascience Practices by Rubin (**2023**) considers a parallel concept to "questionable research practices" called questionable metascience practices (QMPs). Rubin explores 10 QMPs, including ignoring or rejecting criticism of one's own proposed reforms, and the overemphasis on replication. He urges metascientists to "reflect on the ways in which they (a) handle criticism, (b) conceptualize replication, (c) consider researcher bias, (d) avoid sweeping generalizations, and (e) acknowledge the diversity and pluralism of science."

Reflections on Preregistration: Core Criteria, Badges, Complementary Workflows (**2023**) by Thibault, Pennington and Munafò, explores the issues surrounding preregistration, a key pillar of the reform movement. While preregistration promises transparency in theory, challenges emerge in practice, especially in the case of clinical research trials, such as the absence of analysis plans and issues with the awarding of preregistration badges. Their recommendations to the reform and clinical research communities in light of these concerns are straightforward: They propose the consideration of "parallel initiatives to simplify and standardize preregistration (e.g., adopt itemized core preregistration criteria), and to leverage complementary workflows that necessitate open research practices."

Thomas Hostler's article *The Invisible Workload of Open Research* (**2023**) interrogates the burden carried by researchers who practice open science, with concrete examples (a perspective that lends gravitas to existing abstract discourse). Hostler discusses how the adoption of open research practices may exacerbate stress, burnout, and workload pressures in academia, opening a Pandora's box that some science reformers prefer to keep closed. He advocates for awareness from the science reform community (or communities), stating that "It is neither the specific role nor within the capability of open research advocates to tackle the root causes of workload issues, but they must be aware of the potential implications of their calls for systemic changes in incentives for open research."

A team of researchers from the **Quala Lab**, led by Steltenpohl, contributed an article on qualitative open science: *Rethinking Transparency and Rigor from a Qualitative Open Science Perspective* (**2023**). They explain that imposing rigid quantitative standards on all research can have unintended negative consequences for many individuals and groups in the science reform movement, cautioning against the perspective that transparency is a "one-size-fits-all" concept. Qualitative researchers have unique considerations, such as reflexivity and positionality statements. Looking ahead, they argue that through "...expanding open science guidelines to leverage a broader array of rigor and transparency-promoting practices (e.g., reflexivity), we can truly begin to advance practices."

In *A Manifesto for Rewarding and Recognizing Team Infrastructure Roles*, Bennett and colleagues (**2023**) consider team science in the context of the reform movement. They talk about professional team infrastructure roles (such as lab technicians, project managers and data stewards), which involve the crucial activity of supporting research while at the same time do not have responsibility of leading team projects, and therefore typically miss out on reward and recognition, especially under current systems and structures. They underscore the need for fairness in how such personnel are treated, and bring the issue to light using three case studies. "Acknowledging the contri-

butions of all research roles," they argue, "will help retain skill and expertise, and lead to collaborative research ecosystems that are well-positioned to address complex research challenges."

Buzbas and Devezer's (**2023**) article *Tension Between Theory and Practice of Replication* scrutinizes the popular yet somewhat divisive topic of replication. They describe a mismatch between the push for more replication studies and concern about irreproducibility with iterative and slow (but crucial) theoretical developments. They advance theoretical considerations of "non-exact" replication studies and meta-hypothesis testing in multi-lab replications and warn of the problems that enacting reforms without robust theoretical foundations can pose for the reform movement. They emphasize the need for a theoretical framework of metascience to guide science reform, especially in the context of large-scale replication studies. "Theoretical work is still in its early stages of development and needs to continue," they explain, while at the same time noting that "another major challenge arises for the next generation of reform: How do we bridge the gap between theory and practice?"

Finally, in *Reputation without practice? A Dynamic Computational Model of the Unintended Consequences of Open Scientist Reputations*, Linde et al. (**2024**) delve into the dynamics of open science advocacy, exploring how being an advocate for open science can affect academic careers. It introduces a dynamic model to examine two types of open science behaviour (practicing open science and/or advocating open science) and how they can impact the career progression of academic researchers. They found that groups that both practice and advocate open science will be dominant in a scientific community that values open science, and that advocating OS may not have the same advantages as practicing OS. They write: "These results are encouraging to those who feel practicing open science 'is not worth it': in addition to benefits to science at large, our results suggest engaging with OS benefits the individual researcher as well."

Conclusion

With the objective of stimulating a discourse in the realm of scientific reform, together with the structural backing of the Center of Trial and Error and the input of several metascience researchers, we have co-produced a special issue that focuses on "second order" effects or "second-generation" challenges—issues that may arise during or as a consequence of addressing primary reform concerns. The peer-reviewed contributions that appear in the special issue deal with a range of issues: problematic reformer behaviours, the mismatch between preregistration theory and practice in clinical trial research, the hidden workload of open science, open science challenges for qualitative research, neglected yet important roles in team science, the tension between theory and practice of replication, the impact of open science advocacy and practice on researchers' careers, and the role of moral positions on core values in the progress of the science reform movement.

The complexities of scientific reform require thoughtful, well-rounded solutions built on inclusive discussions, and the contributions in this special issue provide a rich tapestry of perspectives to guide our way forward. As we continue to refine our scientific practices, we must ensure that our shared journey towards scientific reform remains reflexive, critical, and balanced.

References

Bennett, A., Garside, D., Praag, C. G., Hostler, T. J., Garcia, I. K., Plomp, E., Schettino, A., Teplitzky, S., & Ye, H. (2023). A manifesto for rewarding and recognizing team infrastructure roles. *Journal of Trial & Error*. https://doi.org/10.36850/mr8 (see p. 2).

Buzbas, E. O., & Devezer, B. (2023). Tension between theory and practice of replication. *Journal of Trial & Error*. https://doi.org/10.36850/mr9 (see p. 3).

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2020). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), Article 200805. https://doi.org/10.1098/rsos.200805 (see p. 1).

Devine, S., Bautista-Perpinya, M., Delrue, V., Gaillard, S., Jorna, T., van der Meer, 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 (see p. 1).

Field, S. M. (2022). *Charting the constellation of science reform.* [Doctoral dissertation, University of Groningen]. Pure. https://doi.org/10.33612/diss.229114775 (see p. 1).

Field, S. M. (2023). Risk reform, or remain within the academic monolith? *The Psychologist*, *36*, 45–47. https://www.bps.org.uk/psychologist/risk-reform-or-remain-within-academic-monolith (see p. 1).

Hostler, T. J. (2023). The invisible workload of open research. *Journal of Trial & Error*. https://doi.org/10.36850/mr5 (see p. 2).

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, *11*(10), Article 1001747. https://doi.org/10.1371/journal.pmed.1001747 (see p. 1).

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 (see p. 1).

Linde, M., Pittelkow, M.-M., Schwarzback, N., & van Ravenzwaaij, D. (2024). Reputation without practice? A dynamic computational model of the unintended consequences of open scientist reputations. *Journal of Trial and Error*. https://doi.org/10.36850/mr10 (see p. 3).

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., & 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. 1).

Rubin, M. (2023). Questionable metascience practices. *Journal of Trial & Error*. https://doi.org/10.36850/mr4 (see p. 2).

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open Science. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *5*, 1–47. https://doi.org/doi.org/10.1002/9781119170174 (see p. 1).

Steltenpohl, C. N., Lustick, H., Meyer, M. S., Lee, L. E., Stegenga, S. M., Reyes, L. S., & Renbarger, R. L. (2023). Rethinking transparency and rigor from a qualitative open science perspective. *Journal of Trial & Error*. https://doi.org/10.36850/mr7 (see p. 2).

Thibault, R. T., Pennington, C. R., & Munafò, M. R. (2023). Reflections on preregistration: Core criteria, badges, complementary workflows. *Journal of Trial & Error*. https://doi.org/10.36850/mr6 (see p. 2).

Tiokhin, L., Panchanathan, K., Lakens, D., Vazire, S., Morgan, T., & Zollman, K. (2021). Honest signaling in academic publishing. *PLOS ONE*, *16*(2), Article 0246675. https://doi.org/10.1371/journal.pone.0246675 (see p. 1).

van Ravenzwaaij, D., Bakker, M., Heesen, R., Romero, F., van Dongen, N., Crüwell, S., Field, S. M., Held, L., Munafò, M. R., Pittelkow, M. M., Tiokhin, L., Traag, V. A., van den Akker, O. R., van 't Veer, A. E., & Wagenmakers, E. J. (2023). Perspectives on scientific error. *Royal Society Open Science*, *10*(7), Article 230448. https://doi.org/10.1098/rsos.230448 (see p. 1).

Wagenmakers, E. J. (2012). A year of horrors. *De Psychonoom*, *27*, 12–13 (see p. 1).

Questionable Metascience Practices

Mark Rubin [1]

Metascientists have studied *questionable research practices* in science. The present article considers the parallel concept of *questionable metascience practices* (QMPs). A QMP is a research practice, assumption, or perspective that has been questioned by several commentators as being potentially problematic for metascience and/or the science reform movement. The present article reviews ten QMPs that relate to criticism, replication, bias, generalization, and the characterization of science. Specifically, the following QMPs are considered: (1) rejecting or ignoring self-criticism; (2) a fast 'n' bropen scientific criticism style; (3) overplaying the role of replication in science; (4) assuming a replication rate is "too low" without specifying an "acceptable" rate; (5) an unacknowledged metabias towards explaining the replication crisis in terms of researcher bias; (6) assuming that researcher bias can be reduced; (7) devaluing exploratory results as being more "tentative" than confirmatory results; (8) presuming that questionable research practices are problematic research practices; (9) focusing on knowledge accumulation; and (10) focusing on specific scientific methods. It is stressed that only *some* metascientists engage in *some* QMPs *some* of the time, and that these QMPs may not *always* be problematic. Research is required to estimate the prevalence and impact of QMPs. In the meantime, QMPs should be viewed as invitations to ask questions about how we go about doing better metascience.

Keywords *metascience, open science, questionable research practices, replication crisis, science reform*

[1] Durham University

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
September 8, 2022
Accepted
December 13, 2022
Published
April 24, 2023
Issued
May 24, 2024

Correspondence
Durham University
mark-rubin@outlook.com

In **2011**, Simmons et al. demonstrated that researchers can present "anything as significant" (p. 1359) by conducting numerous analyses (e.g., using different outcome variables, sample sizes, and/or covariates) and then selectively reporting only those analyses that yield significant results. A year later, John et al.'s (**2012**) published the results of a survey which purported to show that *questionable research practices* (QRPs), such as HARKing and *p*-hacking, are prevalent among psychologists. A few years later, an attempt to replicate 100 psychology studies found that only 39% of effects were rated as replicable (Open Science Collaboration, **2015**).

In light of this and other work, some metascientists have concluded that QRPs play a significant role in increasing the publication of "false positive" results and, therefore, lowering replication rates (e.g., Bishop, **2019**; Bishop, **2020**; Munafò et al., **2017**; Nosek et al., **2012**; Open Science Collaboration, **2015**; Schimmack, **2020**; Spellman et al., **2018**). Partly in response, science reformers have advocated new "open science" research practices that are intended to reveal and reduce QRPs (e.g., preregistered research plans, publicly accessible research data and materials Munafò et al., **2017**).

In the present article, I consider questionable research practices in the field of metascience. A *questionable metascience practice* (QMP) is a research practice, assumption, or perspective that has been questioned by several commentators as being potentially problematic for metascience and/or the science reform movement. I outline ten QMPs that are grouped into the five broad categories of (a) criticism, (b) replication, (c) bias, (d) generalization, and (e) science characterization.

Please note that I have not provided an exhaustive list of QMPs (for some additional QMPs, please see Devezer et al., **2021**, p. 2). In addition, unlike John et al.'s (**2012**) study of QRPs, I have not attempted to estimate the

prevalence of the QMPs that I consider. It is possible that only a *few* metascientists have engaged in the QMPs, and that they have engaged in only a *few* QMPs a *few* times. Nonetheless, under some circumstances, a few low frequency QMPs may be quite influential and problematic, especially when they are undertaken by prominent metascientists who are regarded as leaders and role models in the field. Hence, it is worthwhile considering QMPs even if they have a low prevalence.

Finally, in my view, QMPs are not always problematic. They are merely "questionable" in the sense that they warrant questioning before a conclusion is reached about whether they are problematic in any given situation. Hence, my aim is not to cast aspersions on the field of metascience but, instead, to encourage a deeper consideration of its more questionable research practices, assumptions, and perspectives.[1]

Criticism-Related QMPs

Rejecting or Ignoring Self-Criticism

As several commentators have noted, some metascientists react particularly negatively and defensively towards criticisms of their proposed science reforms (Bastian, **2021**; Gervais, **2021**; Malich & Rehmann-Sutter, **2022**, p. 5; Walkup, **2021**, p. 132). For example, as Flis (**2022**) explained, there was a rather extreme negative reaction on social media to an article by Szollosi et al. (**2020**) that criticized the open science practice of preregistration. Flis suggested that this highly negative reaction may have represented a defensive response that was learned during metascientists' interactions with so-called "status-quoers" who questioned the reality of the replication crisis and opposed the need for science reform. In other words, some first-generation metascientists and reformers may have adopted a particularly negative reaction to self-criticism because they perceived it to be a challenge to their raison d'être.

Instead of rejecting self-criticism, some metascientists may simply ignore it, especially in the more authoritative space of the published literature. For example, as of February 2023, 228 articles have cited a pro-preregistration article by Nosek et al. (**2019**) that was published around the same time and in the same journal as Szollosi et al.'s (**2020**) critical article. However, only 17% of these 228 articles (k = 39) have also cited Szollosi et al. (To identify these 39 articles, I clicked on "cited by" in Google Scholar for the Nosek et al. article and then selected "search within citing articles" and searched for "Is preregistration worthwhile?"). This low co-citation rate may reflect a citation bias against an article that is critical of a prominent science reform (for another example of potential citation bias, please see Flis, **2022**, p. 6). This type of citation bias creates an illusion of consensus in the literature, and it may obstruct the motive for theory improvement by giving the impression that current theories are adequate and undisputed (see also Bishop, **2020**; Hoekstra & Vazire, **2021**, p. 1604). Hence, "failing to cite publications that contradict your beliefs" is regarded as a QRP (Allum et al., **2023**, p. 8). To prevent this QRP from becoming a QMP, metascientists should encourage self-criticism, cite their critics' work, and respond in a thoughtful manner (Altenmüller et al., **2021**; Gervais, **2021**, p. 828; Haig, **2022**, p. 235; Hoekstra & Vazire, **2021**, p. 1604). To be clear, metascientists do not always need to concede to their critics' arguments. However, they do need to engage with those arguments publicly, formally, and carefully (see also Longino, **1990**).

Fast 'n' Bropen Criticism

Concerns have also been raised about the style and tone of some metascientists' interactions with scientists, especially on social media (e.g., Fiske, **2016**; Hamlin, **2017**, p. 692; Pownall & Hoerst, **2022**; Whitaker & Guest, **2020**). For example, Whitaker and Guest (**2020**) coined the term *bropenscience* to refer to a dismissive, mocking, school-yard style of scientific criticism

[1] Researcher positionality statement: I identify as a White, Western, heterosexual, middle-class, cisgender man. My primary field of research is social psychology. However, I have recently published several articles in the field of metascience. Here, I have adopted a relatively nuanced and contextualist approach to issues such as HARKing, *p*-hacking, and multiple testing (e.g., Rubin, **2017a**; Rubin, **2017b**; Rubin, **2021b**; Rubin, **2022**). I have also adopted a relatively critical approach to some science reforms, such as (a) preregistration (e.g., Rubin, **2020**; Rubin, **2022**) and (b) greater adherence to Neyman-Pearson hypothesis testing (e.g., Rubin, **2021a**). My current philosophy of science is closest to perspectival realism (Crețu, **2019**; Giere, **2006**; Massimi, **2022**). For more information about my work, please see https://sites.google.com/site/markrubinsocialpsychresearch/

that some metascientists sometimes use on social media (e.g., Anonymous, **2021**; see also Derksen & Field, **2022**; Pownall et al., **2021**, pp. 529-530). Similarly, Pownall (**2022**) noted that, in contrast to the appeal for more thoughtful and "slower" science, there is a "growing culture of fast, hostile, and superficial critiques of research" on social media.[2]

Although a *fast 'n' bropen* criticism style may be used rarely and by few metascientists, it can be problematic if it is used by relatively prominent metascientists who are regarded as being representative of the field. In particular, it may (a) distract from and/or deter legitimate criticism, (b) cause scientists to feel personally attacked and/or excluded (e.g., Derksen & Field, **2022**; Hamlin, **2017**, p. 692; Pownall et al., **2021**), (c) damage the reputation of metascience, and/or (d) reduce the uptake of beneficial science reforms (Gervais, **2021**). Metascientists should undertake thoughtful, "critical evaluation with civility and mutual respect" (Society for the Improvement of Psychological Science, **2022**).

Replication-Related QMPs

Overplaying Replication

Some metascientists assume that direct replications are a method for assessing the "truth" of a claim or effect. For example, Nosek et al. (**2012**, p. 617) stated that "replication is a means of increasing the confidence in the truth value of a claim"; Nelson et al. (**2018**, p. 520) stated that, "to a scientist, a true effect is one that replicates under specifiable conditions"; and Simmons et al. (**2021**, p. 153) stated that "many published findings do not replicate under specifiable conditions and so are, by the standards of science, untrue" (for further examples, see Devezer et al., **2021**, pp. 6-8). Some metascientists also regard replication as an essential and defining aspect of science. For example, the Open Science Collaboration (**2015**, p. 1) described reproducibility as "a defining feature of science," and Zwaan et al. (**2018**, p. 13) explained that replication is "an essential component of science…a foundational principle of the scientific method" (see also Asendorpf et al., **2013**, p. 108; Chambers, **2017**, p. 48; Nosek et al., **2012**, p. 618; for further examples, see Drummond, **2019**, p. 64; Haig, **2022**, p. 226; Maxwell et al., **2015**, p. 487). In response, critics have argued that these sorts of statements overplay the role of replication in science (De Boeck & Jeon, **2018**; Devezer et al., **2019**; Devezer et al., **2021**; Feest, **2019**; Greenfield, **2017**; Guttinger, **2020**; Haig, **2022**; Iso-Ahola, **2020**; Leonelli, **2018**; Norton, **2015**).

Replication does not indicate whether research claims or findings are true. Replicable results may be "false" due to model misspecification, reliable but invalid measures, or overly liberal evidence thresholds, and "true" results may be nonreplicable due to model misspecification, unreliable methods, or irreversible changes in the population over time (Bak-Coleman et al., **2022**; Buzbas et al., **2023**; De Boeck & Jeon, **2018**; Devezer et al., **2019**; Devezer et al., **2021**; Errington, Mathur, et al., **2021**; Guttinger, **2020**; Iso-Ahola, **2020**; Norton, **2015**; Nosek et al., **2022**, p. 739; Rubin, **2021a**; D. J. Stanley & Spence, **2014**). Furthermore, replication is not an essential component of science. Scientists often use other methods to demonstrate the reliability of their results, such as robustness analyses (Haig, **2022**; Leonelli, **2018**). Alternatively, they may provide a repeat demonstration of the existence of a phenomenon within the same study using a different set of variables that are nonetheless representative of the theoretical constructs that were used in the original demonstration.

Certainly, replication is important in some areas of science. However, it is a QMP to overplay replication as an "essential" aspect of science that indexes the "truth" of findings (Devezer et al., **2021**, p. 10).

Unspecified Replication Rate Targets

Some metascientists claim that replication rates need to be improved. For example, the Open Science Collaboration (**2015**, p. 7) concluded that "there is room to improve reproducibility in psychology," and Munafò et al. (**2017**, p. 1) explained that "data from many fields suggests reproducibility is lower than is desirable." However, it is unclear how replication rates can be judged to be "low" and in need of improvement in the absence of clear

[2] Some critics of the science reform movement also have problematic communication styles at times (for some examples, see Holcombe, **2021**). Nonetheless, two wrongs don't make a right!

targets for "acceptable" replication rates. Logically, this reasoning represents an incomplete comparison.

In their recent review,Nosek et al. (**2022**) found that 64% of 307 replications reported statistically significant evidence in the same direction as the original studies. Is this replication rate "too low" or is it "acceptable?" Nosek et al. were unsure, asking: "what degree of replicability should be expected?" (p. 730) and "what is the optimal replicability rate at different stages of research maturity?" (p. 738). They suggested that these questions should be addressed in future metascience research (see also Open Science Collaboration, **2015**, p. 7). However, the deferral of this question implies that metascientists are trying to solve a problem that they are not yet sure exists. After all, future research may reveal that current replication rates are "acceptable" (Bird, **2020**; Freiling et al., **2021**, p. 692; Guttinger, **2020**, p. 8; Lewandowsky & Oberauer, **2020**). Alternatively, the meaningfulness of quantifying replication rates may be called into question (Buzbas et al., **2023**; Rubin, **2021a**).

In the absence of clear targets for "acceptable" replication rates, it is not surprising that several commentators have questioned whether current replication rates are at "crisis��" levels (e.g., Barrett, **2015**; Bird, **2020**; Buzbas et al., **2023**; Fanelli, **2018**; Firestein, **2016**; Freiling et al., **2021**; Haig, **2022**; Maxwell et al., **2015**; Morawski, **2019**; Shrout & Rodgers, **2018**; Stroebe & Strack, **2014**; Wood & Wilson, **2019**). Certainly, claiming that a replication rate is "too low" without specifying an "acceptable" replication rate represents a QMP.

Bias-Related QMPs

Metabias

As several commentators have observed, contemporary metascientists tend to be concerned with how bias and motivated reasoning influence scientists' methods, analyses, and interpretations (Field & Derksen, **2021**; Flis, **2019**; Morawski, **2019**; Morawski, **2022**; Peterson & Panofsky, **2020**, p. 7; for examples, see Bishop, **2020**; Chambers, **2017**, chapter 1; Chambers & Tzavella, **2022**; Hardwicke & Wagenmakers, **2023**; Ioannidis et al., **2014**; Munafò et al., **2017**; Nosek et al., **2012**; Simmons et al., **2021**, p. 153). Indeed, Morawski (**2022**) has suggested that metascientists may be biased towards explaining the replication crisis in terms of researcher bias because they are overrepresented by psychologists (Moody et al., **2022**; see also Flis, **2019**; Malich & Rehmann-Sutter, **2022**), who tend to be familiar with cognitive and motivational biases (i.e., a type of availability heuristic bias). Consistent with Morawski's interpretation, it is interesting to note that psychologists' metabias may also explain their emphasis on researcher bias during the 1960s-1970s crisis of confidence in social psychology (Peterson & Panofsky, **2021**, p. 600; Rosnow, **1983**). In this previous crisis, psychologists were concerned about researchers biasing the behavior of their participants (e.g., experimenter expectancy effects). In the current replication crisis, they are more concerned about researchers biasing their methods and analyses.

To be consistent with their concerns about researcher bias, metascientists should acknowledge their own *metabias* towards explanations of the replication crisis that refer to researcher bias. There are multiple mutually compatible explanations for failed replications that do not refer to researcher bias, including data errors, fraud, a base rate fallacy, low power, unreliable measurement, poor validity, hidden moderators, and heterogenous effects (e.g., Bird, **2020**; De Boeck & Jeon, **2018**; Fabrigar et al., **2020**; Maxwell et al., **2015**; Rubin, **2021a**; D. J. Stanley & Spence, **2014**). Researcher bias and associated QRPs represent only one potential explanation, yet they have been given a disproportionate amount of attention in explanations of, and solutions to, the replication crisis (e.g., Hardwicke & Wagenmakers, **2023**; Munafò et al., **2017**; Schimmack, **2020**, p. 372). Focusing on researcher bias at the expense of other viable explanations represents a form of causal reductionism (Devezer et al., **2019**, p. 17), and an acknowledgement of metabias may help to produce a more balanced and comprehensive multicausal account of the replication crisis.

The Bias Reduction Assumption

Some metascientists believe that preregistration and registered reports reduce researcher bias. For example, Hardwicke and Wagenmakers (**2023**, p. 15) explained that "preregistration…reduces the risk of bias by encouraging outcome-independent decision-making"; Vazire et al. (**2022**, p. 166) explained that "the aim of the Registered Report format is to reduce bias by eliminating many of the avenues for undisclosed flexibility in research"; and Chambers (**2018**) described "Registered Reports as a vaccine against research bias" (see also Chambers & Tzavella, **2022**, p. 32; Scheel et al., **2021**, p. 2; for commentary, see Field & Derksen, **2021**). There are three problems with this claim.

First, researcher bias influences not only the post hoc selection of hypotheses, data, analyses, and results (i.e., *selective reporting*), but also the a priori selection of hypotheses, methods, analyses, evidence thresholds, and interpretations (i.e., *selective questioning*; Rubin & Donkin, **2022**), and considering selective reporting without also considering selective questioning may lead to a biased evaluation of researcher bias. For example, preregistering the number of times that a researcher will toss a coin may help to identify and reduce any selective reporting of their results (e.g., only reporting when the coin lands heads and not when it lands tails). However, the reduction of this selective reporting will not reduce researcher bias if the researcher's preregistered decision rule is "heads I win, tails you lose!" As Clark et al. (**2022**) put it, "the dice have often been loaded before pre-registration" (p. 13, see also Dellsén, **2020**; Jamieson et al., **2023**). Consequently, it is a QMP to assume that a preregistered study is less biased than a non-preregistered study, because selective questioning in the preregistered study may be more problematic than selective reporting in the non-preregistered study (for similar concerns, see Devezer et al., **2021**, p. 16; Freiling et al., **2021**, p. 698; Jamieson et al., **2023**; McDermott, **2022**; Oberauer, **2019**; Pham & Oh, **2021**, p. 167; Rubin & Donkin, **2022**; Szollosi et al., **2020**, p. 95; Wiggins & Christopherson, **2019**, p. 212).

Second, it might be argued that preregistration reduces selective reporting when all other variables are held constant, including variables associated with selective questioning. However, even if, ceteris paribus, preregistration reduces selective reporting, it may also *increase* other types of researcher bias, such as (a) the *researcher commitment bias* (sticking with a planned research approach, even when it is inappropriate), (b) the *researcher prophecy bias* (misattributing a researcher's lucky, atheoretical prophecy to a theory's predictive power), and (c) a bias towards committing data fraud (for a discussion, please see Rubin & Donkin, **2022**). Again, it is a QMP to consider bias reduction in terms of selective reporting per se and ignore other forms of researcher bias.

Finally, and more generally, the metascientific concept of "bias reduction" assumes that researchers can get closer to an "unbiased" evaluation, which smacks of *naïve objectivism, naïve empiricism, naïve realism,* and *value-free science* (Field & Derksen, **2021**; Morawski, **2019**, p. 228; Reiss & Sprenger, **2020**; Strong, **1991**; van Dijk, **2021**; Wiggins & Christopherson, **2019**). According to these philosophical positions, scientists can observe an immutable reality directly and in an unbiased and objective manner. However, contrary to these positions, research is always undertaken from one perspective or another, so it is always "biased" from one perspective or another, and what are seen as decreases in bias from one perspective may be regarded as increases in bias from another. Consequently, a more tenable position is that open science practices help to *reveal different perspectives* rather than to *reduce bias* (Field & Derksen, **2021**; Grossmann, **2021**; Jamieson et al., **2023**; Pownall, **2022**). For example, a robustness or multiverse analysis allows readers to understand how different analytical approaches produce or "enact" different results (Del Giudice & Gangestad, **2021**; Morey, **2019**; Rubin, **2020**; for a discussion of the "enactment" perspective, see Derksen & Morawski, **2022**). In addition, researcher positionality statements can reveal researchers' perspectives rather than reduce their biases (Jamieson et al., **2023**).

Sweeping Generalization QMPs

Devaluing Exploratory Hypothesis Tests

Some metascientists devalue unplanned exploratory tests of post hoc hypotheses relative to preregistered confirmatory tests of a priori

hypotheses, even when the exploratory tests are correctly reported as being exploratory. For example, relative to the results of confirmatory hypothesis tests, the results of exploratory tests are supposed to have a "higher risk of bias" (Hardwicke & Wagenmakers, **2023**, p. 19) and entail greater "uncertainty" (Nosek et al., **2018**, p. 2601), which makes their associated conclusions more "tentative" (Errington, Denis, et al., **2021**, p. 19; Ioannidis et al., **2014**, p. 238; Nelson et al., **2018**, p. 519; Nosek & Lakens, **2014**, p. 138; Simmons et al., **2021**, p. 154). Consequently, "confirmatory analyses…have much greater evidential impact than exploratory analyses" (Wagenmakers, **2012**, p. 13), and research conclusions should be "appropriately weighted in favour of the confirmatory outcomes" (Chambers & Tzavella, **2022**, p. 36). There are two problems with this perspective.

First, critics have argued that the distinction between exploratory and confirmatory hypothesis tests is unclear and irrelevant, both from a statistical perspective (Devezer et al., **2021**; Rubin, **2020**; Rubin, **2021b**) and from a philosophical standpoint (Rubin, **2020**; Rubin, **2022**; Rubin & Donkin, **2022**; Szollosi & Donkin, **2021**). In particular, it has been shown that the "double use" of the same data to (a) generate hypotheses and then (b) test those hypotheses is not necessarily problematic (Devezer et al., **2021**), and that any "circular reasoning" involved in this process can be identified by checking the *contents* of the reasoning without needing to know the *timing* of the reasoning (Rubin & Donkin, **2022**).

Second, even if we accept the validity of the confirmatory-exploratory distinction and agree that, *all other things being equal,* exploratory results tend to be more tentative than confirmatory results, it would be a fallacy of the general rule to conclude that *all* exploratory results are more tentative than *all* confirmatory results. For example, an exploratory result may be evaluated as being *less* tentative than a confirmatory result when it is based on higher quality theory, methods, and analyses than the confirmatory result and when it is accompanied by greater transparency vis-à-vis robustness analyses and open data and materials (Devezer et al., **2021**; Morey, **2019**; Rubin, **2020**; Szollosi et al., **2020**). Consequently, it would be a QMP to argue that "exploratory studies cannot be presented as strong evidence in favor of a particular claim" (Wagenmakers et al., **2012**, p. 635), because *high quality* exploratory studies can provide stronger evidence than *low quality* confirmatory studies (see also Rubin, **2017b**, p. 314).

Presuming QRPs are Problematic

Another sweeping generalization QMP is to presume that *questionable* research practices are always *problematic* research practices. For example, Hartgerink and Wicherts (**2016**, p. 1) defined QRPs as "practices that are detrimental to the research process…[and that] harm the research process"; Chambers (**2014**) described QRPs as "soft fraud"; and Schimmack (**2020**, p. 372) proposed that "the most obvious solution [to the replication crisis] is to ban the use of questionable research practices and to treat them like other types of unethical behaviours." There are two problems with this position.

First, QRPs can be perfectly acceptable research practices (Fiedler & Schwarz, **2016**; Moran et al., **2022**, Table 6; Rubin, **2022**, p. 551; Sacco et al., **2019**). For example, the QRP of "failing to report all of a study's dependent measures" (John et al., **2012**, p. 525) may not indicate *p*-hacking if (a) there are good reasons to exclude the measures from the research report and (b) the excluded measures are irrelevant to the final research conclusions (Fiedler & Schwarz, **2016**, p. 46; John et al., **2012**, p. 531; Rubin, **2017b**; Rubin, **2020**). As their name implies, QRPs need to be "questioned" by other researchers and interpretated in specific research situations before they can be judged to be potentially problematic.

Second, even potentially problematic research practices such as HARKing and *p*-hacking may not always be problematic for research credibility and replicability (e.g., Bak-Coleman et al., **2022**; Devezer et al., **2019**; Fanelli, **2018**; Leung, **2011**; Rubin, **2017a**; Rubin, **2017b**; Rubin, **2020**; Rubin, **2022**; T. D. Stanley et al., **2018**; Ulrich & Miller, **2020**; Vancouver, **2018**). Hence, a more tenable position is to assume that only *some* QRPs are *potentially* problematic in specific research situations, and only *some* potentially problematic research practices are *actually* problematic under *some* conditions.

Science Characterization QMPs

Focusing on Knowledge Accumulation

Some metascientists assume that the goal of science is to accumulate knowledge (e.g., Errington, Mathur, et al., **2021**, p. 1; Munafò et al., **2017**; Nosek et al., **2012**; Vazire, **2018**). For example, Nosek et al. (**2012**, p. 617) explained that "the primary objective of science as a discipline is to accumulate knowledge about nature," and Vazire (**2018**, p. 416) explained that "the common goal among all scientists is to accumulate knowledge." Commentators have noted that, from this perspective, some metascientists view low replication rates as indicating an "inefficient" accumulation of knowledge (Morawski, **2022**; Peterson & Panofsky, **2021**; for examples, see Errington, Mathur, et al., **2021**; Munafò et al., **2017**; Nosek et al., **2012**; Vazire, **2018**; for discussions, see Hostler, **2022**; Uygun Tunç et al., **2022**). The proposed open science reforms are supposed to improve the efficiency of knowledge accumulation (e.g., Chambers & Tzavella, **2022**, p. 37; Nosek et al., **2012**, p. 626). For example, Nosek et al. (**2012**, p. 626) concluded that "scientific practices can be improved to enhance the efficiency of knowledge building."

However, there are two reasons that knowledge accumulation may not be regarded as the primary objective of science. First, different philosophies of science emphasize different goals. For example, besides knowledge accumulation, Dellsén (**2018**) described three alternative goals of science: truth-seeking, problem-solving, and understanding. Second, any philosophy of science that posits knowledge as a goal should also acknowledge the complementary role of ignorance: "What does this unexpected effect mean?" and "why did we find a null result in this study?" These sorts of known unknowns are essential for scientific progress because they motivate the generation of hypotheses for future studies. Hence, according to this "knowledge-*and*-ignorance" perspective, scientific progress is achieved through not only knowledge accumulation, but also *specified ignorance* (Firestein, **2012**; Merton, **1987**; Open Science Collaboration, **2015**, p. 7; Rubin, **2021a**, p. 5826; Smithson, **1996**).

Importantly, knowledge accumulation and specified ignorance have opposite associations with replicability. Successful replications represent scientific progress by confirming current hypotheses. However, failed replications also represent scientific progress by motivating the generation of new hypotheses that explain why the replications failed (e.g., by positing boundary conditions; for an example, see Firestein, **2012**). Hence, although low replication rates may indicate poor knowledge accumulation, they may also represent scientific progress vis-à-vis greater specified ignorance.

In summary, definitions of scientific progress depend on the types of goals to be achieved (Haig, **2022**, p. 236). Metascientists who assume that knowledge accumulation is central to scientific progress should also acknowledge that (a) other philosophies of science regard other objectives as being more important, and (b) specified ignorance is equally as important as knowledge accumulation.

Homogenizing Science

As several commentators have noted, some metascientists appear to assume that there is a single scientific method rather than a collection of diverse methods (for commentators, see Drummond, **2019**; Malich & Rehmann-Sutter, **2022**; Peterson & Panofsky, **2020**, p. 21; see also Guttinger, **2020**, p. 2). Malich and Rehmann-Sutter (**2022**, pp. 4-6) argued that this "homogenizing view" is apparent every time a metascientist refers to "the scientific method" in the singular and without qualification (e.g., Munafò et al., **2017**, p. 7; Nosek et al., **2012**, p. 618; Zwaan et al., **2018**, p. 13; for further examples, see Drummond, **2019**, p. 64).

In addition, and at the risk of homogenizing metascience (Field, **2022**), some (not all) metascientists focus their concerns on particular aspects of "the scientific method" (Flis, **2019**). In particular, the contemporary metascientific view of science tends to focus on:

1. a priori predictions (e.g., Chambers & Tzavella, **2022**, p. 36; Simmons et al., **2021**, p. 154);
2. quantitative methods (Bennett, **2021**; Hamlin, **2017**, p. 691; Pownall et al., **2021**, p. 530);
3. rigorous statistical analyses (for a review, see Moody et al., **2022**);
4. replicable effects (e.g., Nosek et al., **2012**, p. 617; Simmons et al., **2021**, p. 153);

Table 1 Questionable Metascience Practices

	Name	Definition	Recommended Practice
1	Rejecting or ignoring self-criticism	Rejecting or ignoring criticisms of meta-science and/or science reform	Encourage self-criticism, cite critics' work, and respond in a thoughtful manner
2	Fast 'n' bropen criticism	A quick, superficial, dismissive, and/or mocking style of scientific criticism	Undertake careful "critical evaluation with civility and mutual respect" (Society for the Improvement of Psychological Science, **2022**)
3	Overplaying replication	Assuming that replication is essential to science, and that it indexes "the truth"	Qualify and contextualize claims about the centrality and role of replication in science
4	Unspecified replication rate targets	Assuming that a replication rate is "too low" without specifying an "acceptable" rate	Elaborate on the meaning of "low" when discussing "low replication rates"
5	Metabias	A bias towards explaining the replication crisis in terms of researcher bias	Undertake a more balanced and comprehensive assessment of explanations for the replication crisis
6	The bias reduction assumption	Focusing on selective reporting as the primary form of researcher bias and assuming that it can be reduced without increasing other forms of bias	Consider other forms of researcher bias (e.g., selective questioning, researcher commitment bias) and reveal different research perspectives (e.g., through robustness analyses and researcher positionality statements)
7	Devaluing exploratory hypothesis tests	Devaluing an exploratory result as being more "tentative" than a confirmatory result without considering other relevant issues (e.g., quality of associated theory, methods, analyses, transparency)	Acknowledge that some exploratory results can be *less* tentative than some confirmatory results
8	Presuming QRPs are problematic	Presuming that questionable research practices are always problematic research practices	Acknowledge that only *some* QRPs are *potentially* problematic in specific research situations, and only *some* potentially problematic research practices are *actually* problematic under *some* conditions
9	Focusing on knowledge accumulation	Conceiving knowledge accumulation as the primary objective of science without considering (a) the role of specified ignorance or (b) different objectives in other philosophies of science	Acknowledge that (a) knowledge accumulation and specified ignorance go hand-in-hand and (b) different philosophies of science define scientific progress differently
10	Homogenizing science	Focusing on specific approaches as "the scientific method"	Diversify membership in the metascience community and embrace scientific diversity and pluralism

5. unbiased interpretations (e.g., Hardwicke & Wagenmakers, **2023**; Vazire et al., **2022**, p. 166); and
6. a Popperian philosophy of science (Flis, **2019**; Grossmann, **2021**, p. 74; Morawski, **2019**; Morawski, **2022**; for examples, see Derksen, **2019**).

However, from a critical perspective, these foci may be associated with:

1. *predictivism*: the view that a priori predictions are superior to post hoc inferences (Oberauer & Lewandowsky, **2019**, p. 1605; Rubin, **2017b**; Rubin, **2022**);
2. *methodolatory/methodologism*: the prioritizing of methodological rigor over other research concerns, such as theory (Chamberlain, **2000**; Danziger, **1990**, p. 5; Gao, **2014**);
3. *statisticism/mathematistry*: an overemphasis on statistics as both a problem and a solution in science (Boring, **1919**; Brower, **1949**; Fiedler, **2018**; Proulx & Morey, **2021**);
4. *naïve empiricism* (Strong, **1991**): the view that science progresses through the accumulation of replicable effects (Flis, **2022**; Proulx & Morey, **2021**; van Rooij & Baggio, **2021**);
5. *naïve objectivism*: the view that it is possible for scientists to adopt unbiased and objective perspectives (Field & Derksen, **2021**; Penders, **2022**; Wiggins & Christopherson, **2019**); and
6. a fairly narrow and outdated philosophy of science (Derksen, **2019**; Flis, **2019**, p. 170; Grossmann, **2021**, p. 74; Morawski, **2019**, p. 226, p. 233).

Furthermore, several commentators have noted that these metascientific foci may have the unintended consequence of alienating scientists whose work does not fit with this particular view of science (Bennett, **2021**; Kessler et al., **2021**; Levin & Leonelli, **2017**; Malich & Rehmann-Sutter, **2022**; McDermott, **2022**, p. 58; Penders, **2022**; Pownall et al., **2021**, p. 530; Prosser et al., **2022**; Wentzel, **2021**, p. 170). To address this problem, and to facilitate the recognition of their own biases, metascientists should continue to diversify their membership and embrace scientific diversity and pluralism (Andreoletti, **2020**; Flis, **2022**; Gervais, **2021**; Grossmann, **2021**; Leonelli, **2022**; Pownall, **2022**).

Table 1 summarizes the 10 QMPs that I have discussed and includes recommended practices in relation to each one.

Conclusion

Paralleling John et al.'s (**2012**) concept of questionable research practices, the present article considered a nonexhaustive list of 10 questionable metascience practices. Readers may disagree about the importance of specific QMPs. However, in my view, it remains useful for metascientists to consider the basic concept of QMPs and to reflect on the ways in which they (a) handle criticism, (b) conceptualize replication, (c) consider researcher bias, (d) avoid sweeping generalizations, and (e) acknowledge the diversity and pluralism of science.

In discussing QMPs, we should be careful not to homogenize metascience (Field, **2022**) or to presume that QMPs are necessarily problematic. It is likely that only *some* metascientists engage in *some* QMPs *some* of the time and that QMPs are only problematic in *some* situations. Future metascientific research may wish to assess the prevalence and impact of various QMPs in order to obtain a clearer understanding of these issues. In the meantime, QMPs should be regarded as invitations to reflect on metascientific practices, assumptions, and perspectives and to ask "questions" about how we go about doing better metascience.

References

Allum, N., Reid, A., Bidoglia, M., Gaskell, G., Aubert-Bonn, N., Buljan, I., & Veltri, G. (2023). Researchers on research integrity: A survey of European and American researchers. *F1000Research*, *12*(187), 187. https://doi.org/10.12688/f1000research.128733.1 (see p. 6).

Altenmüller, M. S., Nuding, S., & Gollwitzer, M. (2021). No harm in being self-corrective: Self-criticism and reform intentions increase researchers' epistemic trustworthiness and credibility in the eyes of the public. *Public Understanding of Science*, *30*(8), 962–976. https://doi.org/10.1177/09636625211022181 (see p. 6).

Andreoletti, M. (2020). Replicability crisis and scientific reforms: Overlooked issues and unmet challenges. *International Studies in the Philosophy of Science*, *33*(3), 135–151. https://doi.org/10.1080/02698595.2021.1943292 (see p. 13).

Anonymous. (2021, November 25). *It's 2021... and we are still dealing with misogyny in the name of open science*. University of Sussex School of Psychology Blog. https://blogs.sussex.ac.uk/psychology/2021/11/25/its-2021-and-we-are-still-dealing-with-misogyny-in-the-name-of-open-science/ (see p. 7).

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. V., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919 (see p. 7).

Bak-Coleman, J. B., Mann, R. P., West, J., & Bergstrom, C. T. (2022). Replication does not measure scientific productivity. https://doi.org/10.31235/osf.io/rkyf7 (see pp. 7, 10).

Barrett, L. F. (2015, September 1). Psychology is not in crisis. *The New York Times*. https://www3.nd.edu/~ghaeffel/ScienceWorks.pdf (see p. 8).

Bastian, H. (2021, October 31). The metascience movement needs to be more self-critical. In *Plos blogs: Absolutely maybe*. https://absolutelymaybe.plos.org/2021/10/31/the-metascience-movement-needs-to-be-more-self-critical/ (see p. 6).

Bennett, E. A. (2021). Open science from a qualitative, feminist perspective: Epistemological dogmas and a call for critical examination. *Psychology of Women Quarterly*, *45*(4), 448–456. https://doi.org/10.1177/03616843211036460 (see pp. 11, 13).

Bird, A. (2020). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965–993. https://doi.org/10.1093/bjps/axy051 (see p. 8).

Bishop, D. V. M. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*(7753), 435–435. https://doi.org/10.1038/d41586-019-01307-2 (see p. 5).

Bishop, D. V. M. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *73*(1), 1–19. https://doi.org/10.1177/1747021819886519 (see pp. 5, 6, 8).

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*(10), 335–338. https://doi.org/10.1037/h0074554 (see p. 13).

Brower, D. (1949). The problem of quantification in psychological science. *Psychological Review*, *56*(6), 325–333. https://doi.org/10.1037/h0061802 (see p. 13).

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, *10*(3). https://doi.org/10.1098/rsos.221042 (see pp. 7, 8).

Chamberlain, K. (2000). Methodolatry and qualitative health research. *Journal of Health Psychology*, *5*(3), 285–296. https://doi.org/10.1177/135910530000500306 (see p. 13).

Chambers, C. D. (2014, June 10). Physics envy: Do 'hard' sciences hold the solution to the replication crisis in psychology? *The Guardian*. http://www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology (see p. 10).

Chambers, C. D. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press. (See pp. 7, 8).

Chambers, C. D. (2018, January 25). Registered Reports as a vaccine against research bias: Past, present and future. In *Presentation at registered reports workshop*. https://doi.org/10.23668/psycharchives.797 (see p. 9).

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*, 29–42. https://doi.org/10.1038/s41562-021-01193-7 (see pp. 8, 9, 10, 11).

Clark, C. J., Tetlock, P. E., Frisby, R. E., O'Donohue, W. T., & Lilienfeld, S. O. (2022). Adversarial collaboration: The next science reform. In C. Frisby, R. Redding, W. O'Donohue, & S. Lilienfeld (Eds.), *Political bias in psychology: Nature, scope, and solutions*. Springer. (See p. 9).

Crețu, A.-M. (2019). Perspectival realism. In M. Peters (Ed.), *Encyclopedia of educational philosophy and theory*. Springer. (See p. 6).

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. Cambridge University Press. (See p. 13).

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*(7), 757–777. https://doi.org/10.1037/bul0000154 (see pp. 7, 8).

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 (see p. 9).

Dellsén, F. (2018). Scientific progress: Four accounts. *Philosophy Compass*, *13*(11), 12525. https://doi.org/10.1111/phc3.12525 (see p. 11).

Dellsén, F. (2020). The epistemic impact of theorizing: Generation bias implies evaluation bias. *Philosophical Studies*, *177*, 3661–3678. https://doi.org/10.1007/s11098-019-01387-w (see p. 9).

Derksen, M. (2019). Putting Popper to work. *Theory & Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343 (see p. 13).

Derksen, M., & Field, S. (2022). The tone debate: Knowledge, self, and social order. *Review of General Psychology*, *26*(2), 172–183. https://doi.org/10.1177/10892680211015636 (see p. 7).

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication". *Perspectives on Psychological Science*, *17*(5), 1490–1505. https://doi.org/10.1177/17456916211041116 (see p. 9).

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one*, *14*(5). https://doi.org/10.1371/journal.pone.0216125 (see pp. 7, 8, 10).

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3). https://doi.org/10.1098/rsos.200805 (see pp. 5, 7, 9, 10).

Drummond, C. (2019). Is the drive for reproducible science having a detrimental effect on what is published? *Learned Publishing*, *32*(1), 63–69. https://doi.org/10.1002/leap.1224 (see pp. 7, 11).

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

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *Elife*, *10, Article e71601*. https://doi.org/10.7554/eLife.71601 (see pp. 7, 11).

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366 (see p. 8).

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*(11), 2628–2631. https://doi.org/10.1073/pnas.1708272114 (see pp. 8, 10).

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451 (see p. 7).

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*(4), 433–438. https://doi.org/10.1177/1745691617745651 (see p. 13).

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150 (see p. 10).

Field, S. M. (2022, July 13). *Charting the constellation of science reform*. https://doi.org/10.31219/osf.io/udfw4 (see pp. 11, 13).

Field, S. M., & Derksen, M. (2021). Experimenter as automaton; experimenter as human: Exploring the position of the researcher in scientific research. *European Journal for Philosophy of Science*, *11, Article 11*. https://doi.org/10.1007/s13194-020-00324-7 (see pp. 8, 9, 13).

Firestein, S. (2012). *Ignorance: How it drives science*. Oxford University Press. (See p. 11).

Firestein, S. (2016, February 14). Why failure to replicate findings can actually be good for science. *LA Times*. https://www.latimes.com/opinion/op-ed/la-oe-0214-firestein-science-replication-failure-20160214-story.html (see p. 8).

Fiske, S. T. (2016, October 31). A call to change science's culture of shaming. *APS Observer*, *29*. https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming (see p. 6).

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322 (see pp. 8, 11, 13).

Flis, I. (2022). The function of literature in psychological science. *Review of General Psychology*, *26*(2), 146–156. https://doi.org/10.1177/10892680211066466 (see pp. 6, 13).

Freiling, I., Krause, N. M., Scheufele, D. A., & Chen, K. (2021). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*, *71*(5), 686–714. https://doi.org/10.1093/joc/jqab032 (see pp. 8, 9).

Gao, Z. (2014). Methodologism/methodological imperative. In T. Teo (Ed.), *Encyclopedia of critical psychology*. Springer. https://doi.org/10.1007/978-1-4614-5583-7_614 (see p. 13).

Gervais, W. M. (2021). Practical methodological reform needs good theory. *Perspectives on Psycho-*

logical Science, *16*(4), 827–843. https://doi.org/10.1177/1745691620977471 (see pp. 6, 7, 13).

Giere, R. N. (2006). *Scientific perspectivism*. Chicago Press. (See p. 6).

Greenfield, P. M. (2017). Cultural change over time: Why replicability should not be the gold standard in psychological science. *Perspectives on Psychological Science*, *12*(5), 762–771. https://doi.org/10.1177/1745691617707314 (see p. 7).

Grossmann, M. (2021). *How social science got better: Overcoming bias with more evidence, diversity, and self-reflection*. Oxford University Press. (See pp. 9, 13).

Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science*, *10*(2), 1–17. https://doi.org/10.1007/s13194-019-0269-1 (see pp. 7, 8, 11).

Haig, B. D. (2022). Understanding replication in a way that is true to science. *Review of General Psychology*, *26*(2), 224–240. https://doi.org/10.1177/10892680211046514 (see pp. 6, 7, 8, 11).

Hamlin, J. K. (2017). Is psychology moving in the right direction? An analysis of the evidentiary value movement. *Perspectives on Psychological Science*, *12*(4), 690–693. https://doi.org/10.1177/1745691616689062 (see pp. 6, 7, 11).

Hardwicke, T. E., & Wagenmakers, E. (2023). Reducing bias, increasing transparency, and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*, 15–26. https://doi.org/10.1038/s41562-022-01497-2 (see pp. 8, 9, 10, 13).

Hartgerink, C. H., & Wicherts, J. M. (2016). Research practices and assessment of research misconduct. *ScienceOpen Research*, *0*(0), 1–10. https://doi.org/10.14293/S2199-1006.1.SOR-SOCSCI.ARYSBI.v1 (see p. 10).

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8 (see p. 6).

Holcombe, A. O. (2021). Ad hominem rhetoric in scientific psychology. *British Journal of Psychology*, *113*(2), 434–454. https://doi.org/10.1111/bjop.12541 (see p. 7).

Hostler, T. (2022). Open research reforms and the capitalist university's priorities and practices: Areas of opposition and alignment. *SocArXiv*. https://doi.org/10.31235/osf.io/r4qgc (see p. 11).

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010 (see pp. 8, 10).

Iso-Ahola, S. E. (2020). Replication and the establishment of scientific truth. *Frontiers in Psychology*, *11, Article 2183*. https://doi.org/10.3389/fpsyg.2020.02183 (see p. 7).

Jamieson, M. K., Pownall, M., & Govaart, G. H. (2023). Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass, Article*, *e12735*. https://doi.org/10.1111/spc3.12735 (see p. 9).

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 (see pp. 5, 10, 13).

Kessler, A., Likely, R., & Rosenberg, J. M. (2021). Open for whom? The need to define open science for science education. *Journal of Research in Science Teaching*, *58*(10), 1590–1595. https://doi.org/10.1002/tea.21730 (see p. 13).

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. including a symposium on Mary Morgan: Curiosity, imagination, and surprise. *Research in the History of Economic Thought and Methodology*, *36B*, 129–146. https://doi.org/10.1108/S0743-41542018000036B009 (see p. 7).

Leonelli, S. (2022). Open science and epistemic diversity: Friends or foes? *Philosophy of Science*, *89*(5), 991–1001. https://doi.org/10.1017/psa.2022.45 (see p. 13).

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, *7*(3), 471–479. https://doi.org/10.1111/j.1740-8784.2011.00222.x (see p. 10).

Levin, N., & Leonelli, S. (2017). How does one "open" science? Questions of value in biological research. *Science, Technology, & Human Values*, *42*(2), 280–305. https://doi.org/10.1177/0162243916672071 (see p. 13).

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11, Article 358*. https://doi.org/10.1038/s41467-019-14203-0 (see p. 8).

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press. (See p. 6).

Malich, L., & Rehmann-Sutter, C. (2022). Metascience is not enough – a plea for psychological humanities in the wake of the replication crisis. *Review of General Psychology*, *26*(2), 261–273. https://doi.org/10.1177/10892680221083876 (see pp. 6, 8, 11, 13).

Massimi, M. (2022). *Perspectival realism*. Oxford University Press. (See p. 6).

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400 (see pp. 7, 8).

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 pp. 9, 13).

Merton, R. K. (1987). Three fragments from a sociologist's notebooks: Establishing the phenomenon, specified ignorance, and strategic research materials. *Annual Review of Sociology*, *13*(1), 1–29. https://doi.org/10.1146/annurev.so.13.080187.000245 (see p. 11).

Moody, J. W., Keister, L. A., & Ramos, M. C. (2022). Reproducibility in the social sciences. *Annual Review of Sociology*, *48*, 65–85. https://doi.org/10.1146/annurev-soc-090221-035954 (see pp. 8, 11).

Moran, C., Richard, A., Wilson, K., Twomey, R., & Coroiu, A. (2022). I know it's bad, but i have been pressured into it: Questionable research practices among psychology students in Canada. *Canadian Psychology*, *64*(1), 12–24. https://doi.org/10.1037/cap0000326 (see p. 10).

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Journal of Theoretical and Philosophical Psychology*, *39*(4), 218–238. https://doi.org/10.1037/teo0000129 (see pp. 8, 9, 13).

Morawski, J. (2022). How to true psychology's objects. *Review of General Psychology*, *26*(2), 157–171. https://doi.org/10.1177/10892680211046518 (see pp. 8, 11, 13).

Morey, R. (2019). You must tug that thread: Why treating preregistration as a gold standard might incentivize poor behavior. https://featuredcontent.psychonomic.org/you-must-tug-that-thread-why-treating-preregistration-as-a-gold-standard-might-incentivize-poor-behavior/ (see pp. 9, 10).

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). https://doi.org/10.1038/s41562-016-0021 (see pp. 5, 7, 8, 11).

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836 (see pp. 7, 10).

Norton, J. D. (2015). Replicability of experiment. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, *30*(2), 229–248. https://doi.org/10.1387/theoria.12691 (see p. 7).

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009 (see p. 6).

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 p. 10).

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., & 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 (see pp. 7, 8).

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192 (see p. 10).

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058 (see pp. 5, 7, 8, 11).

Oberauer, K. (2019). Preregistration of a forking path – what does it add to the garden of evidence? https://featuredcontent.psychonomic.org/preregistration-of-a-forking-path-what-does-it-add-to-the-garden-of-evidence/ (see p. 9).

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology [Article]. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2 (see p. 13).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716 (see pp. 5, 7, 8, 11).

Penders, B. (2022). Process and bureaucracy: Scientific reform as civilisation. *Bulletin of Science, Technology & Society*, *42*(4), 107–116. https://doi.org/10.1177/02704676221126388 (see p. 13).

Peterson, D., & Panofsky, A. (2020). Metascience as a scientific social movement. https://osf.io/preprints/socarxiv/4dsqa/ (see pp. 8, 11).

Peterson, D., & Panofsky, A. (2021). Arguments against efficiency in science. *Social Science Information*, *60*(3), 350–355. https://doi.org/10.1177/05390184211021383 (see pp. 8, 11).

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. 9).

Pownall, M. (2022). Is replication possible for qualitative research? *PsyArXiv*. https://doi.org/10.31234/osf.io/dwxeg (see pp. 7, 9, 13).

Pownall, M., & Hoerst, C. (2022). Slow science in scholarly critique. *The Psychologist*, *35*, 2. https://thepsychologist.bps.org.uk/volume-35/february-2022/slow-science-scholarly-critique (see p. 6).

Pownall, M., Azevedo, F., Aldoh, A., Elsherif, M., Vasilev, M., Pennington, C. R., Robertson, O., Tromp, M. V., Liu, M., Makel, M. C., Tonge, N., Moreau, D., Horry, R., Shaw, J., Tzavella, L., McGarrigle, R., Talbot, C., & Parsons., S. (2021). Embedding open and reproducible science into teaching: A bank of lesson plans and resources. In *Scholarship of teaching and learning in psychology*. https://doi.org/10.1037/stl0000307 (see pp. 7, 11, 13).

Prosser, A. M. B., Hamshaw, R. J. T., Meyer, J., Bagnall, R., Blackwood, L., Huysamen, M., Jordan, A., Vasileiou, K., & Walter, Z. (2022). When open data closes the door: Problematising a one size fits all approach to open data in journal submission guidelines. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12576 (see p. 13).

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, *16*(4), 671–681. https://doi.org/10.1177/17456916211017098 (see p. 13).

Reiss, J., & Sprenger, J. (2020). Scientific objectivity. In *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/scientific-objectivity/ (see p. 9).

Rosnow, R. L. (1983). Von osten's horse, hamlet's question, and the mechanistic view of causality: Implications for a post-crisis social psychology. *The Journal of Mind and Behavior*, *4*(3), 319–337. http://www.jstor.org/stable/43852983 (see p. 8).

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 (see pp. 6, 10).

Rubin, M. (2017b). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, *21*(4), 308–320. https://doi.org/10.1037/gpr0000128 (see pp. 6, 10, 13).

Rubin, M. (2020). 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 (see pp. 6, 9, 10).

Rubin, M. (2021a). 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 (see pp. 6, 7, 8, 11).

Rubin, M. (2021b). 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 (see pp. 6, 10).

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 (see pp. 6, 10, 13).

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology*, 1–29. https://doi.org/10.1080/09515089.2022.2113771 (see pp. 9, 10).

Sacco, D. F., Brown, M., & Bruton, S. V. (2019). Grounds for ambiguity: Justifiable bases for engaging in questionable research practices. *Science and Engineering Ethics*, *25*(5), 1321–1337. https://doi.org/10.1007/s11948-018-0065-x (see p. 10).

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/25152459211007467 (see p. 9).

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology*, *61*(4), 364–376. https://doi.org/10.1037/cap0000246 (see pp. 5, 8, 10).

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845 (see p. 8).

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. 5).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2021).

Pre registration: Why and how. *Journal of Consumer Psychology*, *31*(1), 151–162. https://doi.org/10.1002/jcpy.1208 (see pp. 7, 8, 10, 11).

Smithson, M. (1996). Science, ignorance and human values. *Journal of Human Values*, *2*(1), 67–81. https://doi.org/10.1177/097168589600200107 (see p. 11).

Society for the Improvement of Psychological Science. (2022). Mission statement. https://improvingpsych.org/mission/ (see pp. 7, 12).

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. In J. Wixted & E.-J. Wagenmakers (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience: Volume 5 Methodology* (4th ed., pp. 729–775, Vol. 5). Wiley. https://doi.org/10.1002/9781119170174.epcn519 (see p. 5).

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518 (see pp. 7, 8).

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169 (see p. 10).

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450 (see p. 8).

Strong, S. R. (1991). Theory-driven science and naïve empiricism in counseling psychology. *Journal of Counseling Psychology*, *38*(2), 204–210. https://doi.org/10.1037/0022-0167.38.2.204 (see pp. 9, 13).

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 (see p. 10).

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Science*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009 (see pp. 6, 9, 10).

Ulrich, R., & Miller, J. (2020). Meta-research: Questionable research practices may have little effect on replicability. *Elife*, *9, Article e58237*. https://doi.org/10.7554/eLife.58237 (see p. 10).

Uygun Tunç, D., Tunç, M. N., & Eper, Z. B. (2022). Is open science neoliberal? Perspectives on psychological science. https://doi.org/10.1177/17456916221114835 (see p. 11).

Vancouver, J. N. (2018). In defense of HARKing. *Industrial and Organizational Psychology*, *11*(1), 73–80. https://doi.org/10.1017/iop.2017.89 (see p. 10).

van Dijk, T. (2021, June 22). *How to tackle confirmation bias?* Journalistic Platform TU Delft. https://www.delta.tudelft.nl/article/how-tackle-confirmation-bias (see p. 9).

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604 (see p. 13).

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884 (see p. 11).

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779 (see pp. 9, 13).

Wagenmakers, E. J. (2012). A year of horrors. *De Psychonoom*, *27*, 12–13 (see p. 10).

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 (see p. 10).

Walkup, J. (2021). Replication and reform: Vagaries of a social movement. *Journal of Theoretical and Philosophical Psychology*, *41*(2), 131–133. https://doi.org/10.1037/teo0000171 (see p. 6).

Wentzel, K. R. (2021). Open science reforms: Strengths, challenges, and future directions. *Educational Psychologist*, *56*(2), 161–173. https://doi.org/10.1080/00461520.2021.1901709 (see p. 13).

Whitaker, K., & Guest, O. (2020). #Bropenscience is broken science. *The Psychologist*, *33*, 34–37. https://thepsychologist.bps.org.uk/volume-33/november-2020/bropenscience-broken-science (see p. 6).

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137 (see pp. 9, 13).

Wood, W., & Wilson, T. D. (2019). No crisis but no time for complacency. *APS Observer*, *32*(7). https://www.psychologicalscience.org/observer/no-crisis-but-no-time-for-complacency (see p. 8).

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41, Article E120*. https://doi.org/10.1017/S0140525X17001972 (see pp. 7, 11).

The Invisible Workload of Open Research

Thomas J. Hostler[1]

It is acknowledged that conducting open research requires additional time and effort compared to conducting 'closed' research. However, this additional work is often discussed only in abstract terms, a discourse which ignores the practicalities of how researchers are expected to find the time to engage with these practices in the context of their broader role as multifaceted academics. In the context of a sector that is blighted by stress, burnout, untenable workloads, and hyper-competitive pressures to produce, there is a clear danger that additional expectations to engage in open practices add to the workload burden and increase pressure on academics even further. In this article, the theories of academic capitalism and workload creep are used to explore how workload models currently exploit researchers by mismeasuring academic labour. The specific increase in workload resulting from open practices and associated administration is then outlined, including via the cumulative effects of administrative burden. It is argued that there is a high chance that without intervention, increased expectations to engage in open research practices may lead to unacceptable increases in demands on academics. Finally, the individual and systematic responsibilities to mitigate this are discussed.

Keywords *academic capitalism, workload, burnout, administrative burden, open research*

[1]Department of Psychology, Manchester Metropolitan University

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
September 8, 2022
Accepted
April 17, 2023
Published
May 4, 2023
Issued
May 24, 2024

Correspondence
Department of Psychology, Manchester Metropolitan University
t.hostler@mmu.ac.uk

It is widely accepted that conducting open research can improve the endeavour of collaborative human knowledge generation. Here, "open research"[1] refers to a variety of practices that make the plans, procedures, labour, and outputs of research publicly available, although the phrase covers broader meanings elsewhere (Fecher & Friesike, **2014**). Any one individual open practice such as preregistration or data sharing may have a variety of axiological benefits (Uygun Tunç et al., **2022**), but taken together, transparency of the research process improves the epistemic reliability of a piece of research, which facilitates incremental knowledge generation. It also supports an environment by which epistemically unreliable research (whether through errors or bias) can be discounted or ignored (Lakens & Evers, **2014**).

Compared to 'closed' research, where the only publicly available element of research is a final journal report, open research involves transparently cataloguing as much of the research process as possible. It has been argued that transparent workflows can reduce inefficiency and save time once implemented (Lowndes et al., **2017**). However, in practice, conducting open research involves additional time and effort compared to closed research, not only in the process of sharing materials to disciplinary standards, but also in the development of new skills and knowledge to enable this. Generally, the additional work required to conduct open research has been acknowledged (and justified) by proponents of open research reforms (e.g. Allen & Mehler, **2019**; Robson et al., **2021**; Scheliga & Friesike, **2014**; A. J. Stewart et al., **2021**).

However, the current discourse promoting open research fails to engage sufficiently with how additional workload impacts the *practicalities* of academic labour (Callard, **2022**). The majority of literature discussing open research reforms is in the field of 'meta-research' and has typically viewed closed research as a systematic or cultural problem: characterizing researchers as fallible human agents working in systems and contexts that encourage bias

[1]The term 'open research' has been chosen here instead of 'open science' given that many of the practices mentioned are used in disciplines and research paradigms that would not define themselves as sciences (e.g. qualitative research).

and closed practices. Examples include discussion of issues such as the incentives for conducting open practices (Nosek et al., **2012**), biases in publication workflows (Chambers & Tzavella, **2022**), recognition and reward of open research practices (Munafò, **2019**) and compliance with open research mandates (Gabelica et al., **2022**). Whilst these are important issues, the common perspective is that researchers exist solely to conduct research and are primarily judged and motivated by success in this domain. This neglects the fact that the majority of researchers are employed *not* solely as researchers but rather as academics, a role that involves a large number of other activities that compete for time and resources, including teaching, administration, income generation, knowledge exchange, and supervision. Even for academics who are primarily researchers, transparency may not be a priority concern given other important and competing demands such as increasing research regulation (P. M. Stewart et al., **2008**), novel ethical issues (Havard et al., **2012**), or grappling with fundamental issues in theory development (Eronen & Bringmann, **2021**) and methodology (Uher, **2023**). By solely focusing on "open research" as a separable pursuit, meta-research neglects to acknowledge that the additional workload required by open research cannot always be practically accommodated in the day-to-day duties of academics and the time and resources they have available.

This is a critical issue, given the increasing systematic degradation of working standards across academia: There is ample evidence that many academics are already at "capacity" in terms of the amount of work they do (Long et al., **2020**), and yet workloads are still increasing. This has led to endemic levels of stress and burnout in the sector (Urbina-Garcia, **2020**), mental health crises (Nicholls et al., **2022**), recourse to industrial action to protest against overwork (University & College Union, **2022**), and a recognition that the sector is haemorrhaging talent to industry where working conditions are seen to be more favourable (Gewin, **2022**; Seidl et al., **2016**). It is therefore crucial to explore the blind spot in meta-research of how the additional workload of open research may potentially negatively impact on working conditions of academics.

The oversight can be addressed by utilising research from the field of higher education studies, including the theory of academic capitalism (Jessop, **2018**). This theory suggests the scholarly ecosystem can be viewed as a type of market, where institutions are capitalist actors in competition with one another. From this perspective, the way that universities (as the primary employers of most academics) are organised, and subsequently their priorities, policies, and relationship with (and potentially exploitation of) academic labour can be examined. The lens of academic capitalism can therefore offer new insights on the way that open research reforms may be practically prevailed upon academics (Hostler, **2022**), in order to anticipate problems and provide solutions.

The rest of the paper is structured as follows: first, I explain the theoretical framework of academic capitalism, including how academic labour is typically controlled by universities using a "workload model" and exploited via "workload creep". Second, I explain how various open research practices add to the time burden and workload of conducting research, including via the cumulative and unnecessary effect of administrative burden. Third, I will explore how additional open research activities may not be sufficiently accounted for in a workload model, leading to detrimental effects on academics' well-being. Finally, I will conclude with a discussion of potential solutions and the responsibilities of both individuals and institutions to address these issues.

Theoretical Background

Academic Capitalism

The theory of *academic capitalism* comes from higher education research and refers to the tendency for universities to operate in competition with each other in markets, competing over both economic and social capital. This tendency is manifested in their priorities, activities, internal organisation, and management. Whilst there are various specific forms of academic capitalism (Jessop, **2018**), as an overarching theory it provides a framework which enables universities to be considered as strategic actors, rather than passive organizational units (Münch, **2014**).

Through this lens, universities (like all capitalist actors) seek to maximise the utilization of

their resources to remain competitive against one another in a variety of zero-sum 'markets' including student recruitment, research funding, national research evaluation exercises, and national and global university rankings (Collini, **2012**). Primary among a university's resources is its academic labour, which it attempts to "steer" towards its own goals through its internal management policies (Rees, **2015**), and find innovative ways of organizing to improve its (economic) efficiency, for example through the use of fixed-term or part-time contracts (Macfarlane, **2011**).

The strategic deployment of academic labour is not *necessarily* exploitative, and managerialist organization in universities is increasingly tolerated and accepted by academics (Kolsaker, **2008**). However, increased oversight and capitalist logic is also seen to enable normalizing exploitative practices when financial considerations are prioritized over traditional academic professional values (Vican et al., **2020**). This is epitomized by the finding that the majority of casualized academic staff are required to work more hours than which they are paid for in order to complete the work required in their contract (i.e., marking essays to a suitable academic standard; University & College Union, **2019**).

Universities are not purely capitalists and have many competing interests, and the drivers of these interests are dynamic and set by the broader economic and political conditions from which they are created. Often, a university's specific goals are congruous with the metrics and conditions tied to these drivers, for example the criteria used to judge research excellence in national research evaluation exercises. Changes to these metrics subsequently influence the university's strategic plans, leading to the re-allocation of resources and new instructions to academics. The work by open research advocates to change these drivers to reward openness - such as funder mandates for open data (Hefce et al., **2016**), or changes to university ranking criteria (Pagliaro, **2021**) - are therefore some of the most powerful tools for system-wide adoption of open research practices.

However, efforts to change a university's strategic priorities to support open research do nothing to alter the underlying capitalist framework, and so do not tackle the issues of working conditions raised in this article. Academic labour can also potentially be exploited to serve the interests of open research: the competitive pressures of "publish or perish" can easily become "publish open research or perish". Indeed, any changes to working practices or expectations can provide perfect cover (intentionally or not) for increased exploitation (Hostler, **2022**). To understand how the promotion of open research can lead to negative changes in working conditions, a closer look at how academic labour is currently organized is required.

The Workload Model

An academic's job typically consists of a large number of activities in addition to research. A comprehensive analysis of an academic's typical work over a three-month period is provided by Miller (**2019**), covering five main areas of teaching, research, administration, community service, and 'other'. Whilst academics have responsibility for organising *when* they perform each of these duties, universities are increasingly using managerial practices to assign the range and volume of the tasks themselves (Kenny & Fluck, **2022**), in the form of a "workload model": a system for "projectifying" time into limited, measurable slots that can be allocated to different activities (Dollinger, **2020**). Workload models are typically based on an annual measure of time that is allocated across different activities, forming a "split" of work across a year. For example, 40% of time dedicated to research, 40% to teaching and 20% to administration. Certain components of workload models are due to regulatory requirements (Kernohan, **2019**), but they are also a useful tool for a capitalist university seeking to understand and maximise the efficiency of the deployment of its human resources. The use of such models is divisive. Some academics view them as a threat to their autonomy (Boncori et al., **2020**), and whilst they may for some represent a level of protection against being given too many tasks, many others view them as a mechanism for universities to demand unrealistic levels of work from academics by underestimating the time taken for different activities (Papadopoulos, **2017**). There is also a general acknowledgement that workload models are not comprehensive and that a significant proportion of the actual work that academics do is

unaccounted for (Kenny & Fluck, **2019**; Miller, **2019**).

Whilst the time spent on different activities is allocated via workload models, academic performance is typically assessed via departmental or individual targets for outcomes or outputs of academic work. These outcomes are often cascaded from university-level metrics, for example, the number of papers published and in which journals, and the amount of research funding accrued. For researchers employed on temporary contracts, these are targets that are required to secure the next job and "survive" in academia (Anderson et al., **2007**). The discrepancy between the time available to work on different academic activities and the expectations of performance is a key driver of stress and discontent. Many academics already feel that their workloads are at "untenable" levels and that additional time is needed to meet expectations, leading to burnout (Beatson et al., **2021**). Within this context, expectations to perform *additional* duties to conduct open research, whether from formal mandates by funders or universities, or informal social expectations to remain competitive, have the potential to make things worse if insufficiently accounted for (made 'invisible') in a workload model. Unfortunately, historical trends suggest that open research practices will *not* be accounted for in workloads, as I explain below.

Workload Creep

One likely way in which open research may be insufficiently accounted for is via its inclusion in "workload creep". This phenomenon exploits the fact that workload models do not provide a granular breakdown of activities, meaning that expectations around what should be achieved in a given time can be subtly changed without a corresponding change in the amount of hours dedicated to a task (Long et al., **2020**). This is particularly common in the case of research, where changes to research expectations (in terms of quantity or quality of outputs) may be raised without extra time or resources made available. In the case of open research, there is a high potential for expectations to engage in open research to become widespread, but without additional time on academics' workloads dedicated to the activity. These expectations may originate either from employer's performance standards, for example requiring evidence of open research in hiring and promotion criteria (Gärtner et al., **2022**; Robson et al., **2021**), or from mandates for openness from funders, journals, or legislation (Nosek, **2019**).

The broader discourse around academic workload acknowledges the reality of workload creep. Advice to early career academics is to simply learn how to "say no" to requests to perform additional work (Somerville, **2021**), or to ask a manager "what would you like me to stop doing?" (Williams, **2022**). However, it is difficult to apply either of these approaches to changes to research expectations. It is unlikely that open research practices will be explicitly requested by an individual such as a manager to whom one can say "no": academics will either be encouraged or mandated to adopt them by anonymous university or journal policies, which are difficult to contest, or they will do so out of their own volition to remain competitively employable. This then hampers bargaining power in discussions with managers about workloads, making it difficult to secure changes in workload models to accommodate the additional time required.

The issue of workload creep can already be seen in the open research practice of 'open access', where funded research outputs are mandated to be made publicly available, requiring additional work by academics to understand and comply with these requirements (Research Consulting, **2014**). However, this additional work is not reflected in workloads in terms of an increase in the number of hours per year given for research (which in 'full capacity' workloads would require other tasks to be removed), or in changes to performance expectations of an explicit decrease in the number of publications expected per year. Complying with an open access requirement may only take about 30 minutes (Reimer, **2014**), but conducting other open research practices (such as sharing materials, data, code, or preregistrations) would confer a much more significant time burden if expected or mandated in the same way. In the next section I provide several examples of how open research practices may lead to a significant increased time burden.

Additional Workload of Open Research

There are numerous open research practices,

the benefits of which are discussed in detail elsewhere (e.g. Munafò et al., **2017**; Nosek et al., **2018**) and not all need be applied to every piece of research. However, adopting any new practice typically involves making changes to a researchers' existing practices, bringing additional workload. This includes time spent learning and applying 'open' practices, the cumulative workload of novel administrative work, and the indirect labour of teaching and mentoring open research. These are explained in turn below:

Workload of Specific Open Practices

Preregistration

Preregistration involves providing a detailed explanation of a researcher's planned data collection procedure, hypotheses, and analysis plan, so that researcher degrees of freedom in analysis decisions can be observed. However, in order for a preregistration to achieve this functionality, it must be "precise", "specific" and "exhaustive" (Bakker et al., **2020**). This involves communicating plans in a substantially greater level of detail than required in traditional research administration (e.g. for the purposes of ethical review, grant applications), as *multiple* alternative analysis strategies need to be considered and explained (including what will *not* be done) depending on different data collection outcomes including outliers, missing data, and violation of statistical assumptions (Bakker et al., **2020**). The checklist by Wicherts et al. (**2016**) presents 34 different degrees of freedom that researchers should define in advance in a preregistration to prevent *p*-hacking. For the majority of researchers, following such a checklist represents an increase in the explicit planning needed for a research project, where many of their decisions will be based on implicit assumptions. It takes additional time to explicitly articulate these plans and decisions, especially in a form that is understandable to people unfamiliar with the project. The suggestion that a preregistration should typically take only "30–60 minutes" (Aguinis et al., **2020**) is likely to be an inaccurate generalization, depending on the type and complexity of the research and the experience of the researcher, although the time taken should decrease with practice (Nosek et al., **2019**).

Data Sharing

Data sharing is an open research practice that takes significantly more effort compared to a traditional closed approach; insofar as a closed approach takes no effort at all, involving simply ignoring or rebuffing sharing requests (Gabelica et al., **2022**). In contrast, done properly and in line with the principles of Findable, Accessible, Interoperable, Reusable (FAIR; Wilkinson et al., **2016**), data sharing takes a considerable amount of time. First, data must be properly anonymized in order to be shared ethically, a process that is particularly difficult for qualitative data such as interviews (Saunders et al., **2015**), legal documents (Csányi et al., **2021**), and unstructured, high-dimensional data such audio and video recordings (Weitzenboeck et al., **2022**). Second, data should be findable, which means taking time to access and use an appropriate data repository, make sure settings are correct to comply with ethical restrictions, and that sufficient meta-data is provided. Finally, to be accessible, interoperable, and reusable, data must be organized and labelled to community standards and formatted and described in such a way as to be understandable to others who are not familiar with how it was collected or processed. These tasks may represent time-consuming departures from how a researcher typically organises data for their own use.

Analysis Code

Open code refers to sharing the analysis code used to produce the output found in the final report from the research data. This is an open practice that may be unfamiliar to some researchers, especially those who use graphical user interface (GUI) programs such as SPSS where viewing and understanding the underlying code is not necessary to analyse data. Compiling and sharing analysis code may therefore require training and the acquisition of new skills to be able to do this adequately, if a researcher is not experienced in doing this. Some researchers argue that using proprietary programs such as SPSS is not ideal for open research, since it takes more effort for those without access to these programs to utilize and interpret the code (Obels et al., **2020**). This may encourage researchers to utilise open source

alternative programs, such as JASP or *R,* where sharing code is significantly easier, although this in itself involves the development of new software skills and workflows. As with data, the shared code also needs to be written or annotated in such a way to ensure usability for other researchers (Obels et al., **2020**), which takes extra time compared to writing code for one's own use, which may use idiosyncratic shorthand.

Openness Agreements & Administration

In complex projects, additional administration is often required to facilitate the use of open research practices. This is particularly true in research with a large number of collaborators. Here, legal documents may be required to certify sharing agreements for data, materials, or outputs, particularly in cases where different elements of a project have different levels of openness across different time frames. For example, in cases where separable elements of a piece of software may be "owned" by different parties in a collaboration (Levin & Leonelli, **2017**) or where industry collaboration requires delaying the timing of release of results or materials to maintain a competitive edge (Fernández Pinto, **2020**). With the increasing size of research teams and complexity of projects, such agreements become lengthier and take additional time to complete and get approval from all parties. Additional administration (compared to closed research) may also come in the form of recording contributions to research projects (e.g. CREDIT taxonomy; Holcombe, **2019**) or providing meta-data about open research practices in the form of transparency statements or checklists (Aczel et al., **2020**). Administration has a particularly close relationship with workload creep, and below I explain how the theory of administrative burden can further illuminate how the process of integrating minor administrative tasks into existing workflows can exacerbate the time burden of open research.

Administrative Burden

Whilst preparing a large set of audio-visual data to FAIR sharing standards may be a significant technical undertaking, many open research practices may be viewed as essentially administrative tasks involving documenting information about a piece of research. This includes writing preregistrations, data sharing documentation, and statements about the openness (or not) of open practices for different elements of a project. The theory of "administrative burden" (Bozeman, **1993**) can be used to explore how in many cases the additional time spent completing these tasks is unnecessary and unnoticed in workload estimates. Administrative burden theory acknowledges that all administration represents a time burden, but that in many cases it represents unnecessary "red tape" when it does not help to fulfil a regulation's functional objectives. An example of unnecessary research administration might be an ethics form that asks a researcher whether they are using radioactive materials, despite the fact that due to their discipline (e.g. psychology), the answer should be obvious (Bozeman & Youtie, **2020**). In the case of open research, red tape may involve requirements to write transparency statements or complete checklists about the availability of data or materials where none exist (e.g. review papers), or explain the (non)existence of preregistrations for research in which this practice is not required or its use contested (e.g. exploratory or qualitative research).

Red tape can also be seen in the case of "rule redundancy" (Bozeman & Jung, **2017**) resulting from bureaucratic overlap, where administration such as explanations of data sharing arrangements is duplicated across platforms (e.g. for funding applications, ethics applications, and journal requirements), but often with different specifications. Another example of rule redundancy is with preregistrations, documents which may closely mirror elements of existing research administration, such as research protocols required for ethical review. Whilst the existence of a protocol may make completing a separate preregistration easier (or vice-versa), *functionally* the duplication of the information may be unnecessary if one document could potentially serve both purposes, yet generates extra workload when the different formats of each document requires time to adapt the content to move information between the two.

Administrative burden is often exacerbated when technology is used to remotely facilitate administration and therefore lacks the nuance to accurately capture the reality of a particular context. Dialogue with the technology provider

is then required to resolve discrepancies, inconspicuously increasing time burden. For example, platforms or infrastructure to facilitate open practices such as preregistration or sharing data or materials may be unclearly worded or not fit for purpose for particular kinds of research or data (e.g. Borgerud & Borglund, **2020**; Rhys Evans et al., **2021**). This is particularly the case in disciplines such as the humanities, where what constitutes "data" or even "research outputs" may be unusual (including physical objects). The widespread adoption of the ethical norms and terminology from positivist biomedical research in the ethical review process is inappropriate for much social science research and a historical example of creating extra administrative burden for certain groups of scholars (Schneider, **2015**). This foreshadows the potential for open research practices such as preregistration, developed from a similarly narrow statistical perspective (Nosek et al., **2018**), to also be misapplied to other areas of research if administered remotely. Increased administrative burden also occurs when existing technology and systems in the research ecosystem fail to keep pace with developments and trends in research resulting from greater openness. This issue can already be seen in archaic manuscript submission systems which do not accommodate the hundreds of authors found on "big team science" projects enabled by the use of open research practices, requiring significant additional time spent doing administration (Forscher et al., **2022**). The scope for the multitude of open research requirements and applications to novel forms of research to outpace existing technology means that such examples are likely to become more common.

The growth of administrative burden has two facets that make it difficult to combat. The first is that administration can often be convincingly defended on the basis that it collects data that has potential utility or that it is necessary to assure compliance with regulations or mandates. However, whether all the data collected from research administration *is* strictly necessary or ever actually used is contested (O'Leary et al., **2013**), and as many open research practices and associated administration are not yet widely adopted there is a lack of evidence on the actual benefits (e.g., of transparency statements) to consider against potential or actual time costs. There are strong arguments that existing research administration is already excessive and prohibitive, particularly for certain types of research such as clinical trials (D. J. Stewart et al., **2015**). Some estimates put the amount of time allocated to research that is spent on administration at 42% (Rockwell, **2009**) and there is evidence that researchers already employ "workarounds" or exhibit noncompliance to reduce this burden (Bozeman et al., **2021**). More fundamentally, it has been argued that research administration is an ineffective way to ensure compliance with regulations (Schneider, **2015**), as it can easily be falsified (e.g., claiming data is available when it is not; Gabelica et al., **2022**).

The second issue is that administrative burden is a *cumulative* problem. The time cost of any one individual instance of administration can easily be dismissed as trivial: it might take only ten minutes to complete a transparency checklist. In the context of a workload model measuring time annually, this represents <.01% of a researcher's time. However, cumulatively, such administration adds up. In addition to a checklist about methodological transparency (Aczel et al., **2020**), a researcher may also be compelled to complete an ethics transparency checklist (Henry et al., **2018**), a financial conflict of interest checklist (Rochon et al., **2010**), a patient and public involvement checklist (Staniszewska et al., **2017**), and/or a clinical practice guidelines checklist (Cruz Rivera et al., **2020**). When considered in the wider context of an all-round academic job, potential sources of administration multiply even further. Administration is increasing in universities across all elements of an academic role (Hogan, **2011**), all of which compete for importance and time. Minor administrative tasks are constantly introduced to collect data relating to pedagogy, supervision, equality and diversity, technology enhanced learning, financial auditing, health and safety, data protection compliance, employment law, and so on. The cumulative impact of 'trivial' pieces of administration has been described as "death by a thousand 10-minute tasks" (Bozeman et al., **2021**). However, it is only the academic themself that sees the impact of this burden as it is they who need to devote time to completing all of these tasks. The full picture of administrative burden is therefore difficult to detect as it is only visi-

ble when considered *holistically*, a perspective that meta-research and other siloed analyses of individual aspects of academic work often overlook.

Fostering Open Research: Teaching and Supervision

A final way in which open research invisibly adds to workload is through the expectation that academics not only engage in open research themselves but teach and mentor open research practices to junior colleagues, and graduate and undergraduate students. This can be through direct reforms to teaching materials and curriculums, but also through supervision and informal mentoring. Understanding the why and how-to of open research reforms is a big task that necessarily requires knowledge of the philosophy, history, and sociology of science, as well as the practical data science and technological skills discussed previously (Crüwell et al., **2019**). Therefore, instructing graduate and undergraduate students on such topics may require significant reform of existing teaching and supervision practices, which have been described as "largely outdated" (Azevedo et al., **2022**). The issue of updating existing curricula with new knowledge is not one that is unique to research methods, and workloads typically include time to update and rewrite teaching content (although this is often already underestimated). Resources have been developed and shared to reduce this burden (e.g., lesson plans Pownall et al., **2021**), however, the "revolutionary" changes that open research reforms represent (Spellman, **2015**) still make this task considerable and time consuming. Major changes such as shifting to teaching reproducible analysis software like *R* may require significant investments in staff training (e.g. Barr et al., **2019**), again a time sink that is rarely captured in workload models. These issues also apply to the informal mentoring of colleagues and PhD students, work that is typically already neglected in workload models and falls disproportionately on structurally disadvantaged staff (Gordon et al., **2022**). Adding in mentoring of open research skills and knowledge to existing supervision of how to navigate academia and research methods again represents additional activities that are practically time consuming, but that are not reflected in workload models.

Discussion

The root of the issue of workload models is that as models, they are by definition "simulations of measurement" of academic work rather than accurate records of real-world labour, but their use *precedes* the reality in discussions and expectations of the work of academics (Papadopoulos, **2017**). By erroneously "measuring" the reality of the time it takes to perform certain activities or ignoring others entirely, expectations of academic work are unrealistic, yet from a managerial perspective, justified. Open research practices represent a novel type of academic labour with high potential to be mismeasured or made invisible by workload models, raising expectations to even more unrealistic levels.

The actual additional workload of open research is highly dependent on the type of research and the specific practices adopted. Whilst the time burden of tasks such as data sharing or teaching open research may be clear and significant, others such as administration, checklists, or preregistration may be deceptively trivial. However, such trivial tasks can easily add up and multiply, and are thus much more likely to 'creep' into workloads undetected. Taken together, the additional workload required for openness could therefore easily consume any time "saved" by efficiency gains from open workflows (Lowndes et al., **2017**).

The benefits of open research practices have been widely discussed (Munafò et al., **2017**), and generally speaking researchers have positive attitudes towards adopting open research practices, finding them worthwhile (Eynden et al., **2016**; Lowndes et al., **2017**), and recommending their use to others (Sarafoglou et al., **2022**). However, high time cost is repeatedly identified as one of the main barriers to the adoption of open practices (Eynden et al., **2016**; Gownaris et al., **2022**; Tenopir et al., **2011**) and time is the main thing that researchers lack, with workloads at capacity across the sector, having already reached "untenable" levels (Long et al., **2020**; Papadopoulos, **2017**). Historically, institutions have responded to increases in workload and administration by implementing managerial practices such as workload models that aim to increase efficiency by raising expectations of what researchers

should achieve in their existing available time. This has led to endemic levels of stress and mental health issues across academia, with unrealistic expectations and excessive workload cited as the primary causes (Nicholls et al., **2022**; Urbina-Garcia, **2020**).

Without intervention, there is currently no reason to expect that the additional workload required by open research practices will not follow the same pattern of being integrated into existing workloads without a sufficient increase in time available, and thus exacerbate the crisis. The theory of academic capitalism suggests that the responsibility for addressing potential discrepancies between modelled and actual workload will not be taken up voluntarily by university management, who may at best ignore such issues, or at worse tacitly approve of them as a form of capitalist efficiency (Lyons & Ingersoll, **2010**). In other words, if university management can choose not to incorporate open research into workload models, then they won't.

Acknowledging the implications of operating in a capitalist academic system presents a dilemma for open research advocates looking to improve the quality of research without exacerbating existing issues with working conditions. Fairly integrating expectations and incentives to conduct open research into a system which already exploits academic labour is a difficult task, and good intentions on a systematic level can have perverse individual outcomes. On a systematic level it is certainly a desirable outcome if open research is rewarded, thus helping to position responsible researchers into long-term careers and raising the quality of research across the board. But on an individual level it is not a desirable outcome for a researcher already working at maximum capacity and at risk of burnout to be expected to perform extra tasks in order to be able to achieve or retain secure employment. Both of these outcomes can co-occur and uncritical progress towards the former may inadvertently trigger the latter. This duality has implications for both the responsibility of individuals and the design of systems in promoting open research, which I explain below.

Implications for Individuals

First, proponents of open research reforms must acknowledge how the extra work of conducting open research may be practically accommodated in a researcher's existing workload. Although some have attempted to do this (e.g. Robson et al., **2021**), elsewhere the issue is neglected or downplayed. Suggestions that concerns about the workload of open research are an example of a "myth" (Bastiaansen, **2019**) or a "misconception" that can be corrected by "positive advocacy" (Hagger, **2022**) are unhelpful to having honest conversations about the practical negative consequences of conducting open research. Claims that open practices such as sharing resources *reduce* workload (Grahe et al., **2020**) only reflect the perspective of those utilising shared resources, and not those involved in doing the sharing. Efficiency gains from open research which nominally "save time" (Lowndes et al., **2017**) may be inconsequential if open research also results in increased expectations of open outputs in the form of workload creep.

Second, open research advocates should not promote open research practices uncritically. Despite the benefits, all open research practices have an accompanying cost of time, which is a rare and increasingly depleted resource. Even trivial administrative tasks can have a cumulative impact. Whilst it may not be possible to accurately predict potential time costs and benefits in advance of proposing or promoting a new open research initiative, rigorous meta-research should be planned and conducted to evaluate the actual benefits and costs of open research practices. For example, research has investigated the impact of preregistrations on researcher workflow (Sarafoglou et al., **2022**). If a practice is shown to have minimal actual benefit (e.g., if transparency statements go largely unread, or preregistrations fail to prevent researcher bias) then their continued promotion should be re-evaluated or discontinued.

Third, open research advocates, particularly those with influence in universities, should engage more directly with issues of academic labour (Callard, **2022**; Hostler, **2022**). When promoting open research reforms in universities or conducting open practices themselves they should advocate not only for investments in open research infrastructure and training but to receive extra time in workload allocations to acknowledge the additional burden of

open research. Advocates writing about issues of systems and incentives should familiarize themselves with literature on academic capitalism (Jessop, **2018**) and issues of workload modelling (Papadopoulos, **2017**) and acknowledge and address the implications of policy suggestions on workload. They should listen to and support workers' rights groups and trade unions in academia and take a broader interest in how changes to research infrastructure and practice can have negative effects on labour conditions, and take a more holistic view of researchers as part of an increasingly troubled and discontented higher education sector.

Implications for Systems

One proposed solution to the issue of the extra work required for open research is a move to team-based research, and the use of specialists to support academics with open research requirements (A. J. Stewart et al., **2021**). It is a sensible suggestion and one that is likely to be amenable to institutions as it dovetails with many aspects of the 'post-academic' reorganization of research in universities (Ziman, **2000**). In several places it has already been implemented, with an increase in university research professionals and support services to help with open research practices (Carter et al., **2019**), which can directly reduce workload for academics. However, it is a long-term solution, and the availability of such specialist support is not yet consistent across the sector (S. L. K. Stewart et al., **2022**). There is also much work to be done to fairly embed such roles in the infrastructure of university research and to ensure such specialist labour is not exploited itself, and is appropriately funded and rewarded (Bennett et al., **2022**).

Where institutions do implement policies and mandates for current academics to practice open research, these should be designed to minimize unnecessary "red tape" in the form of administration that does not aid the reform's *functional objectives*. This requires a clear understanding and explanation of what the functional objective of the reform actually is, which itself requires evaluating the reforms' axiological position and benefits (Uygun Tunç et al., **2022**). For many reforms (e.g., preregistration or data sharing for certain types or programs of research), such arguments may be contested (e.g., Szollosi et al., **2020**), making suggestions that preregistration should be mandatory for publication (Aguinis et al., **2020**) a clear example of potential 'red tape'. Designing systems to minimize red tape is not an easy task, as both standardization (which does not accommodate non-standard research), as well as diversity (which can result in inconsistent and confusing nomenclature) can be potential sources of administrative burden. In addition, systems should aim to minimize rule redundancy, for example by *replacing* research protocols with preregistration documents during ethical review, to avoid unnecessary duplication of similar documents for different purposes. Policies and interventions should also carefully consider whether administration for the purposes of generating (meta)data about (open) research activities is justified, which whilst *potentially* useful for a number of reasons, may not be worth the cost of the added burden.

If institutions decide to devote funding and resources to support open research, then these should where possible be directed to activities and interventions which practically reduce the time-burden of conducting open research. This includes using funding to provide dedicated workload hours for open research practices, as well as revaluating research performance targets to acknowledge that conducting open research may take significantly longer. Any changes to workload or research expectations should be developed and implemented in consultation with academic staff, made fully transparent, and be flexible and under continuous review (Kenny & Fluck, **2022**). Training and guidance for open research should be as targeted as possible (i.e., disciplinary and methodologically specific) to reduce the time-burden on researchers for interpreting and applying such guidance to their own projects. Grassroots open science communities have an important role in developing and delivering such training and mentoring in an accessible way (Armeni et al., **2021**), although those involved in fostering these communities should also be appropriately workloaded and resourced for these tasks by their institutions.

The broader systems within which universities operate and compete should also be considered in the promotion of open research. Universities respond to market drivers; if a funder requires research data to be made open,

then an astute university will invest in infrastructure to support researchers to do this to ensure continued access to the funding. If university rankings rewarded open research, a competitive university would try to ensure its research outputs performed well on these metrics. By this logic, if funders and rankers valued or mandated fair workloads and working conditions for academics as a condition of eligibility, universities would be inclined to adapt to remain competitive in this new environment.

Finally, it should be remembered that universities are complex, multifaceted, organizations and not solely capitalist actors. They have many competing and sometimes conflicting interests, which can ebb and flow depending on the current climate and conditions. Nevertheless, they are nearly always *strategic*, and direct intervention is often required by those advocating for change to highlight the benefits of one particular priority (e.g., staff wellbeing) over the costs of another (time spent on open research). This has implications for the way in which initiatives relating to promoting open research are designed, described, and advocated for.

Conclusion

The uncritical promotion of expectations to conduct open research within a framework of academic capitalism may inadvertently increase workload for researchers at a time when demands on time are already excessive and academics are struggling to cope. It is neither the specific role nor within the capability of open research advocates to tackle the root causes of workload issues, but they must be aware of the potential implications of their calls for systemic changes in incentives for open research. Understanding how academic labour is organized and viewing universities through the lens of academic capitalism can help open research advocates to promote open research practices in responsible and sustainable ways.

Acknowledgements

My initial work on this topic was conducted whilst undertaking an MA in Higher Education with the University Teaching Academy at Manchester Metropolitan University and I would like to thank my supervisor Bernard Lisewski for his support during this. I would also like to thank Yael Benn for her comments on an initial draft of this manuscript.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., … Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*(1), 4–6. https://doi.org/10.1038/s41562-019-0772-6 (see pp. 26, 27).

Aguinis, H., Banks, G. C., Rogelberg, S. G., & Cascio, W. F. (2020). Actionable recommendations for narrowing the science-practice gap in open science. *Organizational Behavior and Human Decision Processes*, *158*, 27–35. https://doi.org/10.1016/j.obhdp.2020.02.007 (see pp. 25, 30).

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), Article e3000246. https://doi.org/10.1371/journal.pbio.3000246 (see p. 21).

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*(4), 437–461. https://doi.org/10.1007/s11948-007-9042-5 (see p. 24).

Armeni, K., Brinkman, L., Carlsson, R., Eerland, A., Fijten, R., Fondberg, R., Heininga, V. E., Heunis, S., Koh, W. Q., Masselink, M., Moran, N., Ó Baoill, A., Sarafoglou, A., Schettino, A., Schwamm, H., Sjoerds, Z., Teperek, M., van den Akker, O. R., van 't Veer, A., & Zurita-Milla, R. (2021). Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, *48*(5), 605–611. https://doi.org/10.1093/scipol/scab039 (see p. 30).

Azevedo, F., Liu, M., Pennington, C. R., Pownall, M., Evans, T. R., Parsons, S., Elsherif, M. M., Micheli, L., Westwood, S. J., & Framework for Open and Reproducible Research Training (FORRT). (2022). Towards a culture of open scholarship: The role of pedagogical communities. *BMC Research Notes*, *15*(1), Article 75. https://doi.org/10.1186/s13104-022-05944-1 (see p. 28).

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), Article e3000937. https:/ / doi.org/ 10.1371/ journal.pbio.3000937 (see p. 25).

Barr, D., Cleland Woods, H., DeBruine, L., Lai, R., McAleer, P., McNee, S., Nordmann, E., Paterson, H., & Stack, N. (2019). Redesigning methods cur-

ricula for reproducibility. https://psyteachr.github.io/sips2019/ (see p. 28).

Bastiaansen, J. (2019). *10 open science myths – Open Science Community Groningen*. Open Science Community Groningen. https://opensciences-groningen.nl/ 10-open-science-myths/ (see p. 29).

Beatson, N. J., Tharapos, M., O'Connell, B. T., Lange, P., Carr, S., & Copeland, S. (2021). The gradual retreat from academic citizenship. *Higher Education Quarterly*, *76*(4), 715–725. https://doi.org/10.1111/hequ.12341 (see p. 24).

Bennett, A., Garside, D., Gould van Pragg, C., Hostler, T. J., Kherroubi Garcia, I., Plomp, E., Schettino, A., Teplitzky, S., & Ye, H. (2022). A manifesto for rewarding and recognising Team Infrastructure Roles. *Research Equals*. https://doi.org/10.53962/knm3-bnvx (see p. 30).

Boncori, I., Bizjak, D., & Sicca, L. M. (2020). Workload allocation models in academia: A panopticon of neoliberal control or tools for resistance? *Tamara*, *18*(1), 51–69. https://doi.org/DOI: (see p. 23).

Borgerud, C., & Borglund, E. (2020). Open research data, an archival challenge? *Archival Science*, *20*, 279–302. https://doi.org/10.1007/s10502-020-09330-3 (see p. 27).

Bozeman, B. (1993). A theory of government 'red tape'. *Journal of Public Administration Research and Theory*, *3*(3), 273–303 (see p. 26).

Bozeman, B., & Jung, J. (2017). Bureaucratization in academic research policy: What causes it? *Annals of Science and Technology Policy*, *1*(2), 133–214. https://doi.org/10.1561/110.00000002 (see p. 26).

Bozeman, B., & Youtie, J. (2020). Robotic bureaucracy: Administrative burden and red tape in university research. *Public Administration Review*, *80*(1), 157–162. https://doi.org/10.1111/puar.13105 (see p. 26).

Bozeman, B., Youtie, J., & Jung, J. (2021). Death by a thousand 10-minute tasks: Workarounds and noncompliance in university research administration. *Administration & Society*, *53*(4), 527–568. https://doi.org/10.1177/0095399720947994 (see p. 27).

Callard, F. (2022). Replication and reproduction: Crises in psychology and academic labour. *Review of General Psychology*, *26*(2), 199–211. https://doi.org/10.1177/10892680211055660 (see pp. 21, 29).

Carter, S., Carlson, S., Crockett, J., Falk-Krzesinski, H. J., Lewis, K., & Walker, B. E. (2019). The role of research development professionals in supporting team science. In K. L. Hall, A. L. Vogel, & R. T. Croyle (Eds.), *Strategies for team science success* (pp. 375–388). Springer International Publishing. https://doi.org/10.1007/978-3-030-20992-6_28 (see p. 30).

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7 (see p. 22).

Collini, S. (2012). *What are universities for?* Penguin. (See p. 23).

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science: An annotated reading list. *Zeitschrift Für Psychologie*, *227*(4), 237–248. https://doi.org/10.1027/2151-2604/a000387 (see p. 28).

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., SPIRIT AI, & CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The spirit-ai extension. *The Lancet Digital Health*, *2*(10), 549–560. https://doi.org/10.1016/S2589-7500(20)30219-3 (see p. 27).

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and open problems of legal document anonymization. *Symmetry*, *13*(8), 1490. https://doi.org/10.3390/sym13081490 (see p. 25).

Dollinger, M. (2020). The projectification of the university: Consequences and alternatives. *Teaching in Higher Education*, *25*(6), 669–682. https://doi.org/10.1080/13562517.2020.1722631 (see p. 23).

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586 (see p. 22).

Eynden, V. V. D., Knight, G., Vlad, A., Radler, B., Tenopir, C., Leon, D., Manista, F., Whitworth, J., & Corti, L. (2016). Survey of wellcome researchers and their attitudes to open research. *Wellcome Trust*. https://doi.org/10.6084/M9.FIGSHARE.4055448.V1 (see p. 28).

Fecher, B., & Friesike, S. (2014). Open science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening science* (pp. 17–47). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_2 (see p. 21).

Fernández Pinto, M. (2020). Open science for private interests? How the logic of open science contributes to the commercialization of research. *Frontiers in Research Metrics and Analytics*, *5, Article 588331*, Article 588331. https://doi.org/10.3389/frma.2020.588331 (see p. 26).

Forscher, P. S., Wagenmakers, E.-J., Coles, N. A., Silan, M. A. A., Dutra, N. B., Basnight-Brown, D., & IJzerman, H. (2022). The benefits, barriers, and risks of big team science. *PsyArXiv*. https://doi.org/10.31234/osf.io/2mdxh (see p. 27).

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 pp. 22, 25, 27).

Gärtner, A., Leising, D., & Schönbrodt, F. D. (2022). Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology. https://doi.org/10.31234/osf.io/5yexm (see p. 24).

Gewin, V. (2022). Has the 'great resignation' hit academia? *Nature*, *606*(7912), 211–213. https://doi.org/10.1038/d41586-022-01512-6 (see p. 22).

Gordon, H. R., Willink, K., & Hunter, K. (2022). Invisible labor and the associate professor: Identity and workload inequity. *Journal of Diversity in Higher Education*. https://doi.org/10.1037/dhe0000414 (see p. 28).

Gownaris, N. J., Vermeir, K., Bittner, M.-I., Gunawardena, L., Kaur-Ghumaan, S., Lepenies, R., Ntsefong, G. N., & Zakari, I. S. (2022). Barriers to full participation in the open science life cycle among early career researchers. *Data Science Journal*, *21*(1), 2. https://doi.org/10.5334/dsj-2022-002 (see p. 28).

Grahe, J. E., Cuccolo, K., Leighton, D. C., & Cramblet Alvarez, L. D. (2020). Open science promotes diverse, just, and sustainable research and educational outcomes. *Psychology Learning & Teaching*, *19*(1), 5–20. https://doi.org/10.1177/1475725719869164 (see p. 29).

Hagger, M. S. (2022). Developing an open science 'mindset'. *Health Psychology and Behavioral Medicine*, *10*(1), 1–21. https://doi.org/10.1080/21642850.2021.2012474 (see p. 29).

Havard, M., Cho, M. K., & Magnus, D. (2012). Triggers for research ethics consultation. *Science Translational Medicine*, *4*(118). https://doi.org/10.1126/scitranslmed.3002734 (see p. 22).

Hefce, R., Universities UK, & Wellcome Trust. (2016). Concordant on open research data. https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/ (see p. 23).

Henry, B. M., Vikse, J., Pekala, P., Loukas, M., Tubbs, R. S., Walocha, J. A., Jones, D. G., & Tomaszewski, K. A. (2018). Consensus guidelines for the uniform reporting of study ethics in anatomical research within the framework of the anatomical quality assurance (aqua) checklist: Framework of the aqua checklist. *Clinical Anatomy*, *31*(4), 521–524. https://doi.org/10.1002/ca.23069 (see p. 27).

Hogan, J. (2011). Is higher education spending more on administration and, if so, why? *Perspectives: Policy and Practice in Higher Education*, *15*(1), 7–13. https://doi.org/10.1080/13603108.2010.532316 (see p. 27).

Holcombe, A. (2019). Contributorship, not authorship: Use CRediT to indicate who did what. *Publications*, *7*(3), 48. https://doi.org/10.3390/publications7030048 (see p. 26).

Hostler, T. (2022). Open research reforms and the capitalist university's priorities and practices: Areas of opposition and alignment. *SocArXiv*. https://doi.org/10.31235/osf.io/r4qgc (see pp. 22, 23, 29).

Jessop, B. (2018). On academic capitalism. *Critical Policy Studies*, *12*(1), 104–109. https://doi.org/10.1080/19460171.2017.1403342 (see pp. 22, 30).

Kenny, J., & Fluck, A. E. (2019). Academic administration & service workloads in australian universities. *Australian Universities Review*, *61*(2), 21–30 (see p. 24).

Kenny, J., & Fluck, A. E. (2022). Emerging principles for the allocation of academic work in universities. *Higher Education*, *83*(6), 1371–1388. https://doi.org/10.1007/s10734-021-00747-y (see pp. 23, 30).

Kernohan, D. (2019). A beginner's guide to academic workload modelling. https://wonkhe.com/blogs/a-beginners-guide-to-academic-workload-modelling/ (see p. 23).

Kolsaker, A. (2008). Academic professionalism in the managerialist era: A study of english universities. *Studies in Higher Education*, *33*(5), 513–525. https://doi.org/10.1080/03075070802372885 (see p. 23).

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278–292. https://doi.org/10.1177/1745691614528520 (see p. 21).

Levin, N., & Leonelli, S. (2017). How does one "open" science? questions of value in biological research. *Science, Technology, & Human Values*, *42*(2), 280–305. https://doi.org/10.1177/0162243916672071 (see p. 26).

Long, D. W., Barnes, A. P. L., Northcote, P. M., & Williams, P. T. (2020). Accounting academic workloads: Balancing workload creep to avoid depreciation in the higher education sector. *Education, Society and Human Studies*, *1*(2), 55. https://doi.org/10.22158/eshs.v1n2p55 (see pp. 22, 24, 28).

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(6), Article 0160. https://doi.org/10.1038/s41559-017-0160 (see pp. 21, 28, 29).

Lyons, M., & Ingersoll, L. (2010). Regulated autonomy or autonomous regulation? Collective bargaining and academic workloads in australian universities. *Journal of Higher Education Policy and Management*, *32*(2), 137–148. https://doi.org/10.1080/13600800903440592 (see p. 29).

Macfarlane, B. (2011). The morphing of academic practice: Unbundling and the rise of the para-academic. *Higher Education Quarterly*, *65*(1), 59–73. https://doi.org/10.1111/j.1468-2273.2010.00467.x (see p. 23).

Miller, J. (2019). Where does the time go? An academic workload case study at an australian university. *Journal of Higher Education Policy and Management*, *41*(6), 633–645. https://doi.org/10.1080/1360080X.2019.1635328 (see pp. 23, 24).

Munafò, M. (2019). Raising research quality will require collective action. *Nature*, *576*(7786), 183–183. https://doi.org/10.1038/d41586-019-03750-7 (see p. 22).

Munafò, M., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C., 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 pp. 25, 28).

Münch, R. (2014). *Academic capitalism: Universities in the global struggle for excellence*. Routledge. (See p. 22).

Nicholls, H., Nicholls, M., Tekin, S., Lamb, D., & Billings, J. (2022). The impact of working in academia on researchers' mental health and well-being: A systematic review and qualitative meta-synthesis. *PLOS ONE*, *17*(5), Article e0268890. https://doi.org/ 10.1371/ journal.pone.0268890 (see pp. 22, 29).

Nosek, B. A. (2019). *Strategy for culture change*. Centre for Open Science. https://www.cos.io/blog/strategy-for-culture-change (see p. 24).

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009 (see p. 25).

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. 25, 27).

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058 (see p. 22).

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237. https://doi.org/10.1177/2515245920918872 (see pp. 25, 26).

O'Leary, E., Seow, H., Julian, J., Levine, M., & Pond, G. R. (2013). Data collection in cancer clinical trials: Too much of a good thing? *Clinical Trials*, *10*(4), 624–632. https://doi.org/10.1177/1740774513491337 (see p. 27).

Pagliaro, M. (2021). Purposeful evaluation of scholarship in the open science era. *Challenges*, *12*(1), Article 6. https://doi.org/10.3390/challe12010006 (see p. 23).

Papadopoulos, A. (2017). The mismeasure of academic labour. *Higher Education Research & Development*, *36*(3), 511–525. https://doi.org/10.1080/07294360.2017.1289156 (see pp. 23, 28, 30).

Pownall, M., Azevedo, F., Aldoh, A., Elsherif, M., Vasilev, M., Pennington, C. R., Robertson, O., Tromp, M. V., Liu, M., Makel, M. C., Tonge, N., Moreau, D., Horry, R., Shaw, J., Tzavella, L., McGarrigle, R., Talbot, C., Parsons, S., & FORRT. (2021). Embedding open and reproducible science into teaching: A bank of lesson plans and resources. *Scholarship of Teaching and Learning in Psychology*. https:/ / doi.org/ 10.1037/ stl0000307 (see p. 28).

Rees, T. (2015). Developing a research strategy at a research intensive university: A Pro Vice Chancellor's perspective. In R. Dingwall & M. McDonnell (Eds.), *The sage handbook of research management* (pp. 565–580). SAGE Publications Ltd. (See p. 23).

Reimer, T. (2014). Imperial College London submission to the RCUK review on open access. https://doi.org/10.25561/15558 (see p. 24).

Research Consulting. (2014). Counting the costs of open access. http://www.researchconsulting.co.uk/wp-content/uploads/2014/11/Research-Consulting-Counting-the-Costs-of-OA-Final.pdf (see p. 24).

Rhys Evans, T., Branney, P., Clements, A., & Hatton, E. (2021). Improving evidence-based prac-

tice through preregistration of applied research: Barriers and recommendations. *Accountability in Research*, *30*(2), 88–108. https://doi.org/10.1080/08989621.2021.1969233 (see p. 27).

Robson, S. G., Baum, M. A., Beaudry, J. L., Beitner, J., Brohmer, H., Chin, J. M., Jasko, K., Kouros, C. D., Laukkonen, R. E., Moreau, D., Searston, R. A., Slagter, H. A., Steffens, N. K., Tangen, J. M., & Thomas, A. (2021). Promoting open science: A holistic approach to changing behaviour. *Collabra: Psychology*, *7*(1), Article 30137. https://doi.org/10.1525/collabra.30137 (see pp. 21, 24, 29).

Rochon, P. A., Hoey, J., Chan, A.-W., Ferris, L. E., Lexchin, J., Kalkar, S. R., Sekeres, M., Wu, W., Van Laethem, M., Gruneir, A., Maskalyk, J., Streiner, D. L., Gold, J., Taback, N., & Moher, D. (2010). Financial Conflicts of Interest Checklist 2010 for clinical research studies. *Open Medicine: A Peer-Reviewed, Independent Open-Access Journal*, *4*(1), 69–91 (see p. 27).

Rockwell, S. (2009). The FDP Faculty Burden Survey. *Research Management Review*, *16*(2), 29–44 (see p. 27).

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), Article 211997, 211997. https://doi.org/10.1098/rsos.211997 (see pp. 28, 29).

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative Research*, *15*(5), 616–632. https://doi.org/10.1177/1468794114550439 (see p. 25).

Scheliga, K., & Friesike, S. (2014). Putting open science into practice: A social dilemma? *First Monday*. https:/ / doi.org/ 10.5210/ fm.v19i9.5381 (see p. 21).

Schneider, C. E. (2015). *The censor's hand: The misregulation of human-subject research*. MIT Press. (See p. 27).

Seidl, A., Wrzaczek, S., El Ouardighi, F., & Feichtinger, G. (2016). Optimal career strategies and brain drain in academia. *Journal of Optimization Theory and Applications*, *168*(1), 268–295. https://doi.org/10.1007/s10957-015-0747-3 (see p. 22).

Somerville, L. H. (2021). Learn when—and how—to say no in your professional life. *Science*. https://doi.org/ 10.1126/ science.caredit.abg4310 (see p. 24).

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. 28).

Staniszewska, S., Brett, J., Simera, I., Seers, K., Mockford, C., Goodlad, S., Altman, D. G., Moher, D., Barber, R., Denegri, S., Entwistle, A., Littlejohns, P., Morris, C., Suleman, R., Thomas, V., & Tysall, C. (2017). GRIPP2 reporting checklists: Tools to improve reporting of patient and public involvement in research. *Research Involvement and Engagement*, *3*(1), 13. https://doi.org/10.1186/s40900-017-0062-2 (see p. 27).

Stewart, A. J., Farran, E. K., Grange, J. A., Macleod, M., Munafò, M., Newton, P., Shanks, D. R., & the UK Reproducibility Network (UKRN) Local Network Leads. (2021). Improving research quality: The view from the UK Reproducibility Network Institutional Leads for research improvement. *BMC Research Notes*, *14*(1), 458. https://doi.org/10.1186/s13104-021-05883-3 (see pp. 21, 30).

Stewart, D. J., Batist, G., Kantarjian, H. M., Bradford, J.-P., Schiller, J. H., & Kurzrock, R. (2015). The urgent need for clinical research reform to permit faster, less expensive access to new therapies for lethal diseases. *Clinical Cancer Research*, *21*(20), 4561–4568. https://doi.org/10.1158/1078-0432.CCR-14-3246 (see p. 27).

Stewart, P. M., Stears, A., Tomlinson, J. W., & Brown, M. J. (2008). Regulation - the real threat to clinical research. *BMJ*, *337, Article a1732*, Article a1732. https://doi.org/10.1136/bmj.a1732 (see p. 22).

Stewart, S. L. K., Pennington, C. R., da Silva, G. R., Ballou, N., Butler, J., Dienes, Z., Jay, C., Rossit, S., Samara, A., & Leads, U. K. R. N. L. N. (2022). Reforms to improve reproducibility and quality must be coordinated across the research ecosystem: The view from the ukrn local network leads. *BMC Research Notes*, *15*(1), Article 58. https://doi.org/10.1186/s13104-022-05949-w (see p. 30).

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 p. 30).

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), Article e21101. https://doi.org/10.1371/journal.pone.0021101 (see p. 28).

Uher, J. (2023). What's wrong with rating scales? psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass, Article*, *e12740*, Article e12740. https://doi.org/10.1111/spc3.12740 (see p. 22).

University & College Union. (2019). *Counting the costs of casualisation in higher education*. https://www.ucu.org.uk/media/10336/Counting-the-c

osts-of-casualisation-in-higher-education-Jun-19/pdf/ucu_casualisation_in_HE_survey_report_Jun19.pdf (see p. 23).

University & College Union. (2022). *Four fights dispute FAQs*. https://www.ucu.org.uk/article/11818/Four-fights-dispute-FAQs (see p. 22).

Urbina-Garcia, A. (2020). What do we know about university academics' mental health? a systematic literature review. *Stress and Health*, *36*(5), 563–585. https:/ / doi.org/ 10.1002/ smi.2956 (see pp. 22, 29).

Uygun Tunç, D., Tunç, M. N., & Eper, Z. B. (2022). Is open science neoliberal? *Perspectives on Psychological Science*, *174569162211148*. https://doi.org/10.1177/17456916221114835 (see pp. 21, 30).

Vican, S., Friedman, A., & Andreasen, R. (2020). Metrics, money, and managerialism: Faculty experiences of competing logics in higher education. *The Journal of Higher Education*, *91*(1), 139–164. https://doi.org/10.1080/00221546.2019.1615332 (see p. 23).

Weitzenboeck, E. M., Lison, P., Cyndecka, M., & Langford, M. (2022). The GDPR and unstructured data: Is anonymization possible? *International Data Privacy Law*, *ipac008*, Article ipac008. https://doi.org/10.1093/idpl/ipac008 (see p. 25).

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832 (see p. 25).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https:/ / doi.org/ 10.1038/ sdata.2016.18 (see p. 25).

Williams, H. (2022). So, if this is going to be done within my usual hours as part of my current role, what would you like me to stop doing and what reassurances can you offer that this won't adversely affect my career prospects? [Tweet]. https://twitter.com/ alrightPET/ status/ 1534785730995789827?s=20&t=qcWB_lHaL_7tEgU6Z3ym0A (see p. 24).

Ziman, J. M. (2000). *Real science: What it is and what it means*. Cambridge University Press. (See p. 30).

Reflections on Preregistration: Core Criteria, Badges, Complementary Workflows

Robert T. Thibault [1,2,3], Charlotte R. Pennington [4], Marcus R. Munafò [2,3]

[1] Meta-Research Innovation Center at Stanford (METRICS), Stanford University.

[2] School of Psychological Science, University of Bristol.

[3] MRC Integrative Epidemiology Unit at the University of Bristol.

[4] School of Psychology, Aston University.

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
November 17, 2022
Accepted
April 12, 2023
Published
May 15, 2023
Issued
May 24, 2024

Correspondence
Stanford University
robert.thibault@stanford.edu

Funding
Robert T. Thibault is supported by a general support grant awarded to METRICS from Arnold Ventures and a postdoctoral fellowship from the Canadian Institutes of Health Research. Marcus Munafo and Robert Thibault are part of the MRC Integrative Epidemiology Unit (MC_UU_00011/7). The funders have no role in the preparation of this manuscript or the decision to publish.

Clinical trials are routinely preregistered. In psychology and the social sciences, however, only a small percentage of studies are preregistered, and those preregistrations often contain ambiguities. As advocates strive for broader uptake and effective use of preregistration, they can benefit from drawing on the experience of preregistration in clinical trials and adapting some of those successes to the psychology and social sciences context. We recommend that individuals and organizations who promote preregistration: (1) Establish core preregistration criteria required to consider a preregistration complete; (2) Award preregistered badges only to articles that meet the badge criteria; and (3) Leverage complementary workflows that provide a similar function as preregistration.

Keywords *badges, blind data analysis, Open Science, prospective registration, preregistration*

Clinical trials are routinely preregistered[1] (Al-Durra et al., **2020**). However, in other fields such as psychology and the social sciences only a small percentage of studies are preregistered,[2] and they often contain ambiguities in the description of their study design, hypotheses, and analysis plans (Bakker et al., **2020**; van den Akker et al., **2022**). As advocates strive for broader uptake and more effective use of preregistration, the research community could benefit from drawing on the success of preregistration in clinical trials, where preregistration is commonplace.[3] Preregistered clinical trials contain itemized and relatively explicit outcome measures, and most report their results.[4]

[1] Clinical trials research uses the term *prospective registration,* whereas other disciplines (including psychology and the social sciences) use the term *preregistration.* Prospective registration of clinical trials differs from preregistration of research in other disciplines in terms of its history, the functions it was designed to serve, and implementation details (e.g., clinical trial registries do not include sections specifically dedicated to hypotheses or analysis plans) (explanation adapted from: TARG Meta-Research Group & Collaborators, **2022**). To streamline the reading of this commentary, we use the term *preregistration* for both clinical trials and other disciplines.

[2] Randomly sampled publications from the years 2014-2017 found that 1-5% (95% confidence intervals) of studies in psychology and 0-1% (95% CI) of studies in the social sciences were preregistered (Hardwicke et al., **2020**; Hardwicke et al., **2021**). Another study found that 5% of over 90,000 articles in political science and international relations published from 2010-2021 were preregistered, but that 16% were preregistered in 2021 (Scoggins & Robertson, **2023**). Other sources also indicate an increase in the prevalence of preregistration since Hardwicke et al.'s 2014-2017 sample. For example, there were 12,000 OSF preregistrations from 2012-2017 (Nosek & Lindsay, **2018**), but over 100,000 across a similar time period from 2018 to January 2023. Some of these preregistrations, however, may be duplicates (e.g., to update a preregistration, the OSF instructs users to create a new preregistration). The prevalence of preregistration in psychology and the social sciences likely remains far below that of clinical trials research.

[3] For example, 92% of UK clinical trials submitting a final report to the NHS Health Research Authority in Sep 2021 to Sep 2022 are registered in a publicly accessible database (NHS Health Research Authority, **2021**); 75% of German clinical trials that were registered in 2017 were done so prospectively (BIH QUEST, **2023**), and 71% of clinical trials published in PubMed-indexed journals in 2018 were registered, although only 42% were registered before the study began (i.e., preregistered) (Al-Durra et al., **2020**).

[4] As of March 2023, for clinical trials that were completed over 12 months ago, 76% of those covered under the Food and Drug Administration Amendments Act and 84% of those posted on the European Union Clinical Trials Reg-

We propose three actions for the research community to consider to improve the function of preregistration in psychology and the social sciences (see Table 1 & Box 1). These proposals stem from insights developed while conducting research on preregistration across disciplines, including meta-analyses of discrepancies between preregistrations and published manuscripts (TARG Meta-Research Group & Collaborators, **2021**) and a feasibility study of a peer review intervention to address these discrepancies before publication (TARG Meta-Research Group & Collaborators, **2022**). We discuss the function of preregistration in terms of reducing bias and making risk of bias transparent (as outlined in Hardwicke & Wagenmakers, **2023**), as well as the auxiliary benefit of improved research quality.[5] Our proposals are by no means exhaustive; more comprehensive overviews of preregistration are available elsewhere (e.g., Hardwicke & Wagenmakers, **2023**; DeVito, **2022**).

We propose that advocates for preregistration consider to:

1. **Establish core preregistration criteria** (i.e., a minimum amount of information required to consider a preregistration complete—as the World Health Organization's International Clinical Trials Registry Platform has done for clinical trial registration).
2. **Award preregistered badges only to articles that meet the badge criteria** (of which few currently do).
3. **Leverage complementary workflows that provide a similar function as preregistration** (e.g., blinded data analysis to minimize data-dependent analytical decisions).

1. Establish core preregistration criteria

For clinical trials to be considered fully registered, they must provide information regarding 24 specific items, known as the Trial Registration Data Set (World Health Organization, **2017**). Although these 24 items do not include a detailed analysis plan, they set a minimal standard that organizations such as the International Committee of Medical Journal Editors can promote (ICMJE, **2022**; ICMJE, **2023**). This itemized standard laid the foundation for regulations and institutional infrastructure, which in turn drove the widespread uptake of preregistration in clinical trials.[6] It allows for transparent updating of preregistrations and makes comparisons between preregistrations and publications relatively easy (see Figure 1). The structure is sufficiently clear-cut, such that the Health Research Authority in the UK now uses information from ethics applications to register trials on behalf of clinical trialists (NHS Health Research Authority, **2021**). These researchers can go beyond the minimum 24 items, add as many details as they would like to the registration, and append a study protocol.[7]

In contrast to clinical trial registrations which include discrete items followed by short responses (e.g., primary outcome; sample size), preregistration templates in psychology and the social sciences often include broad headers followed by blocks of text (e.g., hypotheses, analysis plan—see Figure 1).[8] Single hypotheses can contain multiple elements that would be better divided into several distinct hypotheses. Preregistrations sometimes list several variables and analyses but provide a sample size calculation for only one analysis. Within a preregistration, aligning a single hypothesis

ister (EUCTR) contained a link in their preregistration pointing to the reported results. Notably, however, reported outcomes do not always match preregistered outcomes (see footnote 7).

[5] We acknowledge that there are ongoing debates about the function, purpose, and goals of preregistration (e.g., McPhetres, **2020**). We feel that this debate does not preclude implementation of our suggestions.

[6] For example, in 2005, the International Committee of Medical Journal Editors (ICMJE) made preregistration a necessary condition for consideration for publication (De Angelis et al., **2004**). Clinical trial registration rates quadrupled in 2005 and thousands of journals now claim to follow this policy (ICMJE, **2023**).

[7] Notably, clinical trial preregistration is not a panacea. Some trials are never preregistered; among those that are, about one-third publish at least one different primary outcome than what was preregistered, and about two-thirds publish at least one different secondary outcome than what was preregistered (TARG Meta-Research Group & Collaborators, **2021**). While some researchers challenge the usefulness of clinical trial preregistration based on this situation (e.g., Lash, **2022**; Abrams et al., **2020**), it is trial preregistration that allows us to identify this risk of bias. Without preregistration, we would not be able to quantify the level of selective reporting or publication bias. Moreover, preregistering protocols—which include statistical analysis plans—remains rare in biomedicine broadly (Serghiou et al., **2021**).

[8] Granted, the scope of preregistration must be expanded to capture the breadth of study designs used in psychology and the social sciences, as compared to the more limited design options used in clinical trials.

Table 1 Problems and proposed solutions for preregistration in psychology and the social sciences.

Problem	Proposed solutions*
1. Low uptake of preregistration **2.** Imprecise and ambiguous preregistrations **3.** Poor alignment between preregistrations and manuscripts	**A.** Establish core preregistration criteria **B.** Award preregistered badges only to articles that meet the badge criteria **C.** Leverage complementary workflows that provide a similar function as preregistration

1. Preregistration provides additional benefits once a substantial proportion of studies are preregistered—such as facilitating evidence synthesis and reducing duplication.
2. Imprecise language and ambiguities in preregistrations leave them open to various interpretations, and in turn, limit their ability to reduce bias and transparently communicate study plans.
3. Poor alignment between preregistrations and manuscripts—both in terms of the overall structure of the documents as well as the specific content—make it difficult to compare the texts to assess risk of bias.

The solutions we propose are partial solutions in the sense that they remain unlikely to fully solve shortcomings in preregistration. They can be implemented individually or alongside other efforts.
*We itemize the *problems* with numbers and the *proposed solutions* with letters to indicate that they are not aligned in a one-to-one manner; each proposed solution could impact each of the three problems to different extents.

to its outcome measure and analysis can be far from trivial. Matching these to text in the manuscript presents an additional challenge. Thus, the less structured information provided in many psychology preregistrations can obscure a reader's understanding of what the researchers planned to do and whether they did it.[9,10]

Given the low prevalence of preregistration in psychology and social sciences research (Hardwicke et al., **2020**; Hardwicke et al., **2021**; Scoggins & Robertson, **2023**), alongside the difficulty of comparing preregistered study details to published study reports (TARG Meta-Research Group & Collaborators, **2022**; van den Akker et al., **2022**) we argue that establishing core preregistration criteria would complement ongoing initiatives that strive for ideal practice.[11]

Efforts have been made to create standard preregistration templates in psychological science (e.g., Open Science Framework, AsPredicted), but these can vary substantially, and there is no broad agreement regarding the details they must include. In an attempt toward standardization, a Preregistration Task Force consisting of the American Psychological Association (APA), the British Psychological Society (BPS), and the German Psychological Society (DGP), supported by the Center for Open Science (COS) and the Leibniz Institute for Psychology, developed a consensus template[12] for the preregistration of quantitative psychology re-

[9] We have no qualms when researchers deviate from a preregistration, so long as they disclose the deviation when reporting results. In many cases, deviations are entirely justifiable and can be necessary to improve a study design or analysis.

[10] One complication for preregistration in psychology and the social sciences, as compared to clinical trials, is that the spectrum of research questions ranges from highly exploratory to purely confirmatory. Some ambiguities in preregistrations may arise when researchers attempt to present exploratory research questions in a confirmatory format.

[11] Some researchers propose that publications of preregistered studies should include a section outlining deviations from the preregistration (e.g., Campbell et al., **2019**). While we agree with these initiatives, sections on deviations will be difficult to interpret if preregistrations are highly ambiguous due to a lack of itemization.

[12] Although the authors use the terms *consensus* and *consensus template* throughout their manuscript, there is no description of the consensus process, which may not have been formalized.

Last Update Posted Date	September 29, 2020
Actual Study Start Date ICMJE	September 2014
Actual Primary Completion Date	April 2019 (Final data collection date for primary outcome measure)
Current Primary Outcome Measures ICMJE **(submitted: August 17, 2020)**	Scores of Teacher and Parent Rated Inattentive Symptoms [Time Frame: Assessed at Baseline, Mid-treatment (2 months), End-treatment (4 months), at 6 month follow up, at 13 month follow up] The primary outcome measure is the composite scores of teacher and parent-rated inattentive symptoms on the Conners-3, rated on a scale of 0-3. Lower scores represent a better outcome, with a maximum score of 3 and a minimum score of 0.
Original Primary Outcome Measures ICMJE **(submitted: September 25, 2014)**	Scores of Teacher and Parent Rated Inattentive Symptoms [Time Frame: up to 60 months] The primary outcome measure will be scores on teacher and parent rated inattentive symptoms on the Conners-3.
Change History	Complete list of historical versions of study NCT02251743 on ClinicalTrials.gov Archive Site
Current Secondary Outcome Measures ICMJE	*Not Provided*
Original Secondary Outcome Measures ICMJE	*Not Provided*

Hypotheses

Priming Hypothesis
Evaluations of Obama's overall performance will be more strongly associated with similarity between people's views on gun control and people's perceptions of Obama's views on gun control among people who watch a video that includes content relevant to the gun control issue than among people who watch a video that does not include content relevant to the gun control issue.

Source Similarity Hypothesis
Among people who watch a video that contains content relevant to the gun control issue, evaluations of Obama's overall performance will be more strongly associated with similarity between people's views on gun control and people's perceptions of Obama's views on gun control among people who identify with the Democratic Party than among people who do not identify with the Democratic Party. Among people who do not watch a video that contains content relevant to the gun control issue, the effect of identification with the Democratic Party on evaluations of Obama's overall performance will not be moderated by the similarity between people's views on gun control and people's perceptions of Obama's views on gun control.

Figure 1 **Comparison of a clinical trial preregistration excerpt (left) to an OSF preregistration excerpt (right)**[*]. The clinical trial preregistration excerpt demonstrates several features that the psychology and social sciences community could benefit from considering. These include: (1) The option for an itemized and tabular format; (2) Clear demarcation of the 24 items contained in the WHO Trial Registration Data Set, which is supported by the International Committee of Medical Journal Editors and demarcated with the superscript ICMJE; (3) Clear demarcation of the primary outcome measure and time frame of assessment; (4) Easy identification of updates to the primary outcome measure (e.g., several time points were added); (5) Easy identification of core items that are not provided (e.g., secondary outcome measures); and (6) A link to a *Change History* log, which looks similar to a Microsoft Word document with track changes. The OSF has recently begun to provide a function allowing researchers to update their preregistration. The updated preregistration identifies sections that were updated (e.g., "Hypotheses"), but does not provide the track-changes style functionality that clinicaltrials.gov does. OSF preregistrations also often contain a statistical analysis plan whereas clinical trial preregistrations rarely do.

[*] *These excerpts are copied from (Arnold & DeBeus, **2013**) and (Berent, **2021**). We selected the clinical trials registration because the lead author (RTT) was familiar with it, and it clearly depicts several benefits of clinical trial preregistrations in a relatively small screenshot. The OSF registration was selected by going to **www.osf.io/registries** and selecting* Provider*: "OSF Registries" and* OSF Registration Type*: "OSF Preregistration," and then choosing a recent preregistration that depicts the text-block response format.*

search (the PRP-QUANT template; Bosnjak et al., **2022**). On the one hand, the template is exhaustive and was purposefully designed to parallel the structure of the APA Style Journal Article Reporting Standards (Appelbaum et al., **2018**); its proper use would present a very effective implementation of preregistration. On the other hand, there is no evidence that user-testing informed the template[13] and its uptake remains limited at this time.[14]

Comparable guidelines in clinical trials have been developed through formal consensus processes that involve diverse stakeholders, include a user-testing stage (i.e., piloting), and are widely used by researchers.[15] These documents were designed to apply across clinical trials research, regardless of the specific disci-

[13] Two of the authors are now testing the usability of the PRP-QUANT template (preregistration: Spitzer et al., **2021**).

[14] This manuscript was first posted as a preprint in February 2021 and has received 27 citations (according to Google Scholar on 31 Jan 2023). These citations appear to mostly reference, rather than use, the PRP-QUANT template. If this template serves a broad user base, is user-friendly, and strongly supported by the organizations involved in making the template (some of the leading psychology organizations in the world), we would predict a greater uptake to date, though increased uptake may occur in the years to come. As of April 2023, it appears the template has been used up to 75 times on the PsychArchives repository.

[15] For example, the Consolidated Standards of Reporting Trials (CONSORT), the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) (Hopewell et al., **2022**).

pline. Their structure is such that researchers can create extensions to the guidelines to target their specific disciplines more fully (e.g., traditional Chinese medicine: Zhang et al., **2020**; pilot trials: Thabane et al., **2016**). Core preregistration criteria could be developed through a similar process and designed to accommodate the diversity of study types in psychology and the social sciences. They could facilitate broad adoption of preregistration by setting a minimum standard that is relatively easy to achieve and a benchmark upon which publishers, funders, and institutions can develop regulations.

2. Award preregistered badges only to articles that meet the badge criteria

As of April 2023, the Center for Open Science website lists 80 journals that award badges to articles that claim to have used open science practices such as preregistration, open materials, and open data (www.cos.io/initiatives/badges). To receive a preregistered badge, a publication should have no undisclosed discrepancies from the preregistration (COS, **2023**). And yet, two studies analyzing psychology publications with preregistered badges found that 89% of 27 articles contained at least one undisclosed discrepancy (Claesen et al., **2021**)[16] and 67% of 258 articles selectively reported at least one hypothesis (van den Akker et al., **2022**).[17] The organization that developed the badges—the Center for Open Science—describes two ways to award badges: author self-disclosure or peer review (COS, **2016**).

Disclosure

Some journals, in their instructions for authors, state that they use the self-disclosure method to award badges (e.g., *Psychological Science*, *Journal of Experimental Social Psychology*), but the disclosure statement provided by the COS which these journals use does not align with the preregistered badge criteria. The four criteria for a preregistered badge are: "(1) A public date-time stamped registration is in an institutional registration system; (2) Registration pre-dates the intervention; (3) Registered design corresponds directly to reported design; and (4) Full disclosure of results in accordance with registered plan" (COS, **2023**). However, the disclosure form used by these journals asks authors to complete five disclosure items (COS, **2016**), none of which match the third and fourth badge criteria. Thus, authors can both truthfully complete the disclosure form and not meet the badge criteria. Even if the disclosure items were realigned to match the badge criteria, it remains unclear whether the proportion of badged papers that fully meet all criteria would rise in the absence of a verification mechanism.

Peer review

We are not aware of any journal that systematically peer reviews articles to ensure they meet the criteria for a preregistered badge. Moreover, based on our experience (TARG Meta-Research Group & Collaborators, **2022**) and that of other researchers who have systematically examined publications awarded with a preregistered badge (Olmo van den Akker, personal communication, 2021; Aline Claesen, personal communication, 2019), we feel it is very difficult to confidently state that the "Registered design corresponds directly to reported design" or "Full disclosure of results in accordance with registered plan." Indeed, one study found that researchers could only agree on the number of hypotheses present in 14% of preregistrations (Bakker et al., **2020**). Another study with a strict operationalization of what constituted a hypothesis had 54% agreement between coders regarding the number of hypotheses (van den Akker et al., **2022**). The lack of itemized core preregistration criteria alongside differences in the structure of preregistrations and manuscripts renders many comparisons ambiguous.

One could argue that the issues we present regarding inaccurate awards are outweighed by the benefit that badges may have on the uptake of preregistration. Indeed, the badge criteria were designed to represent a high aspirational standard, rather than setting a minimal bar. However, there is a possibility that badges in their current implementation have negative effects. Given that most articles awarded a

[16] Researchers replicated and extended this study and found comparable results (Weaver & Rehbein, **2022**).

[17] This sample also includes studies without a preregistered badge, although most articles did have a preregistered badge. They sampled from two populations: articles that contained a preregistered badge and articles that earned a Preregistration Challenge Prize from the Center for Open Science. Their manuscript, however, only reports summary results across both samples.

preregistered badge do not fully meet the criteria for earning that badge, awarding badges can create a false impression that rigorous research practices are being used and therefore lend undue trust to studies awarded a preregistered badge. This practice could also have downstream impacts on the trustworthiness of these types of initiatives more broadly.[18]

Changing the criteria for the preregistered badge could be one way to make clearer what the badge signals. They could be revised, for example, to require only the existence of a permanent and public preregistration in a repository that provides a DOI, without requiring that the preregistration was followed. This criterion would be easy to audit[19] and achieves at least one main function that preregistration was designed to address—putting a timestamp on study plans to help demarcate confirmatory research (Nosek et al., **2018**). An additional criterion could demand that the preregistration include all the items in an established core preregistration criteria, as outlined earlier in this commentary. Machine-readable preregistration statements could also be employed to facilitate automated compliance monitoring from funders or institutions. Based on the current badge criteria, a publication whose preregistration has almost no detail could earn a preregistered badge, whereas a publication with a very detailed preregistration and a minor discrepancy should not earn a badge.[20]

Taken together, current practices for awarding preregistered badges reward researchers even if their preregistration is of low quality and aligns poorly with the associated publication. We commend the development and testing of new initiatives; at the same time, we advocate for follow-up and evaluation to investigate whether they work as intended.

[18] For example, as has happened for "data available upon request" statements, where over 90% of requests go unanswered or are declined (Gabelica et al., **2022**).

[19] Auditing could be conducted by any one of various stakeholders—for example, by the journals who reward the badges, the organization who developed them (COS), funders, researchers (e.g., similar to the FDAAA Trials Tracker DeVito et al., **2019**), or institutions (e.g., similar to the Charité Dashboard on Responsible Research).

[20] Although in practice, badges appear to be awarded in both cases. When appropriately disclosed, we have no qualms with discrepancies.

3. Leverage complementary workflows that provide a similar function as preregistration

Preregistration can reduce bias, increase transparency, and may also improve research quality (Hardwicke & Wagenmakers, **2023**; Sarafoglou, Kovacs, et al., **2022**). However, there are no checks and balances to evaluate whether the study outlined in a preregistration is well designed or clearly described (except when using the Registered Reports format; see Chambers & Tzavella, **2022**). Journal policies and peer review can improve the quality of reporting in relation to a preregistration, but they occur too late in the research pipeline to impact the study design or preregistration quality. Complementary research workflows could achieve some of the same functions as preregistration and may come with additional benefits (e.g., blind data analysis, Experimental Design Assistants, protocol peer review).[21]

For observational research, data management organizations could employ workflows that necessitate open research practices. For example, they could provide researchers with a synthetic dataset, which researchers could use to develop an analysis script. The researchers would then run their analysis in a Trusted Research Environment (TRE) where the results are output, the real data remains hidden, and the analysis is logged and made public (e.g., as done at OpenSAFELY.org). If a Trusted Research Environment is not available, data management organizations could simply provide the complete dataset after the researchers register their analysis script (e.g., as done in Sarafoglou, Hoogeveen, & Wagenmakers, **2022**; and surveyed in Thibault et al., **2023**). These workflows make executable analyses—as opposed to sometimes ambiguous blocks of text—publicly available, while also protecting researchers from making data-dependent analytical decisions. They also overcome arguments raised against preregistration for observational research, including that the data often already exist, knowledge of the data may be necessary to devise a reasonable analysis plan, and registration can inhibit exploration (Lash & Vandenbroucke, **2012**).

[21] See Srivastava (**2018**) for a more in-depth discussion of various alternatives and complementary strategies to typical preregistration.

Researchers can execute a comparable workflow for experimental studies by writing an analysis script based on simulated data and preregistering it before beginning data collection. In other words, the preregistration would include a results section based on a simulated dataset and the numbers are simply updated after running the analysis on the real data.

Another research tool—the Experimental Design Assistant—can be employed in a similar manner. This web application, developed by The National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs), uses a graphical interface to walk preclinical animal researchers through designing their experiment. Unlike preregistration, the EDA is an interactive tool that prompts users to input thorough information and gives warnings when the input fails to align. It then outputs a PDF which the NC3Rs encourages researchers to append to funding applications sent to their organization (NC3Rs, **2021**). This tool holds the potential to simultaneously help researchers design effective experiments and reduce their workloads by using the PDF output as a component of a preregistration.

These examples hold the potential to increase the uptake of preregistration and improve the precision of preregistrations. They do so by embedding the research pipeline with a user-friendly workflow that documents precise study plans.

Box 1. Agents of change

Our three recommendations target the research community broadly and differ in their implementation pathways. Establishing core preregistration criteria would require coordination across various stakeholders including publishers, funders, institutions, learned societies, researchers, and other end-users of research findings. Such an initiative would take a concerted effort and could gain momentum through a grassroots push from researchers or a top-down mechanism from major funders. In contrast, any journal can improve their own use of preregistered badges, and the organization who created them—the Center for Open Science (COS)—holds the ability to redefine the badge criteria. As for our final recommendation, any individual research group, funder, data management organization, or other stakeholder can explore the use of complementary workflows. Highly successful workflows could then be adopted more widely.

Conclusion

In psychology and the social sciences, preregistrations can reduce bias and improve transparency. At the same time, they remain underused, can lack clarity, and are often difficult to compare directly with their associated publication. Current efforts to promote the uptake of preregistration (e.g., badges) and improve preregistration quality (e.g., the PRP-QUANT template) rely largely on the willingness and scrupulousness of research teams alone. We propose that the research community consider parallel initiatives to simplify and standardize preregistration (e.g., adopt itemized core preregistration criteria), and to leverage complementary workflows that necessitate open research practices.

Contributions

RTT wrote an initial draft. All other authors were involved in relevant discussions and contributed to the final draft.

Funding

Robert Thibault is supported by a general support grant awarded to METRICS from Arnold Ventures and a postdoctoral fellowship from the Canadian Institutes of Health Research. Marcus Munafo and Robert Thibault are part of the MRC Integrative Epidemiology Unit (MC_UU_00011/7). The funders have no role in the preparation of this manuscript or the decision to publish.

Acknowledgements

We thank Gustav Nilsonne, Steven Goodman, Mario Malički, Marton Kovacs, and Lisa Spitzer for feedback on earlier drafts of this commentary.

Competing interests

All other authors declare no conflict of interest.

References

Abrams, E., Libgober, J., & List, J. (2020). Research registries: Facts, myths, and possible improvements. *Artefactual Field Experiments, Article*, *00703*. https://ideas.repec.org//p/feb/artefa/00703.html (see p. 38).

Al-Durra, M., Nolan, R. P., Seto, E., & Cafazzo, J. A. (2020). Prospective registration and reporting of trial number in randomised clinical trials: Global cross sectional study of the adoption of ICMJE and declaration of helsinki recommendations. *BMJ*, *369*, 982. https://doi.org/10.1136/bmj.m982 (see p. 37).

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA. *Publications and Communications Board task force report. The American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191 (see p. 40).

Arnold, L. E., & DeBeus, R. (2013). Double-blind 2-site randomized clinical trial of neurofeedback for ADHD, 02251743. www.clinicaltrials.gov. (see p. 40).

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), 3000937. https://doi.org/10.1371/journal.pbio.3000937 (see pp. 37, 41).

Berent, M. (2021). Candidate priming. https://doi.org/10.17605/OSF.IO/F39KX (see p. 40).

BIH QUEST. (2023). *Charité dashboard on responsible research*. https://quest-dashboard.charite.de/#tabStart (see p. 37).

Bosnjak, M., Fiebach, C. J., Mellor, D., Mueller, S., O'Connor, D. B., Oswald, F. L., & Sokol, R. I. (2022). A template for preregistration of quantitative research in psychology: Report of the joint psychological societies preregistration task force. *American Psychologist*, *77*(4), 602. https://doi.org/10.1037/amp0000879 (see p. 40).

Campbell, L., Harris, K., Flake, J. K., Fried, E. I., Beck, E. D., Struhl, M. K., Etz, A., Lindsay, D. S., Feldman, G., van 't Veer, A., & Vazire, S. (2019). https://osf.io/xv5rp/ (see p. 39).

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7 (see p. 42).

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. 41).

COS. (2016). Badges to acknowledge open practices. *OSF*. https://web.archive.org/web/20230420043737/https://osf.io/tvyxz/wiki/2.%20Awarding%20Badges (see p. 41).

COS. (2023). Badges to acknowledge open practices. *OSF*. https://web.archive.org/web/20230508205525/http://web.archive.org/screenshot/https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ (see p. 41).

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P. M., Schroeder, T. V., Sox, H. C., & Weyden, M. B. V. D. (2004). Clinical trial registration: A statement from the. *International Committee of Medical Journal Editors. New England Journal of Medicine*, *351*(12), 1250–1251. https://doi.org/10.1056/NEJMe048225 (see p. 38).

DeVito, N. J. (2022). *Trial registries for transparency and accountability in clinical research* [Doctoral Thesis]. University of Oxford. (See p. 38).

DeVito, N. J., Bacon, S., & Goldacre, B. (2019). FDAAA trialstracker: A live informatics tool to monitor compliance with FDA requirements to report clinical trial results. *BioRxiv*, *266452*. https://doi.org/10.1101/266452 (see p. 42).

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* (see p. 42).

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2021). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *1745691620979806*. https://doi.org/10.1177/1745691620979806 (see pp. 37, 39).

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nat Hum Behav*, *7*, 15–26. https://doi.org/10.1038/s41562-022-01497-2 (see pp. 38, 42).

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017. *Royal Society Open Science*, *7*(2), 190806. https://doi.org/10.1098/rsos.190806 (see pp. 37, 39).

Hopewell, S., Boutron, I., Chan, A.-W., Collins, G. S., de Beyer, J. A., Hróbjartsson, A., Nejstgaard, C. H.,

Østengaard, L., Schulz, K. F., Tunn, R., & Moher, D. (2022). An update to SPIRIT and CONSORT reporting guidelines to enhance transparency in randomized trials. *Nature Medicine*, 1–4. https://doi.org/10.1038/s41591-022-01989-8 (see p. 40).

ICMJE. (2022). Journals stating that they follow the ICMJE recommendations. https://web.archive.org/web/20230508211128/https://www.icmje.org/icmje-recommendations.pdf (see p. 38).

ICMJE. (2023). Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. https://web.archive.org/web/20230508211643/https://www.icmje.org/journals-following-the-icmje-recommendations/ (see p. 38).

Lash, T. L. (2022). Getting over TOP: Epidemiology. https://journals.lww.com/epidem/fulltext/2022/01000/getting_over_top.1.aspx (see p. 38).

Lash, T. L., & Vandenbroucke, J. P. (2012). Should preregistration of epidemiologic study protocols become compulsory? Reflections and a counterproposal. *EPIDEMIOLOGY*, *23*(2), 184–188. https://doi.org/10.1097/EDE.0b013e318245c05b (see p. 42).

McPhetres, J. (2020). What should a preregistration contain? *PsyArXiv*. https://doi.org/10.31234/osf.io/cj5mh (see p. 38).

NC3Rs. (2021). NC3Rs funding schemes applicant and grant holder handbook. https://www.nc3rs.org.uk/sites/default/files/documents/Funding/Handbook.pdf (see p. 43).

NHS Health Research Authority. (2021). Make it public: Transparency and openness in health and social care research. https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/research-transparency/make-it-public-transparency-and-openness-health-and-social-care-research/ (see pp. 37, 38).

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 p. 42).

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31* (see p. 37).

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *PsyArXiv*. https:/ / doi.org/ 10.31234/ osf.io/6dn8f (see p. 42).

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), 211997. https://doi.org/10.1098/rsos.211997 (see p. 42).

Scoggins, B., & Robertson, M. P. (2023). Measuring transparency in the social sciences. *Political Science and International Relations*, *14* (see pp. 37, 39).

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology*, *19*(3), 3001107. https://doi.org/10.1371/journal.pbio.3001107 (see p. 38).

Spitzer, L., Mueller, S., & Bosnjak, M. (2021). Preregistration: Testing the usability of the psychological research preregistration-quantitative (PRP-QUANT) template (see p. 40).

Srivastava, S. (2018). Sound inference in complicated research: A multi-strategy approach. *PsyArXiv*. https://doi.org/10.31234/osf.io/bwr48 (see p. 42).

TARG Meta-Research Group & Collaborators. (2021). Estimating the prevalence of discrepancies between study registrations and publications: A systematic review and meta-analyses. https://doi.org/10.1101/2021.07.07.21259868 (see p. 38).

TARG Meta-Research Group & Collaborators. (2022). Discrepancy review: A feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. https://doi.org/10.1101/2022.01.18.22269507 (see pp. 37, 38, 39, 41).

Thabane, L., Hopewell, S., Lancaster, G. A., Bond, C. M., Coleman, C. L., Campbell, M. J., & Eldridge, S. M. (2016). Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot and Feasibility Studies*, *2*(1), 25. https://doi.org/10.1186/s40814-016-0065-z (see p. 41).

Thibault, R. T., Kovacs, M., Hardwicke, T. E., Sarafoglou, A., Ioannidis, J. P., & Munafo, M. R. (2023). Reducing bias in secondary data analysis via an explore and confirm analysis workflow (ECAW): A proposal and survey of observational researchers. https://doi.org/10.31222/osf.io/md2xz (see p. 42).

van den Akker, O., van Assen, M. A. L. M., Enting, M., de Jonge, M., Ong, H. H., Rüffer, F., Schoenmakers, M., Stoevenbelt, A. H., Wicherts, J., & Bakker, M. (2022). Selective hypothesis reporting in psychology: Comparing preregistrations and corresponding publications. *MetaArXiv*. https://doi.org/10.31222/osf.io/nf6mq (see pp. 37, 39, 41).

Weaver, E. J., & Rehbein, S. T. (2022). Durchgeführt wie geplant? In *Ein detaillierter vergleich zwischen*

studien und ihren prä-registrierten plänen. https://psycharchives.org/en/item/2462b05e-5d58-426b-8b43-a26556294a32 (see p. 41).

World Health Organization. (2017). WHO trial registration data set (version 1.3.1. https://www.who.int/clinical-trials-registry-platform/network/who-data-set (see p. 38).

Zhang, X., Lan, L., Chan, J. C. P., Zhong, L. L. D., Cheng, C.-W., Lam, W.-C., Tian, R., Zhao, C., Wu, T.-X., Shang, H.-C., Lyu, A.-P., & Bian, Z.-X. (2020). WHO trial registration data set (TRDS) extension for traditional chinese medicine 2020: Recommendations, explanation, and elaboration. *BMC Medical Research Methodology*, *20*(1), 192. https://doi.org/10.1186/s12874-020-01077-w (see p. 41).

Rethinking Transparency and Rigor from a Qualitative Open Science Perspective

Crystal Steltenpohl[1], Hilary Lustick[2], Melanie S. Meyer[3], Lindsay E. Lee[4], Sondra M. Stegenga[5], Laurel Standiford Reyes[6], Rachel L. Renbarger[7]

Discussions around transparency in open science focus primarily on sharing data, materials, and coding schemes, especially as these practices relate to reproducibility. This fairly quantitative perspective of transparency does not align with all scientific methodologies. Indeed, qualitative researchers also care deeply about how knowledge is *produced*, what factors influence the research process, and how to share this information. Explicating a researcher's background and role allows researchers to consider their impact on the research process and interpretation of the data, thereby increasing both transparency and rigor. Researchers may engage in positionality and reflexivity in a variety of ways, and transparently sharing these steps allows readers to draw their own informed conclusions about the results and study as a whole. Imposing a limited, quantitatively-informed set of standards on all research can cause harm to researchers and the communities they work with if researchers are not careful in considering the impact of such standards. Our paper will argue the importance of avoiding strong defaults around transparency (e.g., always share data) and build upon previous work around qualitative open science. We explore how transparency in all aspects of our research can lend itself toward projecting and confirming the rigor of our work.

Keywords *open science, transparency, rigor, qualitative, quantitative*

[1]Dartmouth Center for Program Design and Evaluation
[2]University of Massachusetts, Lowell
[3]Johns Hopkins University
[4]East Tennessee State University
[5]University of Utah
[6]University of Southern Indiana
[7]FHI 360

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
September 16, 2022
Accepted
April 1, 2023
Published
May 14, 2023
Issued
May 24, 2024

Correspondence
Dartmouth Center for Program Design and Evaluation
crystal.n.s.young@dartmouth.edu

The social sciences have been undergoing a credibility revolution, also known as the open science movement, within the last decade. This movement emphasizes greater transparency and openness through specific practices, such as preregistration and replication, and improving the quality and quantity of evidence used in making scientific claims (Vazire, **2018**). The concepts of transparency and rigor are important to these conversations. *Transparency* can be defined as "the obligation to make data, analysis, methods, and interpretive choices underlying their claims visible in a way that allows others to evaluate them" (Moravcsik, **2019**). *Rigor* can be defined as "the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting of results" (National Institutes of Health [NIH], **n.d.**). Within the open science movement, discussions around rigor and transparency have largely come from a positivist, quantitative perspective that focuses on the transparency of outputs, namely open data, open materials, open code, and open access to manuscripts (Hagger, **2019**; Lyon, **2016**; Powers & Hampton, **2019**). For example, Fecher and Friesike (**2014**) describe five schools of open science which focus on (1) creating openly available platforms, tools, and services for scientists; (2) making science accessible for citizens; (3) developing alternative measures of impact; (4) making knowledge freely available for everyone; and (5) making the knowledge creation process more efficient. Notably, their description of open science does not explicitly discuss transparency regarding the decisions researchers make during the re-

search process. More inclusive definitions of open science have emerged over time (e.g., United Nations Educational, Scientific and Cultural United Nations Educational, Scientific and Cultural Organization, **2021**).

Indeed, much of the conversation around transparency in the open science movement has focused on standardizing outputs or processes related to data sharing, such as the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., **2015**), authored by researchers who conduct largely, if not entirely, quantitative research projects. The TOP Guidelines include recommendations for citations, data transparency, analytic methods (code) transparency, research materials transparency, design and analysis transparency, preregistration of studies, preregistration of analysis plans, and replication. These guidelines feature three levels of implementation; lower levels recommend practices or require only statements about a measure of transparency and reproducibility (e.g., articles state whether code is available), while higher levels require authors to engage in a specific practice (e.g., code must be posted to a trusted repository).

When standards are created, it is important to consider both who is at the table and who is not. Open science guidelines fail to account for research based on epistemologies that are not strictly positivist and methods that are not strictly quantitative in nature, such as qualitative and interpretivist approaches. As such, they have unfortunately had limited applicability to those kinds of research. Feminist and qualitative scholars have long maintained that there are multiple ways of understanding, yet evangelists of the open science movement have commonly made assumptions that there is a shared understanding of a specific type of research (e.g., empiricism, deductive reasoning). This problematic assumption halts progress in the integration of perspectives for open science and contributes to possibly ignoring systematically marginalized voices (see Bennett et al., **2022**).

Default standards can be useful because they automate processes and reduce cognitive load when making decisions. However, they can also be dangerous, because if researchers are not making thoughtful and informed decisions, then these defaults could result in individuals and research teams moving forward under assumptions that could ultimately cause harm to research participants (Sakaluk, **2021**; Steltenpohl et al., **2021**). Strict guidelines about data sharing, for example, may create problems for researchers working with qualitative data, which without careful attention to confidentiality safeguards (e.g., pseudonyms, redacting personally identifiable information) may be more identifiable than, for example, Likert-scale responses on a survey. These concerns are also relevant to researchers who have built trust through prolonged engagement with a participant community (Ross et al., **2018**), and those whose research perspective differs from the seemingly common conceptualization of research as being strictly "right" or "wrong" (Lash, **2015**).

Previous work within the qualitative research community suggests it may be helpful to think about the relationship between transparency and rigor (e.g., Billups, **2014**; Davies & Dodd, **2002**; Mill & Ogilvie, **2003**; Rolfe, **2006**) and to examine how transparency in all aspects of research can project and confirm the rigor of qualitative inquiry work. Opening the black box of the research process in this way also allows readers from all educational backgrounds to better understand how research is done, what kinds of decisions are made during a research project, and best practices within the researchers' respective fields. As such, practices that promote transparency and rigor in qualitative science could (and, we would argue, should) be considered in line with the tenets of open science. For example, qualitative researchers are encouraged to provide detailed descriptions of their methods (e.g., sampling, data collection, analysis) to the extent that other researchers could follow their methods with other samples in other contexts (Creswell et al., **2018**). This guidance aligns with guidance for the open science practice of replication, where researchers redo a study to see if the same results arise in a different population or at a different time.

Due to the use of purposive sampling and sociocultural contextual factors, replicating qualitative inquiry methods with different samples in different contexts is not as straightforward as replicating null hypothesis significance testing, which in and of itself is not always straightforward. What may be construed as a "failed" replication (often implying the pre-

vious study's results were "wrong") within quantitative paradigms is often interpreted simply as evidence that warrants further study in qualitative methodologies, because qualitative paradigms do not stem from a hypothesis but from inquiry. In qualitative research, we tend to sample to understand context, rather than to generalize; we think of validity in terms of trustworthiness, rather than replicability. At the end of the day, quantitative, qualitative, and mixed methods researchers share the goal of advancing knowledge in the field through rigorous and transparent processes. However, guidelines intended to promote these processes may have the unintended consequence of becoming gatekeepers that limit the ability of qualitative researchers to publish and obtain funding.

Taking a closer look at "the why" behind these open science practices could help researchers across all methodologies understand the benefits of these practices and aspects of each methodology that should be considered before engaging in the practices. This conversation may not only help qualitative researchers engage in open science practices, but also open doors for quantitative researchers to employ rigor and transparency strategies, such as positionality or reflexivity statements, that have traditionally been used exclusively by qualitative researchers (e.g., Jamieson et al., **2022**). The purpose of the current paper is to reframe the conversation within the open science movement around transparency and rigor to include considerations from qualitative research that may be of use in determining best practices for rigor and transparency in open science. These concerns may or may not be unique to qualitative research; we are only speaking from this perspective as researchers who have engaged with qualitative and mixed methods research.

Common Concerns about the Open Science Views of Transparency and Rigor

Much of the discussion about transparency in open science revolves around data, which is relevant to both qualitative and quantitative research. There are increasing requirements for sharing data, particularly through major funding agencies (e.g., National Institutes of Health [NIH], **2020**). For example, a recent U.S. memo dictates that scientific data from projects funded through federal agencies must be available upon publication of the manuscript unless the data fall under specific limitations (Office of Science and Technology Policy, **2022**).

Additionally, some researchers have argued that secondary analysis of qualitative data holds the potential to relieve the burden on vulnerable populations and community partners who collaborate with researchers (e.g., Ruggiano & Perry, **2019**). It also may help foster new findings through the application of a different lens or through a focus on particular elements of the dataset (Long-Sutehall et al., **2011**). Publicly available qualitative data could potentially be beneficial when individuals or groups are particularly difficult to access (Fielding, **2004**) or when working with groups that might face trauma from multiple rounds of research participation (Ruggiano & Perry, **2019**). Some research has revealed that participants are generally willing to allow researchers to archive and share their qualitative data (e.g., Cummings et al., **2015**; Mozersky et al., **2020**; VandeVusse et al., **2022**) out of a desire to help others and improve research.

Although the TOP guidelines for data sharing do suggest that materials should be stored in a trusted digital repository (Nosek et al., **2015**, page 5), these guidelines do not address potential issues that could arise in the future (e.g., guidelines for updating data; Lash, **2015**) or other concerns around data sharing. Mozersky et al. (**2020**) found that less than four percent of researchers reported sharing qualitative data, largely due to concerns related to breaching participant confidentiality and anticipation of wavering participation in future qualitative studies. The sharing of sensitive and potentially identifiable data, even when approved by participants, comes with immense ethical responsibility, both from the researchers sharing the data and those using the data. The rush to mandate open data leaves many ethical ambiguities unaddressed. Researchers are not routinely trained on how to ethically share data, evident in part from researchers' anecdotes of finding identifiable data on data repositories (e.g., Elson, **2021**). Even when researchers are careful, participants may be more identifiable than they believe, especially in certain circumstances, such as being a part of a small identity

group or discussing place-specific content in interviews (Gow et al., **2020**). It is vital that participants understand how likely they are to be identifiable, especially vulnerable participants or those living in dangerous conditions (Ross et al., **2018**; Small et al., **2014**). In fact, ensuring participants truly understand what data sharing *means* is likely to be one of the largest barriers to ethically sharing data (VandeVusse et al., **2022**).

A more fundamental issue complicating data transparency is the question of what constitutes data. At first glance, it may seem obvious that a qualitative researcher's data would comprise transcripts and direct observational notes. However, qualitative researchers are ideally trained to provide detailed audit trails that incorporate notes from the research process (e.g., reflexivity, field notes, methodological notes), instrument development (e.g., observation protocols), and evidence of how codes, categories, and themes are synthesized. These secondary data may be necessary for observers to fully understand research findings and how conclusions were drawn from the data. Secondary data (e.g., memos, protocols) support the reporting process, but traditional journal space limitations may not provide a way to share these data.

To further complicate matters, many qualitative researchers would posit that while secondary data are a combination of the researcher's perceptions and observations, even primary data, such as interview transcripts, are filtered to some extent through the researcher. This is because, in qualitative research, the researcher is an instrument of both data collection and analysis (e.g., Peredaryenko & Krauss, **2013**). Open science standards do not take these complications into account, nor would many quantitatively-focused researchers and reviewers necessarily know what to do with these types of data, were qualitative researchers mandated to share them. Moving to open science practices and transparency will highlight the need to manage extensive and complicated data and various data types more effectively. For example, data management is often an ignored nuance in longitudinal quantitative data and many studies use "trial and error" to determine best practices (Youngblut et al., **1990**). Data collection and storage declarations must be determined as part of the research design before the study begins, and participants and researchers will need to explore and understand risks. Furthermore, large sets of data can be difficult to maintain and manage–or even remember. Khan et al. (**2018**) determined that when researchers reviewed the study files they had stored in their cloud storage, more than 50% of the files were "forgotten" and 14% were unrecognizable. These results reveal the need for extensive training around perpetual data sharing requirements.

The researcher-as-instrument tradition also complicates discussions around reproducibility (i.e., the ability for another researcher to look at someone's data and reproduce the analyses), one of the key components of rigor as it is currently discussed in the open science movement (NIH, n.d.). Quantitative researchers' focus on reproducibility is often contrary to the tenets of qualitative research, particularly in methodologies aiming to uncover new ways of knowing, such as constructivist and grounded theory approaches. If one understands the researcher as a data collection instrument and a filter through which data is processed, strict quantitative-focused reproducibility becomes less likely—not through misconduct or error, but because ultimately, *people* conduct research, and people are not likely to have exactly the same perspectives. Guidelines that reinforce reproducibility without addressing this tension are not going to be useful for all researchers.

A common refrain within the open science movement is that "preregistration is not a prison," (e.g., DeHaven, **2017**; Mellor, **2021**) which we ultimately think is true in intent. However, the goal of preregistration does appear to be putting guardrails on research processes, to ensure researchers do not significantly deviate from their analysis plans. Indeed, in their development of the useful qualitative preregistration template, Haven et al. (**2020**) describe preregistrations as being useful for qualitative researchers who are involved in some level of testing, where it would be important to communicate hypotheses that were drawn prior to the start of the study. For researchers who do not do this kind of research, however, preregistration may have limited utility *as a method for introducing guardrails to the research process.* Preregistration may, however, be useful

in other contexts, which we describe later. Journal requirements for preregistration, then, may exclude entire areas of research if not carefully crafted.

The narrow view of rigor described by the National Institutes of Health [NIH] (**n.d.**) and others can lead people to assume there is one right way to conduct research. There are numerous decision points in quantitative research design that are susceptible to researcher bias, contextual factors such as community history or current events, and conceptual assumptions that may be a result of different training paradigms. Adhering to only one limited set of standards in the pursuit of rigor could therefore prioritize the positivist position that there could be one objective reality but *only if* the proper, rigorous methods were used, which would limit theory development and the advancement of knowledge generally.

Concerns arise if all qualitative research is subject to quantitative expectations of reproducibility to obtain funding or be published in peer-reviewed journals. This would eliminate important qualitative methodologies, many of which have historically been a foundation for amplifying voices of understudied individuals and groups. Focusing on only one metric, such as replicability or reproducibility, will likely lead to increased oppression of research and knowledge, instead of meeting a primary aim of open science to improve access to information which would enable people to better examine scientific claims. Further, it is our belief that many quantitative researchers would benefit from considering qualitative-born methods for determining rigor and transparency in their own work, which we will now discuss.

Rigorous Transparency through a Qualitative Lens

It is difficult to separate rigor from transparency. One cannot fully determine the rigor of work that is not transparent, and not all transparent work is rigorous. Importantly, reproducibility alone cannot determine the rigor of a study. When discussing rigor, qualitative researchers often focus on trustworthiness, which includes confirmability, credibility, dependability, and transferability (Lincoln & Guba, **1986**; Stahl & King, **2020**). Qualitative researchers establish *confirmability* by providing evidence that the conclusions they have made are drawn from the data, but this evidence relies on attention to credibility, dependability, and transferability (Lincoln & Guba, **1985**; Nowell et al., **2017**; Tobin & Begley, **2004**). Using approaches such as extended engagement, triangulation, and member checking, qualitative researchers can address *credibility* by ensuring alignment between participants' views and how those views are presented in the research report. *Dependability* can be established by explicitly and transparently documenting the research process, which can include codebooks, tables, or figures that show the research process. *Transferability* requires researchers to provide enough information to allow research consumers to make determinations about whether the findings might be relevant to other contexts. A common thread between each of these elements of trustworthiness is transparency. As interest in open science practices grows, it may be useful to reexamine the research process with these concepts in mind and integrate qualitative and quantitative perspectives as appropriate.

Research Processes

Importantly, qualitative researchers are often concerned not just with research results, but also with the process by which those results are produced. This could be an exciting area of growth for many quantitative scholars who are open to learning from qualitative researchers; most, if not all, practices would be just as beneficial for quantitative researchers to employ (see, for example, Hope et al., **2019**). For example, positionality (i.e., an examination of who is doing the interpreting) and reflexivity (i.e., an examination of how researchers will reflect on, consider the impact of, and/or mitigate bias where necessary) are staple concepts in qualitative research courses.

Positionality statements are prose that allow researchers to reflect on their role within the data interpretation process (Clancy, **2013**; Guillemin & Gillam, **2004**; Lazard & McAvoy, **2020**; Makel et al., **2022**; Patton, **2014**; Rooney, **2015**; Savin-Baden & Howell-Major, **2013**). This is an opportunity for researchers to think about and discuss decisions that were made, why they were made, and how that might have impacted the research process. Researchers are encouraged to consider both strengths *and* weaknesses of their approaches. Position-

ality statements do not give researchers permission to conduct a certain kind of research, but rather provide additional context to help readers understand the decisions a research team has made. Many journals do not require positionality statements to go beyond reflecting on one's background *vis a vis* their research topic, but rigorous transparency compels us to confront how we will address potential biases, which is where reflexivity becomes especially important.

Researchers may engage with reflexive practices in a variety of ways, including but not limited to member checking (Caretta, **2016**; Creswell & Miller, **2000**; Goldblatt & Band-Winterstein, **2016**; Lincoln & Guba, **1986**; Stake, **1995**), critical team discussions Mao et al. (**2016**), memos (Birks et al., **2008**), and external audits (Wolf, **2003**). Most researchers do not use all of these methods in every single study, though a qualitative paper with none of these practices would likely raise questions from careful reviewers.

Member checking can help researchers ensure they have collected and are interpreting data in a way that is true to participants' experience (e.g., Caretta, **2016**). Bornstein et al. (**2022**) engaged in member checking by returning to participants after a first round of interviews, sharing the themes the researchers had deduced, and asking participants if these themes were accurate reflections of their experiences. This allowed for deviation, as well as nuance and deeper reflection, from participants. Caretta and Perez (**2019**) note that member checking is especially crucial for establishing validity in qualitative methods, because participants may (and likely will) not all agree on a topic. By understanding the context of their opinions and differences, researchers can approach sharper validity; however, the ever-present potential for disagreement reminds us that no one researcher or participant can or should hold epistemological authority over the data.

Critical team discussions, particularly when colleagues come from a variety of backgrounds, are another practice for sharing epistemological authority. While member checking involves sharing data with actual participants, critical team discussions can involve sharing data among colleagues who either share a background with the participant or, at the very least, can offer a different perspective from that of the research team. (Mao et al., **2016**) designed a practice for critical team discussions in which a group of graduate students met regularly to share dissertation data and provide feedback on each other's interpretations of these data. In this way, the students were able to benefit from others' impressions of their participants and reflect on their own blind spots. The authors call this *critical reflexivity* and recommend it for any researchers conducting qualitative research (e.g., Fook & Askeland, **2007**; Mao et al., **2016**).

Through memos, researchers can incorporate relevant literature, context, or background knowledge to analyze a particular set of qualitative data (e.g., interview transcripts). Similarly, through journaling, researchers record their thoughts and feelings in response to data and determine how they can keep these personal reactions from biasing their data interpretation (Meyer & Willis, **2019**). Thick description (Geertz, **1973**) also allows us to be as context-specific about our data and ourselves as possible, and, thus, be as rigorously transparent as possible. While thick descriptions provide detail, they can be difficult to fit within traditional publication word and page limits. However, with online publications and supplementary opportunities, thick descriptions of nuanced situations (e.g., historical events that may impact analysis for a small number of participants) can be included to provide additional transparency.

External audits, where a research team invites an external expert to review research-generated data, can also enhance rigor by providing an outside examination of the research process and products – including meeting notes, memos, and other reflections – and the extent to which interpretations and/or findings are supported by the data (Creswell & Miller, **2000**; Lincoln & Guba, **1986**; Rodgers & Cowles, **1993**; Wolf, **2003**). These audits can be beneficial by adding to the synthesis of ideas and bringing additional perspectives to a specific research topic. Similarly, a research team may engage with peer debriefing, where they review the research thus far with someone who is familiar with the work but not on the team; the peer reviewer is encouraged to challenge the research team on their findings (Creswell & Miller, **2000**). External audits and

peer debriefing lend themselves to establishing the credibility of the findings and require researchers to be transparent about how they have conducted their research and are thinking about the data.

It is clear to us that there is potential for alignment between current open science practices, such as preregistration, and practices in which qualitative researchers already engage. Haven et al. (**2020**) briefly discuss one such alignment in their development of the qualitative preregistration template, when they say that preregistration could be helpful to "make visible the connections between analytical assumptions, evidence, and decisions that form a particular interpretation of the data" (p. 2). We interpret this to mean that preregistration could be helpful as a means to encourage reflexivity—for researchers to document how they conceptualized their study at the beginning of the project and reflect on those assumptions, perspectives, and potential misunderstandings throughout the rest of the research process. Too often, across all methodologies, the process for the *why* behind study decisions is not documented clearly (Mackenzie & Knipe, **2006**). From the inception of a study, it should be clear what the theoretical framework and paradigm are, since these drive, or should drive, methodology and methods. Specifically,

> *It is the paradigm and research question, which should determine which research data collection and analysis methods (qualitative/quantitative or mixed methods) will be most appropriate for a study. In this way researchers are not quantitative, qualitative or mixed researchers, rather a researcher may apply the data collection and analysis methods most appropriate for a particular research study. (Mackenzie & Knipe, **2006**, page 7-8)*

The preregistration template, then, is an immensely helpful tool in any researcher's toolbox, if it is viewed as a space for clearly documenting the *why* in addition to the *how* of a study, and if journal editors and reviewers understand that qualitative researchers are not using the preregistration as guardrails, but rather as a mirror or window to look back at how the study team conceptualized their research at the beginning of the study. If reviewers punish authors for deviating from what was preregistered – for example, if a researcher changes their interview protocols because they realized participants were understanding a question differently than how they meant it to be understood – preregistration templates will not be useful to qualitative researchers. Two key elements are required to ensure the value of the preregistration process to all researchers, regardless of methodological approach. First, the TOP Guidelines and other guidance on preregistration must clearly outline preregistration requirements for qualitative research in addition to quantitative research so that a positivist framework is not pushed onto qualitative research. Second, reviewers, editors, and other partners in the publication process must be clearly trained in these differences so that qualitative researchers are not inadvertently denied opportunities for publication due to inappropriate requirements, such as requiring hypotheses or denying researchers the ability to add or change data collection methods. Ultimately, preregistration is not the only way to achieve the goal of reflexivity or transparency in the research process; however, if journals subscribe to Level 3 of the TOP guidelines, all studies must be preregistered. This kind of guideline may be less useful to qualitative researchers than a broader guideline to provide an audit trail (or similar concept), which would allow researchers to provide many kinds of evidence, including preregistration and preregistration updates (Corker et al., **2022**) illuminating how they got from study conceptualization to their results.

Data

Although privacy, participant rights, consent, and the potentially identifiable nature of qualitative data must always be at the forefront of data practices, there has been an increasingly detailed discussion of what transparency and rigor mean in qualitative research.

Transparency.

One traditional method for increasing transparency is data sharing. When considering qualitative data sharing, we have ethical obligations to participants regarding *how*, with *whom,* and *when* this data can and should be shared.

Existing recommendations include (a) professional data curation and archiving processes to optimize privacy (*how data is shared*), (b) ensuring necessary materials and details are included in the data storage/archiving to enhance contextual and data understanding (*how data is shared*), and (c) providing graded data access (*when and with whom data is shared*).

Regarding data curation and archiving, the Qualitative Data Repository (QDR) first provides professional curation to optimize de-identification prior to storage. This is an important step, though it can easily be counteracted by the "thick description" characteristic of rigorous contemporary (i.e., interpretivist) qualitative data (Geertz, **1973**). When we consider *when and with whom* data can be shared, we must always balance privacy with transparency. One method that holds promise for this balance is graded sharing access to the data. This means that data are available based on researcher qualifications (e.g. ICPSR, **n.d.**) and the ability to meet security standards (e.g. ICPSR, **n.d.**; Qualitative Data Repository, **2022**). For example, data stored on the Qualitative Data Repository can only be used for research or teaching, but depositors can designate the data as standard access (i.e., all registered users can access), special access (i.e., conditional, depositor-approved, restricted offline, or embargoed access), or depositor-approved access (i.e., QDR staff and the depositor review access requests) which allows researchers to understand how potentially sensitive data may be used ("Access controls," **2020**). Setting gradations for how much access researchers have to data could allow researchers to meet open science requirements without unnecessarily compromising participant anonymity, an important balance to strike when working with sensitive data and with organizations that are protective of their data (e.g., governmental agencies, industry partners). Additionally, researchers can share data with the participants to provide transparency to the community of focus (Humphreys et al., **2021**).

Next, QDR recommends having clear and detailed guidance for archiving qualitative data, such as detailed codebooks, processes, and contextual information. Specifically, they suggest including files and artifacts that help to "document the context in which information was gathered and/or data were created, the collection and generation processes, and (when applicable) how the data were analyzed" (Qualitative Data Repository, **n.d.**, Documentation Files section). This is critical to optimizing future understanding of the larger context foundational to qualitative data and ensuring transparency. Taken together, these techniques for qualitative data sharing are critical for both maintaining open science standards and protecting the identities of research participants.

Rigor

A recent article in the *American Journal of Pharmaceutical Education* defines rigor in qualitative research as "ensuring that the research design, method, and conclusions are explicit, public, replicable, open to critique, and free of bias" (Johnson et al., **2020**). Qualitative research demonstrates rigor not only by sharing raw transcripts, but also—if not primarily—through thick description of data and detailed explication of the research design. Journal word limits often mean these details get squeezed out of articles during the editing process. One promising solution is to produce a supplementary "data paper" (Schöpfel et al., **2019**), which describes data in much more detail than a traditional manuscript. For instance, in a qualitative data paper, a researcher could provide detailed context about the community and historical context in which data were collected, how the data were collected, and any additional information about positionality and reflexivity that other researchers would need to know before using the data. This meets many researchers' professional obligations to publish in peer-reviewed journals, while also upholding ethical obligations to provide data *and* contextual information about said data.

Another strategy often used by qualitative researchers to ensure rigor is that of triangulation, or the process by which "researchers search for congruence among multiple and different sources of information" (Creswell & Miller, **2000**, page 126). The data produced from such methods are necessary for knowledge acquisition and theory advancement. Scholars might triangulate multiple methods of data collection (e.g., interviews, observations, archival data), sources (e.g., different communities), researchers' accounts, and/or theoretical approaches, to see whether a particular

finding holds up across contexts. This strategy informs the rigor of our work by seeing how well it holds up to scrutiny under different conditions. Engaging in triangulation may reveal important disconfirming evidence that advances, questions, or adds nuance to a theory.

Materials

Quantitative researchers often use the term "materials" to refer to stimuli, models, programs, and other tools through which raw data are processed. In qualitative research, we might additionally consider journal entries, analytic or reflective memos, critical conversations, and member checking to all be materials – that is, products we create while processing data. How we interpret our data and draw conclusions depends on how well we understand our participants and mitigate any bias we may bring to the analysis. These materials are therefore as important as the data itself; they are the process by which researchers process raw data into findings. Part of our responsibility in qualitative open science—and, arguably, all open science—is to ensure that we are transparent not only with our data, but also with the materials, context, and strategies we use to interpret findings and draw conclusions. While these recommendations are echoed in current quantitative critical theory (QuantCrit; e.g., Gillborn et al., **2018**), they still tend to be heeded more often by qualitative researchers than quantitative and mixed methods researchers (Hope et al., **2019**).

Looking Ahead

Researchers have a habit of creating divisions where there need not be, a notorious example being the myth of incompatibility between qualitative and quantitative methods (Ercikan & Roth, **2006**; Malterud, **2001**). While these two approaches do bring different perspectives to the research process, there are opportunities to enhance the quality of our work by learning from *both* approaches. The open science movement and the research community in general can benefit from many of the practices qualitative researchers use to maintain rigor and transparency – namely, attention to providing a high level of contextual detail and reflexivity practices for mitigating bias. By expanding open science guidelines to leverage a broader array of rigor and transparency-promoting practices (e.g., reflexivity), we can truly begin to advance practices.

The rigor and transparency we call for is different from, and at times in opposition to, ideas of rigor that rely on generalizability alone. A study need not be replicable or reproducible to be rigorous. Instead, researchers need to be transparent about the context and reflexive processes used to draw conclusions from the data so readers can see their line of reasoning, determine the extent to which they trust the findings, and whether those findings might transfer to other contexts (e.g., trustworthiness; Lincoln & Guba, **1985**). While another researcher may not be able to replicate a study in another context and arrive at identical findings, they may be able to align their study design if they understand the contextual factors the analysis should incorporate. Documentation, such as preregistrations, journal entries, memos, member checking, and critical conversations are all materials that a researcher can upload, along with data, to an open science repository with the appropriate permissions clearly described. These materials would help justify how the data were interpreted and facilitate conceptual or theoretical replications. Given the diversity of practices across disciplines, one may question, as Clarke (**2022**) does in her review of Heidi Levitt's *Reporting Qualitative Research in Psychology: How to Meet APA Style Journal Article Reporting Standards*, whether general reporting standards are even necessary or possible. It may be that a one-size-fits-all approach will not work for the social sciences, let alone science more broadly. If one comes to this conclusion, we would argue that guidelines need to *clearly articulate the kinds of research to which they apply*. If one agrees with Levitt (**2020**) that general reporting standards are necessary and possible, then hopefully they also agree that when guidelines are advertised as "general science guidelines," then they need to reflect *all* forms of scientific inquiry, not just lab-based quantitative research.

Intentions alone are not enough to move science forward. Creating responsible, considered processes for rigorously transparent open science requires involving interested parties from a wide range of backgrounds,

perspectives, research areas, and training paradigms. If open science practitioners truly want their practices to become more mainstream, they must invite researchers with very different perspectives to the table, and everyone at the table must discuss these issues in good faith. If we can come to a mutual understanding of our various paradigms and agree upon guidelines that respect each other's epistemologies, we will be much more successful in moving the field forward.

References

Access controls. (2020). https://qdr.syr.edu/guidance/human-participants/access-controls (see p. 54).

Bennett, C., Fitzpatrick-Harnish, K., & Talbot, B. (2022). Collaborative untangling of positionality, ownership, and answerability as white researchers in indigenous spaces. *International Journal of Music Education*, *40*(4), 628–641 (see p. 48).

Billups, F. D. (2014). Trustworthiness and the quest for rigor in qualitative research. *NERA Researcher*, *52*, 10–12. https://www.nera-education.org/docs/TNR_Fall_2014_Color_Final.pdf (see p. 48).

Birks, M., Chapman, Y., & Francis, K. (2008). Memoing in qualitative research: Probing data and processes. *Journal of Research in Nursing*, *13*(1), 68–75. https://doi.org/10.1177/1744987107081254 (see p. 52).

Bornstein, J., Lustick, H., Shallish, L., Hannon, L., & Okilwa, N. (2022). Active accountability for disproportionate discipline and disability classification highlights student agency, contextualization, and racialization. *American Educational Research Association 2022 Conference* (see p. 52).

Caretta, M. A. (2016). Member checking: A feminist participatory analysis of the use of preliminary results pamphlets in cross-cultural, cross-language research. *Qualitative Research*, *16*(3), 305–318. https://doi.org/1177/1468794115606495 (see p. 52).

Caretta, M. A., & Perez, M. A. (2019). When participants do not agree: Member checking and challenges to epistemic authority in participatory research. *Field Methods*, *31*(4), 359–374. https://doi.org/10.1177/1525822X19866578 (see p. 52).

Clancy, M. (2013). Is reflexivity the key to minimising problems of interpretation in phenomenological research? *Nurse Researcher*, *20*(6), 12–16. https://doi.org/10.7748/nr2013.07.20.6.12.e1209 (see p. 51).

Clarke, V. (2022). Navigating the messy swamp of qualitative research: Are generic reporting standards the answer? *Qualitative Research in Psychology*, *19*(4), 1004–1012. https://doi.org/10.1080/14780887.2021.1995555 (see p. 55).

Corker, K. D.-K., Whylly, P. E., K, & Steltenpohl, C. N. (2022). *The importance of updating registrations: A round table discussion*. Center for Open Science. https://www.youtube.com/watch?v=6JfsBC31en4 (see p. 53).

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory Into Practice*, *39*(3), 124–130. https://doi.org/10.1207/s15430421tip3903_2 (see pp. 52, 54).

Creswell, J. W., Miller, D. L., & Poth, C. N. (2018). *Qualitative inquiry & research design: Choosing among five approaches*. SAGE. (See p. 48).

Cummings, J. A., Zagrodney, J. M., & Day, T. E. (2015). Impact of open data policies on consent to participate in human subjects research: Discrepancies between participant action and reported concerns. *PLoS One*, *10*(5), 0125208. https://doi.org/10.1371/journal.pone.0125208 (see p. 49).

Davies, D., & Dodd, J. (2002). Qualitative research and the question of rigor. *Qualitative Health Researcher*, *12*(2), 279–289. https://doi.org/10.1177/104973230201200211 (see p. 48).

DeHaven, A. (2017). *Preregistration: A plan, not a prison*. Center for Open Science. https://www.cos.io/blog/preregistration-plan-not-prison (see p. 50).

Elson, M. (2021). I'm all in favor of data sharing, even mandatory where possible. https://web.archive.org/web/20210929155622/https://twitter.com/maltoesermalte/status/1390758338321952770 (see p. 49).

Ercikan, K., & Roth, W. M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23. https://doi.org/10.3102/0013189X035005014 (see p. 55).

Fecher, B., & Friesike, S. (2014). Open science: One term, five schools of thought. In S. Bartling & S. Friesike (Eds.), *Opening science*. Springer. https://doi.org/10.1007/978-3-319-00026-8 (see p. 47).

Fielding, N. (2004). Getting the most from archived qualitative data: Epistemological, practical and professional obstacles. *International Journal of Social Research Methodology*, *7*(1), 97–104. https://doi.org/10.1080/1364557032000164069 9 (see p. 49).

Fook, J., & Askeland, G. A. (2007). Challenges of critical reflection: 'nothing ventured, nothing gained.

Social Work Education, *26*(5), 520–33. https://doi.org/10.1080/02615470601118662 (see p. 52).

Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In C. Geetz (Ed.), *The interpretation of cultures: Selected essays* (pp. 3–30). Basic Books. (See pp. 52, 54).

Gillborn, D., Warmington, P., & Demack, S. (2018). Quantcrit: Education, policy, 'big data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/ 10.1080/ 13613324.2017.1377417 (see p. 55).

Goldblatt, H., & Band-Winterstein, T. (2016). From understanding to insight: Using reflexivity to promote students' learning of qualitative research. *Reflective Practice*, *17*(2), 100–113. https://doi.org/10.1080/14623943.2015.1134471 (see p. 52).

Gow, J., Moffatt, C., & Blackport, J. (2020). Participation in patient support forums may put rare disease patient data at risk of re-identification. *Orphanet Journal of Rare Diseases*, *15*(1), 1–12. https://doi.org/10.1186/s13023-020-01497-3 (see p. 50).

Guillemin, M., & Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. *Qualitative Inquiry*, *10*(2), 261–280. https://doi.org/10.1177/1077800403262360 (see p. 51).

Hagger, M. S. (2019). Embracing open science and transparency in health psychology. *Health Psychology Review*, *13*(2), 131–136. https://doi.org/10.1080/17437199.2019.1605614 (see p. 47).

Haven, T. L., Errington, T. M., Gleditsch, K. S., van Grootel, L., Jacobs, A. M., Kern, F. G., Piñeiro, R., Rosenblatt, F., & Mokkink, L. B. (2020). Preregistering qualitative research: A delphi study. *International Journal of Qualitative Methods*, *19*, 1–13. https:/ / doi.org/ 10.1177/ 1609406920976417 (see pp. 50, 53).

Hope, E. C., Brugh, C. S., & Nance, A. (2019). In search of a critical stance: Applying qualitative research practices for critical quantitative research in psychology. *Community Psychology in Global Perspective*, *5*(2), 63–69. https://doi.org/10.1285/i24212113v5i2p63 (see pp. 51, 55).

Humphreys, L., Lewis Jr, 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 p. 54).

ICPSR. (n.d.). Accessing restricted data at ICPSR. https://www.icpsr.umich.edu/web/pages/ICPSR/access/restricted/ (see p. 54).

Jamieson, M. K., Govaart, G., & Pownal, M. (2022). Reflexivity in quantitative research: A rationale and beginner's guide. *PsyArXiv*. https://doi.org/10.31234/osf.io/xvrhm (see p. 49).

Johnson, J. L., Adkins, D., & Chauvin, S. (2020). A review of the quality indicators of rigor in qualitative research. *American Journal of Pharmaceutical Education*, *84*(1), 138–146. https://doi.org/10.5688/ajpe7120 (see p. 54).

Khan, M. T., Hyun, M., Kanich, C., & Ur, B. (2018). Forgotten but not gone: Identifying the need for longitudinal data management in cloud storage. *ACM Proceedings*, 1–12. https:/ / jglobal.jst.go.jp/en/detail?JGLOBAL_ID=202002290369007676 (see p. 50).

Lash, T. L. (2015). Declining the transparency and openness promotion guidelines. *Epidemiology*, *26*(6), 779–780. https:/ / doi.org/ 10.1097/EDE.0000000000000382 (see pp. 48, 49).

Lazard, L., & McAvoy, J. (2020). Doing reflexivity in psychological research: What's the point? what's the practice? *Qualitative Research in Psychology*, *17*(2), 159–177. https://doi.org/10.1080/14780887.2017.1400144 (see p. 51).

Levitt, H. M. (2020). *Reporting qualitative research in psychology: How to meet APA style journal article reporting standards*. American Psychological Association. (See p. 55).

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage. (See pp. 51, 55).

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation*, *30*, 73–84 (see pp. 51, 52).

Long-Sutehall, T., Sque, M., & Addington-Hall, J. (2011). Secondary analysis of qualitative data: A valuable method for exploring sensitive issues with an elusive population? *Journal of Research in Nursing*, *16*(4), 335–344. https://doi.org/10.1177/1744987110381553 (see p. 49).

Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly*, *25*(4), 153–171. https://doi.org/10.18352/lq.10113 (see p. 47).

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in educational research*, *16*(2), 193–205. http:/ / www.iier.org.au/ iier16/ mackenzie.html (see p. 53).

Makel, M. C., Meyer, M. S., Pei, M. A., Roberts, A. M., & Plucker, J. A. (2022). Replication is relevant in qualitative research. *Educational Research and Evaluation*, *27*(1-2), 215–219. https://doi.org/10.1080/13803611.2021.2022310 (see p. 51).

Malterud, K. (2001). Qualitative research: Standards, challenges, and guidelines. *The Lancet*, *358*(9280), 483–488. https://doi.org/10.1016/S0140-6736(01)05627-6 (see p. 55).

Mao, L., Mian Akram, A., Chovanec, D., & Underwood, M. L. (2016). Embracing the spiral: Researcher reflexivity in diverse critical methodologies. *International Journal of Qualitative Methods*, *15*(1), 1–8. https://doi.org/10.1177/1609406916681005 (see p. 52).

Mellor, D. T. (2021). Preregistration and transparency in the research process. *PsyArXiv*. https://doi.org/10.31219/osf.io/8rq3t (see p. 50).

Meyer, K., & Willis, R. (2019). Looking back to move forward: The value of reflexive journaling for novice researchers. *Journal of Gerontological Social Work*, *62*(5), 578–585. https://doi.org/10.1080/01634372.2018.1559906 (see p. 52).

Mill, J. E., & Ogilvie, L. D. (2003). Establishing methodological rigour in international qualitative nursing research: A case study from ghana. *Journal of Advanced Nursing*, *41*(1), 80–87. https://doi.org/10.1046/j.1365-2648.2003.02509.x (see p. 48).

Moravcsik, A. (2019). *Transparency in qualitative research*. SAGE Publications Ltd. https://doi.org/10.4135/9781526421036 (see p. 47).

Mozersky, J., Walsh, H., Parsons, M., McIntosh, T., Baldwin, K., & DuBois, J. M. (2020). Are we ready to share qualitative research data? knowledge and preparedness among qualitative researchers, IRB members, and data repository curators. *IASSIST quarterly*, *43*(4), 13–27. https://doi.org/10.1002/eahr.500044 (see p. 49).

National Institutes of Health [NIH]. (n.d.). *Enhancing reproducibility through rigor and transparency*. U.S. Department of Health; Human Services. https://grants.nih.gov/ policy/ reproducibility/ index.htm (see pp. 47, 51).

National Institutes of Health [NIH]. (2020). *NIH data sharing policy and implementation guidance*. U.S. Department of Health; Human Services. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/data-management-and-sharing-policy-overview#after (see p. 49).

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., & Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374 (see pp. 48, 49).

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13. https://doi.org/10.1177/1609406917733847 (see p. 51).

Office of Science and Technology Policy. (2022). Ensuring free, immediate, and equitable access to federally funded research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf (see p. 49).

Patton, M. Q. (2014). *Qualitative research & evaluation methods*. Sage Publications. (See p. 51).

Peredaryenko, M. S., & Krauss, S. E. (2013). Calibrating the human instrument: Understanding the interviewing experience of novice qualitative researchers. *Qualitative Report*, *18*(43), 1–17. https://doi.org/10.46743/2160-3715/2013.1449 (see p. 50).

Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*, *29*(1), 01822. https://doi.org/10.1002/eap.1822 (see p. 47).

Qualitative Data Repository. (n.d.). Preparing data files. https://qdr.syr.edu/guidance/managing/preparing-data (see p. 54).

Qualitative Data Repository. (2022). Human participants general guidance. https://qdr.syr.edu/guidance/human-participants (see p. 54).

Rodgers, B. L., & Cowles, K. V. (1993). The qualitative research audit trail: A complex collection of documentation. *Research in Nursing & Health*, *16*(3), 219–226. https://doi.org/10.1002/nur.4770160309 (see p. 52).

Rolfe, G. (2006). Validity, trustworthiness and rigour: Quality and the idea of qualitative research. *Journal of Advanced Nursing*, *53*(3), 304–310. https://doi.org/10.1111/j.1365-2648.2006.03727.x (see p. 48).

Rooney, V. M. (2015). Consent in longitudinal intimacy research: Adjusting formal procedure as a means of enhancing reflexivity in ethically important decisions. *Qualitative Research*, *15*(1), 71–84. https:/ / doi.org/ 10.1177/ 1468794113501686 (see p. 51).

Ross, M. W., Iguchi, M. Y., & Panicker, S. (2018). Ethical aspects of data sharing and research participant protections. *American Psychologist*, *73*(2), 138–145. https://doi.org/10.1037/amp0000240 (see pp. 48, 50).

Ruggiano, N., & Perry, T. E. (2019). Conducting secondary analysis of qualitative data: Should we, can we, and how? *Qualitative Social Work*, *18*(1), 81–97. https://doi.org/10.1177%2F1473325017700701 (see p. 49).

Sakaluk, J. K. (2021). Response to commentaries on sakaluk (2020. *Archives of Sexual Behavior*, *50*(5), 1847–1852. https://doi.org/10.1007/s10508-021-02020-w (see p. 48).

Savin-Baden, M., & Howell-Major, C. (2013). *Qualitative research: The essential guide to theory and practice*. Routledge. (See p. 51).

Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2019). Data papers as a new form of knowledge organization in the field of research data. *Knowledge Organization*, *46*(8), 622–638. https://halshs.archives-ouvertes.fr/halshs-02284548 (see p. 54).

Small, W., Maher, L., & Kerr, T. (2014). Institutional ethical review and ethnographic research involving injection drug users: A case study. *Social Science & Medicine*, *104*, 157–162. https://doi.org/10.1016/j.socscimed.2013.12.010 (see p. 50).

Stahl, N. A., & King, J. R. (2020). Expanding approaches for research: Understanding and using trustworthiness in qualitative research. *Journal of Developmental Education*, *44*(1), 26–28. https://files.eric.ed.gov/fulltext/EJ1320570.pdf (see p. 51).

Stake, R. E. (1995). *The art of case study research*. Sage. (See p. 52).

Steltenpohl, C. N., Montilla Doble, L. J., Basnight-Brown, D. M., Dutra, N. B., Belaus, A., Kung, C. C., Onie, S., Seernani, D., Chen, S., Burin, D. I., & Darda, K. (2021). Society for the improvement of psychological science global engagement task force report. *Collabra: Psychology*, *7*(1), 22968. https://doi.org/10.1525/collabra.22968 (see p. 48).

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing*, *48*(4), 388–396. https://doi.org/10.1111/j.1365-2648.2004.03207.x (see p. 51).

United Nations Educational, Scientific and Cultural Organization. (2021). UNESCO recommendation on open science. *UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000379949 (see p. 48).

VandeVusse, A., Mueller, J., & Karcher, S. (2022). Qualitative data sharing: Participant understanding, motivation, and consent. *Qualitative Health Research*, *32*(1), 182–191. https://doi.org/10.1177/10497323211054058 (see pp. 49, 50).

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884 (see p. 47).

Wolf, Z. R. (2003). Exploring the audit trail for qualitative investigations. *Nurse Educator*, *28*(4), 175–178. https://journals.lww.com/nurseeducatoronline/Fulltext/2003/07000/Exploring_the_Audit_Trail_for_Qualitative.8.aspx (see p. 52).

Youngblut, J. M., Loveland-Cherry, C., & Horan, M. F. (1990). Data management issues in longitudinal research. *Nursing Research*, *39*(3), 188–189. https://doi.org/10.1097/00006199-199005000-00019 (see p. 50).

A Manifesto for Rewarding and Recognising Team Infrastructure Roles

Arielle Bennett[1], Daniel Garside[2], Cassandra Gould van Praag[3], Thomas J. Hostler[4], Ismael Kherroubi Garcia[5], Esther Plomp[6], Antonio Schettino[7], Samantha Teplitzky[8], Hao Ye[9]

[1]The Alan Turing Institute; The Turing Way
[2]National Eye Institute, National Institutes of Health, USA
[3]Wellcome Centre for Integrative Neuroimaging, University of Oxford
[4]Manchester Metropolitan University, UK
[5]Kairoi Ltd
[6]Delft University of Technology, Faculty of Applied Sciences; The Turing Way
[7]Erasmus University Rotterdam; IGDORE
[8]University of California, Berkeley
[9]University of Florida

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
September 30, 2022
Accepted
June 26, 2023
Published
August 14, 2023
Issued
May 25, 2024

Correspondence
Delft University of Technology, Faculty of Applied Sciences
e.plomp@tudelft.nl

The Scientific Reform Movement has highlighted the need for large research teams with diverse skills. This has necessitated the growth of professional team infrastructure roles (TIRs) who support research through specialised skills, but do not have primary responsibility for conceiving or leading research projects. TIRs such as Lab Technicians, Project Managers, Data Stewards, Community Managers, and Research Software Engineers all play an important role in ensuring the success of a research project, but are commonly neglected under current reward and recognition procedures, which focus on the individual academic researcher instead of the teams involved. Without meaningful identification and recognition of TIR contributions, we risk reinforcing the conceptual and practical division between academic researchers and TIRs. This situation is inequitable and detrimental to the research enterprise: the limited potential for career advancement for TIRs may cause them to leave for other occupations, ultimately leading to a loss of institutional skill, expertise, and memory. This contribution explores the evolution of specialist TIRs and the status of these positions in various settings. We provide three case study descriptions of TIR activities, so that readers may become more familiar with the breadth and depth of their work. We then propose system level changes designed to embed meaningful recognition of all contributions. Acknowledging the contributions of all research roles will help retain skill and expertise, and lead to collaborative research ecosystems that are well-positioned to address complex research challenges.

Keywords *Team Infrastructure Roles, Rewards and Recognition, Research Evaluation, Team Science, Career*

The social and technological developments of recent decades have reinforced the notion of science as a team-based enterprise. As we tackle increasingly complex scientific questions (Coles et al., **2022**), we leverage the strengths of diverse research teams, recognising that we cannot solve the significant challenges of our time through isolated endeavours. This increased diversity in practice is part and parcel of the Scientific Reform Movement, which seeks to promote the uptake of practices that improve the transparency of the research process (Penders, **2022**), as well as to provide recognition for these practises (Coles et al., **2023**). Reform of academic publication and authorship practices are one route to address such issues, but we see authorship (or contributorship, see Rennie et al. (**1997**)) as a symptom of entrenched inequity, rather than the source of it. The Scientific Reform movement should go beyond reformation of publishing and aim instead to address fundamental roots of academic inequity, such as the perceptions of what it means to be a researcher and participate in research. In this piece we will explore a broad range of factors which may lead to inequity in the academic workforce and suggest changes to research systems to improve equitable practices.

The emergence of TIRs

To illustrate the increasingly diverse and team-

based approaches to research, consider that over 5,000 named authors across the globe collaborated in the detection of the Higgs Boson at CERN (Castelvecchi, **2015**), how successful climate models require expertise in atmospheric physics, soil science, meteorology, and more (Huebner et al., **2017**), or the integration of research into artificial intelligence with moral philosophy (Jobin et al., **2019**). With increasing collaboration and growing research complexity, new specialised roles have emerged to support research processes. We call these *team infrastructure roles* (TIRs), making explicit their structural function in the research process. TIRs bring vital expertise to the process of research, but they are not well integrated in traditional academic organisational structures.

TIRs contributing to the research process include laboratory technicians, project managers, grant officers, finance managers, privacy officers, patent officers, and internal review board members (Heffner, **1979**; UKRI, **2023**). These roles are known collectively as "professional service staff" or "research professionals". Their position in between supporting roles and academic researchers has been referred to as the "third space" (Whitchurch, **2008**). While some contributions of these roles may appear to be solely bureaucratic, one cannot deny the value of a skilled project manager, finance manager or technician in handling their respective responsibilities. We provide some examples of TIRs and their diverse areas of speciality below and in **Table 1**. These examples and perspectives are primarily informed by our academic experience in the US and Europe. The challenges, case studies, and changes that we suggest may be less applicable, or necessary, in other contexts. For example, low/middle income countries may prioritise other forms of research reform rather than dedicate resources to these types of positions (Bezuidenhout & Chakauya, **2018**; Bezuidenhout et al., **2017**; Onie, **2020**).

The emergence of new TIRs has introduced unmapped complexity into the academic ecosystem, particularly in relation to recognition, reward, and development. We argue that successful integration of TIRs in the academic system will require naming, exploring, and resolving frictions associated with these new roles.

Challenges

Lack of autonomy within TIR roles

Academic researchers are afforded substantial freedom in determining their career paths. This stems from historical positioning of academic researchers as "appointees" who perform scholarship as a public duty, rather than "employees" who are a means of production for a university (Finkin & Post, **2011**). This legitimises autonomy in the management of day-to-day activities and professional development (Wolf & Jenkins, **2021**), contributing to an internally recognised credit system.

In contrast, many TIRs are employed as "technical staff", with a specific remit in their job description to perform support activities, governed by the requirements of academic researchers or the broader goals of the research institute. Consequently, pursuing projects or publications outside of this support remit can be seen as a distraction. This lack of autonomy limits the ability of TIRs to prioritise the growth of their skills alongside evolving research disciplines or methodology, constrains their opportunities for progression towards leadership roles, and ultimately squanders their ability to inform the direction of the research agenda.

Limited formalisation of career pathways

Many TIR careers lack development pathways (NCRIS, **2022**; Virágh et al., **2019**). This is in contrast to academic research careers, where the criteria for promotion up to the highest levels are well documented, clearly advertised, and often supported by formal and informal systems of mentoring. For example, the *Vitae Researcher Development Framework* (Vitae, **2014**) maps out academic researchers' expected skill development across all facets of scholarly activity. Individuals employed in Human Resources or Finance positions can also access industry-specific accreditation and qualifications to support their progression (for example, training offered through the Chartered Institute of Personnel and Development for Human Resources professionals, or the Association of Chartered Certified Accountants for accountants).

In contrast, conventional opportunities for career development, such as increasing job responsibility and resulting uplifts in remuneration (UKRI-Research England, **2022**; Virágh et

al., **2019**), are inconsistent for TIRs. Individuals in TIR positions may therefore look outside of the academy for progression, with subsequent departures leading to institutional memory loss (Bossu & Brown, **2018**; McInturff & Adenis, **2022**). A lack of professional recognition also introduces challenges in funding TIRs, especially where salaries are not competitive with similar roles outside of academia (UKRI-Research England, **2022**). The restriction of developmental opportunities, lack of established profiles and compensation, and limited funding routes leave TIRs to act as lone advocates for their own positions, a stressful and complicated task due to their unique niche within the academic organisational structures.

Prejudice against TIR activities and career choices

The growing availability of TIRs in research institutes means that academic researchers can increasingly "outsource" some of the research responsibilities that were traditionally theirs alone. Passing those tasks to professionals may be viewed by some as "a hollowing out of [...] what it means [...] to be an academic" (Macfarlane, **2011**, p. 71). By this account, whilst specialisation of roles and responsibilities may increase efficiency, it may also negatively impact traditional academic values and identity, reinforcing a working culture geared only towards maximum productivity (Beatson et al., **2021**; Limas et al., **2022**; Wellcome Trust, **2020**). Thus, the mere existence of TIRs may be viewed negatively by some within the academy.

Prejudice can also result from changes to the status of roles within an institution. Harloe and Perry (**2005**) suggest that moving to a "co-operative form of production" akin to co-creation, rather than one in which TIRs simply facilitate the work of academics, may undermine the "collegial culture" in universities. In this culture, research academics have traditionally had exclusive responsibilities in determining their university's governance and organisation through engagement with institutional decision-making systems (such as committees). TIRs may thus be viewed as yet another non-academic staff member whose increasing influence dilutes academics' autonomy and authority, and/or increases their already heavy workload. This perspective highlights current tensions in the system: TIRs may be perceived as not sufficiently qualified to exert influence in the system, despite the fact that many TIRs are highly skilled researchers with doctoral degrees and years of academic experience (Teperek et al., **2022**; UKRI-Research England, **2022**).

TIRs may also be stigmatised as "failed academics" because they do not pursue traditional academic careers (ARMA, **2020**; Gould van Praag, **2022**; Sever & Janssen, **2017**). This parallels the prejudice against "leaving academia" for industry, often viewed as a last resort for those who "couldn't hack it" (Gewin, **2022**).

These prejudices towards the activities and career choices of TIRs make it more difficult to enact changes to infrastructure and reward systems which could benefit them. It also contributes to "imposter syndrome", with the barriers to reward and progression implicitly reinforcing the message that TIRs are of lower status than academic researchers (Sims, **2021**; UKRI-Research England, **2022**). Relatedly, the prejudice can also go the other way: TIRs may believe that academics' reluctance to engage with their help is limiting the potential of an institution (Harloe & Perry, **2005**). These tensions can negatively impact attempts at institutional change.

Recognition of TIR contributions

Academic incentives are often focused on the contributions of the individual, and the image of a "lone academic genius" (Elkins-Tanton, **2021**). This is reinforced by prizes awarded to singular "outstanding" academic researchers, the common practice of naming a research group by the lead Professor (for example, the "Smith lab"), and apparent ownership of team members ("[Person X] is *my* PhD student" or "*my* postdoc"). The power to confer authorship is generally enacted by senior researcher(s) and, in many disciplines, only the first and last authors are deemed to have done the actual work. Practically, however, research builds on previous work as well as a diversity of contributions that do not always lead to authorship and are therefore not formally recognised (Coles et al., **2022**; Forscher et al., **2020**; Shirazi, **2014**; Tiokhin et al., **2021**). By focusing solely on individuals and first/last authorship positions on publications, the academic research system neglects the value of a broader set of contributors - with their own unique skills and expertise

(Baum et al., **2022**). This results in precarious positions for TIRs, as their work rarely translates directly to authorship, let alone a first or last authorship position. TIRs are therefore not fully participating in the credit economy (Zollman, **2018**), where prestige from authorship and awards can bring further rewards in the form of downstream funding success and access to high-status jobs (Huebner & Bright, **2020**).

Growth of TIRs

Some emerging TIRs have been exemplary in handling the challenges outlined above. These examples may serve to illustrate the utility of making TIR duties, performance expectations and influence more explicit, along with the merits of forming professional communities of practice. These roles have been listed in order of more established (Research Software Engineer) to relatively recent (Research Application Manager). These roles exemplify how well-resourced TIRs can bring substantial value to the academic workflow. In **Table 1** we additionally summarise career trajectories and opportunities for recognition in each role.

Example 1: Research Software Engineer

Research software engineering represents an established specialised research role: a hybrid between researcher and programmer which requires expertise in both research and programming. Similar roles have existed for decades with a variety of titles, but the specific title – Research Software Engineer (RSE) – was conceived at Collaborations Workshop in Oxford in 2012 (Hettrick, **2016**), followed by the formation of the RSE Association in 2013. The rise of RSEs demonstrates the power of naming and defining a role, providing an identity and focal point for action (Sims, **2021**). Hettrick (**2016**) summarises the first four years of actions by the RSE Association, including numerous articles, market analysis, and policy work. Today, there are RSE networks on every continent, an international council of RSE associations, and an emerging, standardised career path for RSEs. Many institutions have established RSE groups, independent of research labs, while the Netherlands eScience Centre is an example of an independent organisation which centres the role of RSEs in the research process. This is the result of sustained, organised advocacy efforts by both researchers and RSEs.

RSEs function both as individuals in embedded roles as well as consolidated groups who provide expertise on a project-by-project basis within their institutions. This "consultant" model provides access to RSE expertise for groups who do not have the budget for longer term investment.

Example 2: Research Community Manager

Research Community Managers (also known as Scientific Community Managers) foster collaboration, engagement, connection, and productivity among members of a community, where a *community* is a group of people united by a common tool, discipline, location, service, or interest. Only in recent years the coordination and management of scientific communities has become formalised, as cross-institutional and international collaborations have become more common. The *Center for Scientific Collaboration and Community Engagement* (CSCCE) was established in 2016 to provide training, support infrastructure, and advocacy for Research Community Managers, formalising it as a distinct professional role (CSCCE, **2022a**). The first Community Engagement Fellowship cohort in 2017 kick-started the conversation around the nature of scientific community management and its unique challenges and considerations compared to communities outside academia. The CSCCE provides a space where Research Community Managers can receive support, domain-specific updates, and opportunities for collaboration and professional development. The CSCCE is now developing a community manager certification (CSCCE, **2022b**), so that individuals who are expected to foster community engagement can perform their role with confidence and a thorough understanding of the technical and theoretical basis of community activities.

Example 3: Research Application Manager

Research Application Managers (RAMs; The Turing Way Community, **2022b**) bring product thinking and stakeholder engagement to research outputs. For example, RAMs at The Alan Turing Institute address the need for sustainability of research infrastructure, extend

Table 1 A summary of each of the example roles described in the main text, highlighting whether there is an established professional advocacy organisation, expected career trajectories and professional development, comparisons to roles outside of research, and how these roles can be recognised.

	Research Software Engineer (RSE)	Research Community Manager (RCM)	Research Application Manager (RAM)
Summary of Role	Creates and/or maintains software specifically intended for research purposes	Fosters collaboration and engagement among a specific scientific community	Guides research projects (including infrastructure) for sustained impact and reuse through user community engagement
Professional Organisation	National and regional RSE associations	CSCCE	None yet
Sources of Professional Development	Software development training; Software Sustainability Institute	Community management training; CSCCE	Product management training
Career Pathways	Increasing rank, management of other RSEs or RSE teams	Director of organisations, scientific organisation administration, programme/network management	None yet
Non-research Equivalents	Software development	Community/outreach manager, developer advocate	Developer relations, product manager, developer advocate
Reward/Recognition Opportunities	Conferences, software publications, software citation, awards	Conferences, informal praise, training and development opportunities, contributorship on publications, awards	Conferences, inter-institute interactions, wider uptake of projects

existing research outputs and software, and seek opportunities to reuse and reproduce these outputs in new scenarios (The Turing Way Community, **2022b**). RAMs think beyond the research project cycle, cultivate a broader understanding of a discipline's trajectory, and understand the interconnectedness of scientific research more broadly. This role is still emerging as distinct from a Product Manager in industry or an academic Innovation Officer, with little formal documentation or organised advocacy in place. RAMs represent an interesting example of a newly emerging TIR which may experience a similar trajectory as RSEs and Research Community Managers.

Pathways forward

Here we present pathways through the challenges described and towards the successes of the highlighted case studies. We identify first steps towards a vision in which all TIRs are appropriately rewarded, recognised, and integrated with the work and priorities of research academics. An appropriate next stage will be the evaluation of costs and practicality of each intervention in supporting immediate or long-term change, with iterative piloting and refinement towards the idealised vision.

Re-imagine the research system to emphasise the process, not only the outcomes

Although research is primarily viewed in terms of knowledge production, we take inspiration from the values described in the SCOPE framework (INORMS, **2022**) and recommend that individual *outputs* (such as publications, discoveries, technologies) be deprioritised in favour of elevating the *process*. More specifically, many research activities do not directly lead to outputs that are commonly measured and rewarded in academia, such as those of the TIR

case studies described previously. Additionally, efforts that improve the research process by increasing transparency, reproducibility, and cooperation may not lead to journal publications. A narrow focus on publications as a reward mechanism will necessarily draw time away from such improvements. The focus on individual outputs additionally encourages implicit or explicit "gaming" of the system, with production metrics being prioritised above all other concerns (Goodhart's law; Goodhart, **1984**).

One way to emphasise the research process is through normalising the sharing of research artefacts (such as protocols, data objects, code, preprints) produced through the process. A move to more frequent or continuous publishing will alleviate some of the pressures associated with precarious contracts, such as the lag between contribution and traditional journal authorship. Expanding *incremental publications to include research artefacts, broadly defined,* can also reduce gatekeeping around authorship—research groups may be more willing to acknowledge a named contribution where there is a clearer connection between the work and the published object. For example, a lab technician working on a protocol will have a stronger claim to be a named contributor on a published protocol than a research paper that uses that protocol. Alongside systems that are specific for one type of output (for example, arXiv for preprints or PREreview for published peer reviews), general-purpose platforms such as ResearchEquals, PubPub, and Octopus enable the creation of a timely and persistent record of broad research contributions. By affording attention and credit to a broader range of output types, the primacy of the final journal article in evaluation metrics will be reduced and each contribution will garner respect in its own right.

An expansive system for recognising contributions

We imagine a future where research is inclusive and participatory, with each contribution being valuable to the process and subsequent outcomes. This requires the acknowledgment that different individuals bring a diverse and meaningful array of skills and expertise, including those from backgrounds that lack traditional academic credentialing. Contributions can be in the form of materially-visible work (for example writing, data collection, software development), workflow improvements, ideation, and more. A thorough and accurate accounting of all contributions will require moving beyond quantifiable metrics such as datasets curated or lines of code written. As TIRs can support the research process in a myriad of ways, integrating qualitative descriptions of their contributions will be necessary to properly recognise their efforts.

The Contributor Roles Taxonomy (CRediT; (Brand et al., **2015**) is an increasingly popular framework for recognising contributions. However, even with 14 codified roles, the CRediT system does not fully address the problem of recognising diverse contributions. As previously noted, it is too common that "research" is synonymous with "peer-reviewed publication", when there are many other contributions that are impactful within the research endeavour. For example, Harris et al. (**2020**) published on the decades-long collaborative NumPy programming library project. There was a notable lack of gender diversity among the listed authors of the published report (Gallant, **2022**), despite gender diversity among the more recent code and documentation contributors (Weber Mendonça, **2020**), raising the question of how to recognise indirect contributions. If research is conducted in a version control system that tracks all changes (such as the Open Science Framework), one might assume all contributions would be observable and easily collated. But such a system will overlook efforts that are not readily recorded in said system (such as coordination and planning efforts, or offline discussions). The Turing Way's 'Record of Contributions' (The Turing Way Community, **2022a**) demonstrates one way to recognise all forms of contributions, where indirect contributions can be nominated into the tracking system: namely, using the all-contributors bot (All Contributors, **2022**). In addition, systems for tracking impact via citations will need to be much more comprehensive. For example, even with Digital Object Identifiers (DOI) emerging as a de facto standard, a DOI generated using Zenodo is only recorded as a citation if it is properly indexed, which is currently not always the case.

Furthermore, a focus on publications will neglect some TIR contributions entirely, especially for roles where the primary responsibilities do

not include research. Indeed, TIR contributions can include teaching, training, mentorship, lab supervision, and consultations provided by specialised experts in funding acquisition, outreach, project management, statistics, data analysis, or software development. These contributions rely on research content expertise yet are not easily folded into publishable research objects. Although some of these activities are performed within the remit of high-level leadership, appointment to such positions often requires evidence of a "successful research career", ignoring the expertise accumulated in TIR roles. Although it is unrealistic to expect any single system for recognising contributions to be ideal for every context, a credit framework that is customisable for different institutions and locales is an important first step towards addressing these challenges.

A system to validate research outputs

The above framework presupposes a large expansion in the types of research outputs. However, there may be resistance in recognising these outputs as "valid" because many lack formal systems for external peer review. Indeed, a system which incentivises "productivity" without an assessment of quality (no matter the output type) could lead to decreased trust in research. To ensure the quality of research outputs, and the ability for researchers to build effectively upon each other's works, systems should be established for expert review of all research outputs. Mirroring the peer review system for publications, TIRs could then participate by contributing their experience and skills to the review process.

Notwithstanding the complex debates about open peer review (Heesen & Bright, **2021**; Ross-Hellauer, **2017**), unremitted labour (Aczel et al., **2021**), and power dynamics (Huber et al., **2022**), peer review can serve a useful purpose in validating research outputs. Realising an appropriate system for peer review of diverse research outputs, however, will require large infrastructural and behavioural shifts. In the case of research software, such systems have already emerged in venues such as rOpenSci (**2022**), pyOpenSci (Holdgraf et al., **2022**), and the Journal of Open Source Software (**2022**). For other types of outputs, a peer review system would need to be designed to integrate effectively with how the outputs are used. For example, research protocols cannot be easily modified following reviewers' suggestion, so there would have to be a well-specified role or aim for reviewer feedback beyond the suggestion of changes.

Standardised roles and pathways for career development

As demonstrated in the TIR examples above, and by Jetten et al., (Jetten et al., **2021**) for the Data Stewards in the Netherlands, the trend to professionalise TIRs leads to improvements in the visibility of their work, increased opportunities for training and networking with peers, and role-specific rewards and recognition. We argue that professionalisation also improves the integration of TIRs within research organisational structures. As seen with Research Software Engineers, TIRs may operate in fully independent teams that consult with academic researchers. This structure necessitates leadership responsibility, creating the opportunity for parity in responsibility and compensation between an academic researcher managing a lab group and a TIR managing a team of research support specialists. TIR leadership will also invite a degree of autonomy to direct activities and professional development within the team, including the opportunity to contribute to larger infrastructural change through service on institutional committees. The demarcation of specific responsibilities also supports negotiations to command a salary commensurate with expertise and makes it easier for individuals to move across institutions.

Professionalisation is, however, hampered by variability in the recognition and career support available to TIRs across institutions. This variability could be addressed through the creation of a new job family and pathway which parallels the development of the distinction between "Research", "Teaching and Research", or "Teaching and Scholarship" grades found in many UK institutions (for example the University of Sussex (**2019**) and University of St. Andrews (**2015**)), and the work by the National Collaborative Research Infrastructure Strategy (NCRIS, **2022**). Promotion levels in these new job families should match academic and managerial roles, in contrast to the Technical and Operational or Facilities profiles that only go as high as a standard post-doctoral grade. We

note that these job families were legitimised in the UK following negotiation between campus trade unions (University and Colleges Union (UCU), Unite and Unison) and representatives of the employers. Such a change may therefore require engagement of Unions across the sector to advocate on behalf of all research institution employees.

The professionalisation of TIRs could be further accelerated if larger mainstream funders created TIR fellowships (see similar recommendations by Teperek et al. (**2022**) and UKRI-Research England (**2022**)). This would require a cultural change from funders to value long-term investment in individual TIRs, and infrastructural change in how funds are distributed. In our idealised future, once role profiles are professionalised and standardised, institutions may ensure the continuity of support without the need for individual fellowships, through dedicated structural funding. A recent report by the UK Science, Innovation and Technology Government committee (U. K. Science, Innovation and Technology Committee, **2023**) on Reproducibility and Research Integrity recommended that "Funders and universities should develop dedicated funding for the presence of statistical experts and software developers in research teams. In tandem, universities should work on developing formalised, aspirational career paths for these professions." showing fledgling support for this idea at the highest level (U. K. Science, Innovation and Technology Committee, **2023**).

Conclusion

Recent socio-technical advancements have brought attention to the opportunities and needs surrounding research teams with diverse expert skills. Nevertheless, there is considerable work to be done to ensure that all individuals who make significant contributions to research teams are appropriately acknowledged and rewarded. TIRs are a unique facet of this problem, as positions dedicated to support research, but existing outside the typical researcher career structure. As a result, TIRs experience a lack of autonomy, have limited opportunities for career development, and face prejudice for deviating from the traditional academic credit system.

While acknowledging that there are significant challenges faced by TIRs in the current academic model, we highlighted three cases where there have been efforts to professionalise TIR profiles, thereby creating communities, recognisable standards in training, development opportunities, and collective advocacy: Research Software Engineers, Research Community Managers, and Research Application Managers.

Drawing from the successes and learnings of these examples, we suggest four system-level changes to address issues in the systems of reward and recognition available to TIRs, and their integration with the work and priorities of research academics. A summary of each proposal is provided below:

1. Shift the focus of academic research to appropriately value the *process* of the endeavour, not only the *prestige* of the outputs. Acknowledging that no output is necessarily final, we advocate for frequent or continuous public documentation (publication) of every stage of research, allowing for recognition of various contributions at each stage.
2. Expand the system for recognising contributions, going beyond the implementation of CRediT, by acknowledging contributions that are not visible in the form of authorship.
3. Create mechanisms for validating the quality and impact of non-journal outputs akin to peer review, noting that this will require infrastructural development in the delivery of review, and agreement on review standards for different output types.
4. Standardise and professionalise roles and pathways for career development, culminating in an academic career track which is distinct from the current "researcher" versus "non-researcher" dichotomy and, importantly, not restricted in the level of influence or reward achievable.

These proposals are offered at a time of increasing focus on increasing support for the open dissemination of research outputs (Concordat Working Group, **2016**; Office of Science and Technology Policy, **2022**; UNESCO, **2021**), calls to improve the broader culture of academia (COARA, **2022**; Wellcome Trust, **2020**), improving the bureaucratic efficiency of academia (Independent Review of Research Bureaucracy, **2022**), and the existing commitments to improve TIR positions (NCRIS, **2022**;

Technician Commitment, **2020**). If we seek to actualise the reform and ambitions of motions such as the San Francisco Declaration on Research Assessment (DORA, **2012**), we must acknowledge that there is significant scope to modernise the culture and tools we use to recognise and reward contributions. Systemic changes that improve the access of TIRs to career satisfaction will impact the reward and recognition processes relevant to the entire academy, making room to acknowledge, value and celebrate more diverse contributions and contributors to research.

Contributions

CRediT contributions were established using tenzing (Holcombe et al., **2020**):

- **Arielle Bennett**: Conceptualisation, Project administration, Supervision, Writing - original draft, and Writing - review & editing.
- **Daniel Garside**: Conceptualisation, Visualisation, and Writing - review & editing.
- **Cassandra Gould van Praag**: Conceptualisation, Visualisation, Writing - original draft, and Writing - review & editing.
- **Thomas J. Hostler**: Conceptualisation, Writing - original draft, and Writing - review & editing.
- **Ismael Kherroubi Garcia**: Conceptualisation, Writing - original draft, and Writing - review & editing.
- **Esther Plomp**: Conceptualisation, Project administration, Supervision, Visualisation, Writing - original draft, and Writing - review & editing.
- **Antonio Schettino**: Conceptualisation and Writing - review & editing.
- **Samantha Teplitzky**: Conceptualisation, Writing - original draft, and Writing - review & editing.
- **Hao Ye**: Conceptualisation, Project administration, Supervision, Visualisation, Writing - original draft, and Writing - review & editing.

Acknowledgements

We thank Dylan Roskam-Edris for helpful comments and Julien Colomb for sharing resources. Many thanks to Sarahanne Field for editing this issue and for her support and patience in the process. Thanks to Yo Yehudi for input on the abstract and title of this work. We thank Sander van der Laan and Theodosios Famprikis for their input on the preprint of this work. We thank Natalia B. Dutra and Christopher R. Chartier for their helpful and constructive reviews.

Funding

Arielle Bennett's contributions were supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/W006022/1, particularly the "Tools, Practices & Systems" theme within that grant & The Alan Turing Institute'. Cassandra Gould van Praag was supported by the NIHR Oxford Health Biomedical Research Centre and funded in whole, or in part, by the Wellcome Trust. Antonio Schettino was employed at Erasmus Research Services as Senior Advisor Open Science. Daniel Garside's contributions were supported by the Intramural Research Program of the NIH, National Eye Institute.

References

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*(1), 14. https://doi.org/10.1186/s41073-021-00118-2 (see p. 66).

All Contributors. (2022). *All contributors* (Version v2.17.0). Retrieved September 27, 2022, from https:/ / web.archive.org/ web/ 20220927025500/https://github.com/all-contributors/all-contributors (see p. 65).

ARMA. (2020). The ARMA survey on research culture. https://arma.ac.uk/wp-content/uploads/2021/03/ARMA-Research-Culture-Survey-2020.pdf (see p. 62).

Baum, M. A., Braun, M. N., Hart, A., Huffer, V. I., Meßmer, J. A., Weigl, M., & Wennerhold, L. (2022). The first author takes it all? Solutions for crediting authors more visibly, transparently, and free of bias. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12569 (see p. 63).

Beatson, N. J., Tharapos, M., O'Connell, B. T., Lange, P., Carr, S., & Copeland, S. (2021). The gradual retreat from academic citizenship. *Higher Education Quarterly*. https://doi.org/10.1111/hequ.12341 (see p. 62).

Bezuidenhout, L. M., & Chakauya, E. (2018). Hidden concerns of sharing research data by low/middle-income country scientists. *Global Bioethics*, *29*(1), 39–54. https://doi.org/10.1080/11287462.2018.1441780 (see p. 61).

Bezuidenhout, L. M., Leonelli, S., Kelly, A. H., & Rappert, B. (2017). Beyond the digital divide: Towards a situated approach to open data. *Science and Public Policy*, *44*(4), 464–475. https://doi.org/10.1093/scipol/scw036 (see p. 61).

Bossu, C., & Brown, N. (Eds.). (2018). *Professional and support staff in higher education*. Springer Singapore. https://doi.org/10.1007/978-981-10-1607-3 (see p. 62).

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211 (see p. 65).

Castelvecchi, D. (2015). Physics paper sets record with more than 5,000 authors. *Nature*, nature.2015.17567. https://doi.org/10.1038/nature.2015.17567 (see p. 61).

COARA. (2022). *The agreement on reforming research assessment*. Retrieved September 30, 2022, from https://web.archive.org/web/20220930124600/https://coara.eu/agreement/the-agreement-full-text/ (see p. 67).

Coles, N. A., DeBruine, L. M., Azevedo, F., Baumgartner, H. A., & Frank, M. C. (2023). Big team' science challenges us to reconsider authorship. *Nature Human Behaviour*, *7*(5), 665–667. https://doi.org/10.1038/s41562-023-01572-2 (see p. 60).

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. https://doi.org/10.1038/d41586-022-00150-2 (see pp. 60, 62).

Concordat Working Group. (2016). *Concordat on open research data*. UK Research and Innovation. Retrieved September 1, 2022, from https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-ConcordatonOpenResearchData.pdf (see p. 67).

CSCCE. (2022a). *About the center* [CSCCE]. Retrieved April 27, 2022, from https://web.archive.org/web/20220427165526/https://www.cscce.org/about/ (see p. 63).

CSCCE. (2022b). *CSCCE community manager certification program* [CSCCE]. Retrieved May 29, 2022, from https://web.archive.org/web/20220729002352/https://www.cscce.org/trainings/cscce-community-manager-certification-program/ (see p. 63).

DORA. (2012). *San Francisco declaration on research assessment* [DORA]. Retrieved September 3, 2022, from https://web.archive.org/web/20220903151339/https://sfdora.org/read/ (see p. 68).

Elkins-Tanton, L. (2021). Time to say goodbye to our heroes? *Issues in Science and Technology*, *37*(4), 34–40. Retrieved September 10, 2022, from https://web.archive.org/web/20220910202704/https://issues.org/say-goodbye-hero-model-science-elkins-tanton/ (see p. 62).

Finkin, M. W., & Post, R. C. (2011). *For the common good: Principles of American academic freedom*. Yale University Press. (See p. 61).

Forscher, P. S., Wagenmakers, E.-J., Coles, N. A., Silan, M. A. A., Dutra, N. B., Basnight-Brown, D., & IJzerman, H. (2020, May 20). *The benefits, barriers, and risks of big team science* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/2mdxh (see p. 62).

Gallant, L. (2022, September 16). *A team of 26 authors and there appears to be 0 gender diversity... that is an active choice and [tweet by @lissgallant]* [Twitter]. https://web.archive.org/web/20200916222153/https://twitter.com/lisgallant/status/1306357619712577537 (see p. 65).

Gewin, V. (2022). Has the 'great resignation' hit academia? *Nature*, *606*(7912), 211–213. https://doi.org/10.1038/d41586-022-01512-6 (see p. 62).

Goodhart, C. A. E. (1984). Problems of monetary management: The UK experience. In C. A. E. Goodhart (Ed.), *Monetary theory and practice: The UK experience* (pp. 91–121). Macmillan Education UK. https://doi.org/10.1007/978-1-349-17295-5_4 (see p. 65).

Gould van Praag, C. (2022, May 22). Off the beaten PI track [Organisation for Human Brain Mapping (OHBM) 2022 annual conference, Glasgow, UK.]. https://doi.org/10.5281/ZENODO.6651963 (see p. 62).

Harloe, M., & Perry, B. (2005). Repenser l'université sans la vider de son sens : Engagements externes et transformations internes de l'université dans l'économie du savoir. *Politiques et gestion de l'enseignement supérieur*, *17*(2), 31–45. https://www.cairn.info/revue-politiques-et-gestion-de-l-enseignement-superieur-2005-2-page-31.htm (see p. 62).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy [Number: 7825 Publisher: Nature Publishing Group]. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2 (see p. 65).

Heesen, R., & Bright, L. K. (2021). Is peer review a good idea? *The British Journal for the Philosophy of Science*, *72*(3), 635–663. https://doi.org/10.1093/bjps/axz029 (see p. 66).

Heffner, A. G. (1979). Authorship recognition of subordinates in collaborative research. *Social Studies of Science*, *9*(3), 377–384. https://doi.org/10.1177/030631277900900305 (see p. 61).

Hettrick, S. (2016, August 17). *A not-so-brief history of research software engineers* [Software sustainability institute]. Retrieved April 7, 2022, from https://web.archive.org/web/20220407191258/https://www.software.ac.uk/blog/2016-08-17-not-so-brief-history-research-software-engineers-0 (see p. 63).

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing (C. R. Sugimoto, Ed.). *PLOS ONE*, *15*(12), e0244611. https://doi.org/10.1371/journal.pone.0244611 (see p. 68).

Holdgraf, C., Solvik, K., Ogasawara, I., Brett, M., Sundell, E., gaow, Chen, Z., Joseph, M., Lau, S., Rokem, A., Willing, C., Nicholson, D., Mason, J., Wasser, L., Bantilan, N., Moss, S., & Kashyap, S. (2022, September 21). *pyOpenSci/contributing-guide: Pre release 0.3* (Version v0.3). Zenodo. https://doi.org/10.5281/ZENODO.7101778 (see p. 66).

Huber, J., M. Inoua, S., Kerschbamer, R., König-Kersting, C., Palan, S., & Smith, V. L. (2022). Nobel and novice: Author prominence affects peer review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4190976 (see p. 66).

Huebner, B., & Bright, L. K. (2020). Collective responsibility and fraud in scientific communities. In S. Bazargan-Forward & D. Tollefsen (Eds.), *The Routledge handbook of collective responsibility* (1st ed.). Routledge. https://doi.org/10.4324/9781315107608 (see p. 63).

Huebner, B., Kukla, R., & Winsberg, E. (2017). *Making an author in radically collaborative research* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190680534.003.0005 (see p. 61).

Independent Review of Research Bureaucracy. (2022). *Independent review of research bureaucracy - final report*. UK Government - Department for Business, Energy & Industrial Strategy. (See p. 67).

INORMS. (2022). *The SCOPE framework, a five-stage process for evaluating research responsibly* (No. 10). Retrieved August 30, 2022, from https://web.archive.org/web/20220801134009/https://inorms.net/scope-framework-for-research-evaluation/ (see p. 64).

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M., & Van Gelder, C. W. G. (2021, March 19). *Professionalising data stewardship in the Netherlands. competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship*. Zenodo. https:/ / doi.org/ 10.5281/ ZENODO.4320504 (see p. 66).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2 (see p. 61).

Journal of Open Source Software. (2022). *Review criteria*. Retrieved May 11, 2022, from https://web.archive.org/web/20220511204643/https://joss.readthedocs.io/en/latest/review_criteria.html (see p. 66).

Limas, J. C., Corcoran, L. C., Baker, A. N., Cartaya, A. E., & Ayres, Z. J. (2022). The impact of research culture on mental health & diversity in STEM. *Chemistry – A European Journal*, *28*(9). https://doi.org/10.1002/chem.202102957 (see p. 62).

Macfarlane, B. (2011). The morphing of academic practice: Unbundling and the rise of the para-academic. *Higher Education Quarterly*, *65*(1), 59–73. https://doi.org/10.1111/j.1468-2273.2010.00467.x (see p. 62).

McInturff, S., & Adenis, V. (2022). It takes a laboratory to avoid data loss. *Nature*. https://doi.org/10.1038/d41586-022-02967-3 (see p. 62).

NCRIS. (2022). Towards better recognition for research infrastructure specialists. the Australian national fabrication facility. https://web.archive.org/web/20230425094426/https://anff.org.au/news/towards-better-recognition-for-research-infrastructure-specialists/ (see pp. 61, 66, 67).

Office of Science and Technology Policy. (2022, August 25). Public access memo (A. Nelson, Ed.). https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf (see p. 67).

Onie, S. (2020). Redesign open science for Asia, Africa and Latin America. *Nature*, *587*(7832), 35–37. https://doi.org/10.1038/d41586-020-03052-3 (see p. 61).

Penders, B. (2022). Process and bureaucracy: Scientific reform as civilisation. *Bulletin of Science, Technology & Society*, *42*(4), 107–116. https://doi.org/10.1177/02704676221126388 (see p. 60).

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails. A proposal to make contributors accountable. *JAMA: The Journal of the American Medical Association*, *278*(7), 579–585. https://doi.org/10.1001/jama.278.7.579 (see p. 60).

rOpenSci. (2022). *Software peer review* [rOpenSci]. Retrieved July 4, 2022, from https://web.archive.org/web/20220704125950/https://ropensci.org/software-review/ (see p. 66).

Ross-Hellauer, T. (2017). What is open peer review? a systematic review. *F1000Research*, *6*, 588. https://doi.org/10.12688/f1000research.11369.2 (see p. 66).

Sever, R., & Janssen, K. (2017). Career options for scientists. *Cold Spring Harbor Perspectives in Biology*, *9*(9), 032755. https://doi.org/10.1101/cshperspect.a032755 (see p. 62).

Shirazi, R. (2014, July 15). *Reproducing the academy: Librarians and the question of service in the digital humanities* [Roxanne shirazi]. Retrieved July 6, 2022, from https://web.archive.org/web/20220617010749/https://roxanneshirazi.com/2014/07/15/reproducing-the-academy-librarians-and-the-question-of-service-in-the-digital-humanities/ (see p. 62).

Sims, B. H. (2021). *Research software engineer as an emergent professional identity: A sociological perspective*. https://www.osti.gov/servlets/purl/1784685 (see pp. 62, 63).

Technician Commitment. (2020). *Technicians make it happen* [Technicians make it happen]. Retrieved September 1, 2022, from https://web.archive.org/web/20200809162757/https://www.technicians.org.uk/technician-commitment (see p. 67).

Teperek, M., Cruz, M., & Kingsley, D. (2022). Time to re-think the divide between academic and support staff. *Nature*. https://doi.org/10.1038/d41586-022-01081-8 (see pp. 62, 67).

The Turing Way Community. (2022a). *Record of contributions*. Retrieved June 4, 2022, from https://web.archive.org/web/20220604150908/https://the-turing-way.netlify.app/afterword/contributors-record.html (see p. 65).

The Turing Way Community. (2022b). *Research application managers: Overview* [The Turing Way] [https://doi.org/10.5281/zenodo.6533831]. Retrieved July 12, 2022, from https://web.archive.org/web/20220712021649/https://the-turing-way.netlify.app/collaboration/research-infrastructure-roles/ram.html (see pp. 63, 64).

Tiokhin, L., Panchanathan, K., Smaldino, P. E., & Lakens, D. (2021). *Shifting the level of selection in science*. MetaArXiv. https://doi.org/10.31222/osf.io/juwck (see p. 62).

U. K. Science, Innovation and Technology Committee. (2023). *Reproducibility and research integrity report (sixth report of the session 2022-2023)*. https://committees.parliament.uk/publications/39343/documents/194466/default/ (see p. 67).

UKRI. (2023). 101 jobs that change the world. https://www.ukri.org/news-and-events/101-jobs-that-change-the-world/ (see p. 61).

UKRI-Research England. (2022). *Research culture: A technician lens*. Retrieved September 28, 2022, from https://www.mitalent.ac.uk/Research-Culture (see pp. 61, 62, 67).

UNESCO. (2021). *UNESCO recommendation on open science*. United Nations Educational, Scientific and Cultural Organization. Retrieved August 20, 2022, from https://web.archive.org/web/20220820070614/https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en (see p. 67).

University of St. Andrews. (2015). *Job families and generic role descriptors - guidance notes* [University of St Andrews - human resources]. Retrieved September 24, 2022, from https://web.archive.org/web/20151007123310/https://www.st-andrews.ac.uk/hr/gradingrewardandconditions/jobfamiliesgenericroledescriptors/jobfamiliesguidancenotes/ (see p. 66).

University of Sussex. (2019). *Academic role profiles* [University of Sussex]. Retrieved December 20, 2019, from https://web.archive.org/web/20191222011201/https://www.sussex.ac.uk/humanresources/business-services/jobevaluation/academicroleprofiles (see p. 66).

Virágh, E., Zsár, V., & Balázs, Z. (2019). *Research management and administration: A profession still to be formalized*. HÉTFA Research Institute, Center for Economic, and Social Analysis. Budapest, Hungary. https://hetfa.eu/wp-content/uploads/2019/04/Research-managers_final_0408.pdf (see p. 61).

Vitae. (2014). *Vitae researcher development framework*. Retrieved September 1, 2022, from https://web.archive.org/web/20220901044422/https://www.vitae.ac.uk/researchers-professional-development/about-the-vitae-researcher-development-framework (see p. 61).

Weber Mendonça, M. (2020). *Hi, just note that the authorship in this paper reflects contributions of the past 20 years, and the community has [Tweet by @melissawm]*. Retrieved September 16, 2020, from https://web.archive.org/web/20200916224502/https://twitter.com/melissawm/status/1306363367825776640 (see p. 65).

Wellcome Trust. (2020, January 15). *What researchers think about the culture they work in*. Wellcome Trust. https://wellcome.org/reports/what-researchers-think-about-research-culture (see pp. 62, 67).

Whitchurch, C. (2008). Shifting identities and blurring boundaries: The emergence of third space professionals in UK higher education. *Higher Education Quarterly*, *62*(4), 377–396. https://doi.org/10.1111/j.1468-2273.2008.00387.x (see p. 61).

Wolf, A., & Jenkins, A. (2021). *Managers and aca-*

demics in a centralising sector: The new staffing patterns of UK higher education. Nuffield Foundation. Retrieved September 28, 2022, from https://www.kcl.ac.uk/policy-institute/assets/managers-and-academics-in-a-centralising-sector.pdf (see p. 61).

Zollman, K. J. S. (2018). The credit economy and the economic rationality of science. *The Journal of Philosophy*, *115*(1), 5–33. https://doi.org/10.5840/jphil201811511 (see p. 63).

Tension Between Theory and Practice of Replication

Erkan Buzbas[1,*], Berna Devezer[1,2,*]

A core problem that has been addressed in the scientific reform movement so far is the low rates of reproducibility of research results. Mainstream reform literature has aimed at increasing reproducibility rates by implementing procedural changes in research practice and scientific policy. At the sidelines of reform, theoreticians have worked on understanding the underlying causes of irreproducibility from the ground up. Each approach faces its own challenges. While the mainstream focus on swift practical changes has not been buttressed by sound theoretical arguments, theoretical work is slow and initially is only capable of answering questions in idealized setups, removed from real life constraints. In this article, we continue to develop theoretical foundations in understanding non-exact replications and meta-hypothesis tests in multi-site replication studies, juxtapose these theoretical intuitions with practical reform examples, and expose challenges we face. In our estimation, a major challenge in the next generation of the reform movement is to bridge the gap between theoretical knowledge and practical advancements.

Keywords *formal theory, scientific reform, reproducibility, replication, multi-site replications, meta-hypothesis*

*Both authors contributed equally to this work.

[1]Department of Mathematics and Statistical Science, University of Idaho

[2]Department of Business, University of Idaho

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
November 28, 2022
Accepted
August 22, 2023
Published
November 2, 2023
Issued
May 24, 2024

Correspondence
University of Idaho, Department of Business
bdevezer@uidaho.edu

Funding
Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The scientific reform movement that was propelled by exposing a scientific replication crisis has attracted interest of scholars spanning a wide range of disciplines from social and biomedical sciences to humanities and formal sciences. A core problem attacked has been low rates of results reproducibility[1] (given a result from an original study, the probability of obtaining the same result in replication experiments) in some scientific disciplines. We have argued elsewhere that first generation approaches have predominantly aimed at reforming scientific policy and practice to improve results reproducibility, at times based on hasty diagnoses of problems and normative implementations of procedural solutions (Devezer et al., **2021**). This approach of reforming scientific practices via bureaucratic innovations (Penders, **2022**) has effectively, if unintentionally, derailed the investigation of underlying causes of results reproducibility and deprioritized a rigorous theoretical foundation. We fear that this may have led to the illusion that fundamentally mathematical, complex issues can be solved with procedural fixes. Perhaps such mechanistic solutions may indeed help improve the efficiency of certain scientific processes (Peterson & Panofsky, **2023**). But if we aim to address mathematical problems, the ability to identify and attack them as mathematical seems critical. For example, it should be clear that the true reproducibility rate of a true result in a sequence of exact replication studies with independent and identically distributed random samples can take any value on $[0, 1]$ depending on the elements defining that study. Thus, we should not have a default expectation of high rates of reproducibility or assume that low reproducibility is indicative of false discoveries (Buzbas et al., **2023**). A designation of *crisis* could as well point to a problem with our expectations. How do we calibrate our expectations from replication studies, especially regarding reproducibility rates?

As statisticians, explanations and solutions that we deem satisfactory for a procedure require clear and precise reporting about models and analyses. In contrast, many high-profile,

[1]Also referred to as *replicability* in the literature.

multi-site replication studies reporting on results reproducibility fail to even state the statistical model under which inference is made (e.g., Klein et al., **2018**; Open Science Collaboration, **2015**). We would be in a precarious position to interpret any results or make statistical recommendations without providing a statistical model and tight theoretical reasoning for it. Such precision is not trivial or merely decorative. As many analyst studies (Breznau et al., **2022**; Silberzahn et al., **2018**) show with striking clarity, our inference depends on and may drastically vary with the statistical models we assume.

On the other hand, precision is useless without theoretical understanding. As John von Neumann said: "There's no sense in being precise when you don't even know what you're talking about." So not only do we need precise reporting of model specification but also a statistical understanding of its assumptions, performance, and implications. In this sense, there is a growing tension between the practical and theoretical foci in science reform, and from a theoretical perspective, there is a lot of room for improvement with regard to the rigor of statistical reasoning in the first generation reform literature. Such theoretical reasoning stands to provide much needed clarity and precision in making sense of the practical advances the reform movement has brought about. However, there are outstanding challenges on the theory side of things as well.

In this article, we aim to provide insight into the process of thinking about some key statistical issues regarding the reproducibility of scientific results, particularly in current big team applications of multi-site replication studies. We strive to expose what it takes to make precise statistical statements, what kind of questions are raised and need to be addressed on the way, and why this matters. Key points of theoretical exposition are the importance of distinguishing between exact and non-exact replications, and their proper statistical treatment in the analyses of replication data. We believe that a strong theoretical understanding of non-exact replications should inform our interpretation of results reproducibility in replication studies. We further evaluate the consequences of this distinction on testing meta-hypotheses in large-scale multi-site replication studies using well-established statistical theory. Our treatment will hopefully make clear what exactly can be gained by stronger theoretical foundations in science reform moving forward. Finally, as a next generation challenge, we propose the daunting yet necessary interdisciplinary work of bringing the theory and practice closer together.

Distinguishing between exact and non-exact replications

We set out to work with an idealized version of the real-life problem regarding the divergences between replication studies and the originals. An idealized study comprises background knowledge, an assumed statistical model, and statistical methods to analyze a sample, which is generated under the true model independent of all else (Devezer et al., **2021**). A *result* is a function from the space of analysis to the real line. To evaluate the reproducibility rate of a result obtained from a sequence of studies, all replications must randomly sample the same sampling distribution of results. This is guaranteed if a sequence of replication studies are *exact* replications, that is, only the random sample is new across the replications. Statistically, the best way to assess whether a given result from one study is reproduced in its replications is by the relative frequency of that result in all replication studies. A major value of this idealization is advancing understanding of the theory of results reproducibility under exact replications.

In practice, this idealized setup is unrealistic, given uncontrollable sources of variability across replications. No sensible person would argue that they have replicated another study exactly, except perhaps under rare, extremely well-controlled studies. The main issue is that we expect the natural phenomenon represented to be distorted by specific conditions contributing to the design in different ways because every controlled study imposes constraints of its own. Some concrete evidence showing this comes from purpose-built and performed large-scale replication studies. These studies have not been able to fix the design parameters so as to perform exact replication studies of each other. Examples of variations in study parameters purported to be replications of each other may include:

- studies sampling only subsets of populations (i.e., non-random sampling of a larger population),
- unequal sample sizes,
- missing data of different kinds,
- differences in sampling methods,
- differences in post-processing of data,
- differences in statistical analysis methods.

Therefore, with the exception of specialized studies where extremely well-controlled designs involving few random variables can be implemented—such as in particle physics—we argue that exact replication experiments are unlikely to be operationalized in daily science. Consequently, the sampling distribution of the results in replication studies will differ from that of the original study or from each other, and drawing conclusions about result reproducibility becomes a challenging problem. Leaving aside the (perhaps more important) issue of why the sampling distributions differ from each other, how they differ from each other becomes the primary objective to understand if we are to study results reproducibility properly. Hand-waving at a sequence of non-exact replications as "conceptual replications" and claiming to test the generalizability of an underlying effect won't do as it is begging the question. Therefore, to develop a broad theory of results reproducibility, we must work with a sequence of not-necessarily-exact replications, and treat the case of exact replications as a special case.

Unfortunately, for non-exact replications, we have very limited theory. What we know can be summarized as follows. A sequence of replication studies can be treated as a proper stochastic process. If all the elements of replication studies are exactly equivalent to each other except the sample generated under the true model, then the case of exact replications can be treated as a special case, yielding a single sampling distribution of reproducibility rate. When replications are non-exact, there is no single true reproducibility rate for results from all studies and therefore, no single sampling distribution of reproducibility rate. The mean of the process, then, becomes the target, but it needs to be interpreted carefully (see Buzbas et al., **2023**, for more information).

A critical matter is that theoretical conclusions about reproducibility depend on the choice of the *result*, and hence the mode of statistical inference in studying results reproducibility. Harking back to L. J. Savage, we believe that the *result* needs to be "as big as an elephant" to allow for generalizability of conclusions. The least useful mode of statistical inference to study results reproducibility is the null hypothesis significance test under the frequentist approach due to its inflexible interpretation of results probabilistically (e.g., cannot talk about probability of a result unconditional of a hypothesis) and its forced dichotomization of results in hypothesis testing (e.g., reject or fail to reject). These properties obscure the measurable quantitative signal in studies of reproducibility, thereby decreasing the resolution in results which could be used to understand the mechanisms driving irreproducibility. If there exists a problem of result reproducibility in hypothesis testing, it will likely exist under other major modes of inference such as point estimation, interval estimation, model selection, and prediction of future observables.

An efficient way to investigate results reproducibility information theoretically is under a more mathematically refined mode of inference than null hypothesis significance testing. Estimation theory is the best understood mode of statistical inference, equipped with most well-established theorems. It is our safety net. Based on estimation theory, model selection is a challenging but ultimate target in modern statistics (as well as in Devezer et al., **2019**).

Now that we have laid out our foundations, let us turn to the theoretical challenges that arise when trying to understand results reproducibility from non-exact replications.

Evaluating the results from non-exact replications with respect to results reproducibility

In this section we discuss the problem of evaluating results reproducibility from replication studies. Our perspective involves reference to the true data generating mechanism M_T, which we assume exists and we would like to operate under[2]. For convenience and purposes

[2] If M_T does not exist, we chase a moving target. There are some tools of mathematical statistics to study these cases (see for example M-open versus M-closed discussion in Bernardo and Smith, **2000** and references therein), but these are out of our scope for the treatment of results reproducibility.

of illustration, we assume a linear model with additive errors $M_T := \{Y|\mathbf{X}_T\beta_T = \mathbf{X}_T\beta_T + \epsilon\}$, which is well-studied and accommodates common research studies. Here, Y is $n \times 1$ vector of response variables, $\mathbf{X}_T$ is $n \times p$ matrix of predictors (e.g., design matrix in experiments), β_T is $p \times 1$ vector of model parameters, and ϵ is $n \times 1$ vector of stochastic errors. It is convenient to assume zero mean, constant variance, and uncorrelated errors, which are collectively known as the ordinary least squares (OLS) assumptions. Often, a full model specification is required for analysis, so these assumptions are augmented with normal distribution of errors. Thus, M_T is an exactly determined representation of a natural phenomenon of our interest or in other words, the *target true* data generating mechanism. In practice, we may not know the space on which M_T exists. It is worth mentioning that only if the set of models considered by a researcher includes M_T, we can hope to study results reproducibility of true results. More precisely, we should be able to include M_T in our consideration of data generating mechanisms with possible null values for some parameters. Here, we assume that we can satisfy this assumption. However, we invite the reader to ponder how challenging it is to satisfy this assumption in practice. For the discussion below, we also define M_O, the assumed model in the original (first) study, and M_R, the assumed model in a subsequent replication study.

Hypothetically, in an ideal study there should be no *intervention by the study* which distorts the data generated by M_T. If the study assumes that it is operating under M_T as the data generating mechanism, the model is correctly specified and the solution $\hat{\beta}_T = (\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{X}_T'Y$ are best linear unbiased estimators. Almost all statistical inferential goals require some version of $\hat{beta}_T$ under linear models. Therefore, without loss of generality, here we set it as an example of a target *result*.

We believe that the hypothetical scenario of no intervention by a study is not realistic to evaluate results reproducibility from a sequence of studies because it requires the equivalence of M_T, M_O, M_R, and so on for other replication studies. This equivalence means that all variables that are indeed in M_T are included in M_O and M_R, and no variable that is not in M_T are excluded in M_O and M_R[3]. Again, in practice any study introduces its constraints on the data generating mechanism—see the list in the preceding section for some examples of constraints. If we proceed on this assumption, then M_O is likely not equivalent to M_T. So we ask: In which aspects M_O differ from M_T? Naturally, we also ask: In which aspects M_R differ from M_O?

A reviewer has raised the question of why M_T needs to be invoked at all to compare results reproducibility from M_O and M_R, since they do not involve M_T directly, with good reason. That is, it is sufficient to assess whether M_O and M_R are equivalent without reference to M_T. This assessment is fair but it invites further difficulties of statistics to the study of results reproducibility. M_O equivalent to M_R but not equivalent to M_T is the case of inference from replications under *model misspecification*. We defer discussing some of the issues related to this case to the end of this section drawing from a broader lore of mathematical statistics. In the next paragraph, we briefly focus on the consequences of all three models being different from each other on estimates.

We assume that the original study includes only some of the variables that are indeed in M_T, and also includes some variables that are not part of M_T in the assumed model. The relevant matrix of predictors (e.g., design matrix) which we denote by $\mathbf{X}_O$ in the assumed model M_O has now changed. If it is correctly identified, this study operates under the data generating model $M_O := \{Y|\mathbf{X}_O\beta_O = \mathbf{X}_O\beta_O + \epsilon\}$, and the solution for OLS estimates in M_O is now $\hat{\beta}_O = (\mathbf{X}_O'\mathbf{X}_O)^{-1}\mathbf{X}_O'Y$. There is no convincing argument to assume that M_O is equivalent to M_T. Thus, $\beta_T \neq \beta_O$ implying that $\hat{\beta}_T \neq \hat{\beta}_O$ given the same data. Even if M_O is equivalent to M_T, it is more likely for the replication studies in the sequence to generate data under models different than M_T or M_O via the process of inclusion and exclusion of variables and due to examples of variations listed in the previous section. Let the assumed model (if identified correctly) in a replication study be $M_R := \{Y|\mathbf{X}_R\beta_R = \mathbf{X}_R\beta_R + \epsilon\}$, and the solution for OLS estimates

[3] Note that we have already made the simplifying assumption of operating in a universe of linear models. In reality, the functional form of the relationship among variables is another factor that can vary across M_T, M_O, M_R.

in M_R is now $\hat{\beta}_R = (\mathbf{X}_R'\mathbf{X}_R)^{-1}\mathbf{X}_R'Y$. Thus, $\hat{\beta}_T$, $\hat{\beta}_O$, and $\hat{\beta}_R$ are not equivalent estimates even when they see the same data.

For the case where M_O and M_R are equivalent to each other but different from M_T, we would have $\hat{\beta}_O = \hat{\beta}_R$, and we can assess the reproducibility rate of a specific result since we have exact replications. Assuming the number of studies all under M_O increases, the estimated reproducibility rate converges to the true reproducibility rate of the specific result. The problem with this rate is that it is the true reproducibility rate of a false or true result obtained under the misspecified model M_O. There is no systematic approach that can tell how the misspecification affects the truth or falsity of the obtained result. That is, even if we assume that M_O is close enough to M_T to pass, say a pre-determined criterion modelwise, we do not have a theory that would guarantee that the reproducibility rate of a specific (true) result obtained under M_O would be close to the reproducibility rate of the same type of result obtained under M_T. The closeness of the models does not guarantee the closeness of the reproducibility rates of all results and the relationship is often a function of the result and the type of the model misspecification itself. In other words, we cannot tell what rate of reproducibility to expect for a given result under model misspecification unless we study the particular kind of misspecification in a given context. Nor can we tell what a successful or failed replication means exactly. If, however, a theory of dependence between reproducibility rate of results under correctly specified models and misspecified models can be built, then the condition of direct equivalence of M_T and M_O can be bypassed. Building such a theory would require measuring the discrepancy between M_T and M_O which still requires knowledge of M_T to some extent. The consequences of model misspecification are often case specific, and strong results for broad classes of models have been elusive to frustration of statisticians. This issue is beyond the scope of this paper and by invoking M_T as our target, we believe that we can have a more fruitful first step in discussing results reproducibility, at least at this early theoretical stage of the field.

A simple example to show drastic effects of non-exact replications

Let us assume that the true model is $M_T := E(Y|\beta_T) = \beta_T$, that is a one-parameter model with no predictors. Our interest is to estimate the population mean β_T using the sample $Y_1, Y_2, \cdots, Y_n$ independently generated under the assumed model in the original study $M_O := E(Y|\beta_O) = \beta_O$. As we have indicated in general $\beta_O \neq \beta_T$. Without loss of generality, let us define a very simple relationship $\beta_O = \beta_T + c$, where $c \neq 0$ is some constant. The best estimator of the population mean is the sample mean and so we use the estimator $\hat{\beta} = n^{-1}\Sigma_{i=1}^{n} Y_i$. However, we quickly realize that $E(\hat{\beta}) = E(Y|\beta_O) = \beta_O = \beta_T + c$. Thus, the sample mean obtained in the study with assumed model M_O is a *biased estimator* for the true population mean of interest β_T under the true model. And it gets worse. We also quickly realize that by weak law of large numbers $\hat{\beta}$ is a consistent estimator of β_O the population mean from which the sample is drawn and $\beta_T = \beta_O - c$ implies that the probability $P(|\hat{\beta} - \beta_T| > \delta)$ does not converge to 0 for any small $\delta > 0$ as the sample size increases. Thus, the sample mean of the data obtained in the study with assumed model M_O is also an inconsistent estimator for the true population mean of interest β_T under the true model. An immediate implication of this argument is on the choice of sample size as the divergent factor between a true model, an original study, and its replications. To assess the reproducibility of a given result, the result must be obtained from studies with equal sample sizes across original and replication studies and identical methods must be applied to arrive to a result. Researchers in practice might be tempted to increase the sample size in a replication (e.g., for higher statistical power as a better standard to meet). However, larger sample sizes do not necessarily imply that true results will be more reproducible. To extend the theoretical example given above, we consider testing the true hypothesis $\beta_T = 0$ using intervals $\hat{\beta} \pm SE(\hat{\beta})$, where $SE(\hat{\beta})$ is the standard error of the sampling distribution of $\hat{\beta}$. The reproducibility rate defined as the proportion of the number of true results to the number of total results in a sequence of studies decreases with larger sample sizes, converging to zero (Figure 1). This

leads to a dilemma: Do we strive to conduct very few replication studies but making them as exact as possible at a great expense of resources because we know how to interpret their results theoretically, or do we try and run studies that are non-exact even if we do not quite understand what the implications of this non-exactness may be?

We conclude that to connect the results of *non-exact* replication studies, we need theoretical results on how non-exact they are and what the exact effect of non-exactness is on the results. Once again, hand-waving at perceived closeness or likeness will further obscure our vision and make it more difficult to properly interpret the replication results. In Buzbas et al. (**2023**), we show how easily true or false results can be made more or less reproducible with small perturbations in study design.

Many analyst studies (e.g., Breznau et al., **2022**; Hoogeveen et al., **2022**; Silberzahn et al., **2018**) provide an example of how this type of non-exactness may play out in practice. In these studies, independent research teams are given the same data set and asked to perform analyses to test a given scientific hypothesis. What is striking is that not only inferential procedures but also assumed models vary greatly across analysts, resulting in a range of results not necessarily consistent with each other. Note that these studies use the same data set that samples a given population. In non-exact replication studies, inference is performed from independent data sets often sampling different populations. Also worth noting is that the many analyst results confirm our earlier point about the weakness of focusing on hypothesis tests instead of a more generic modeling framework.

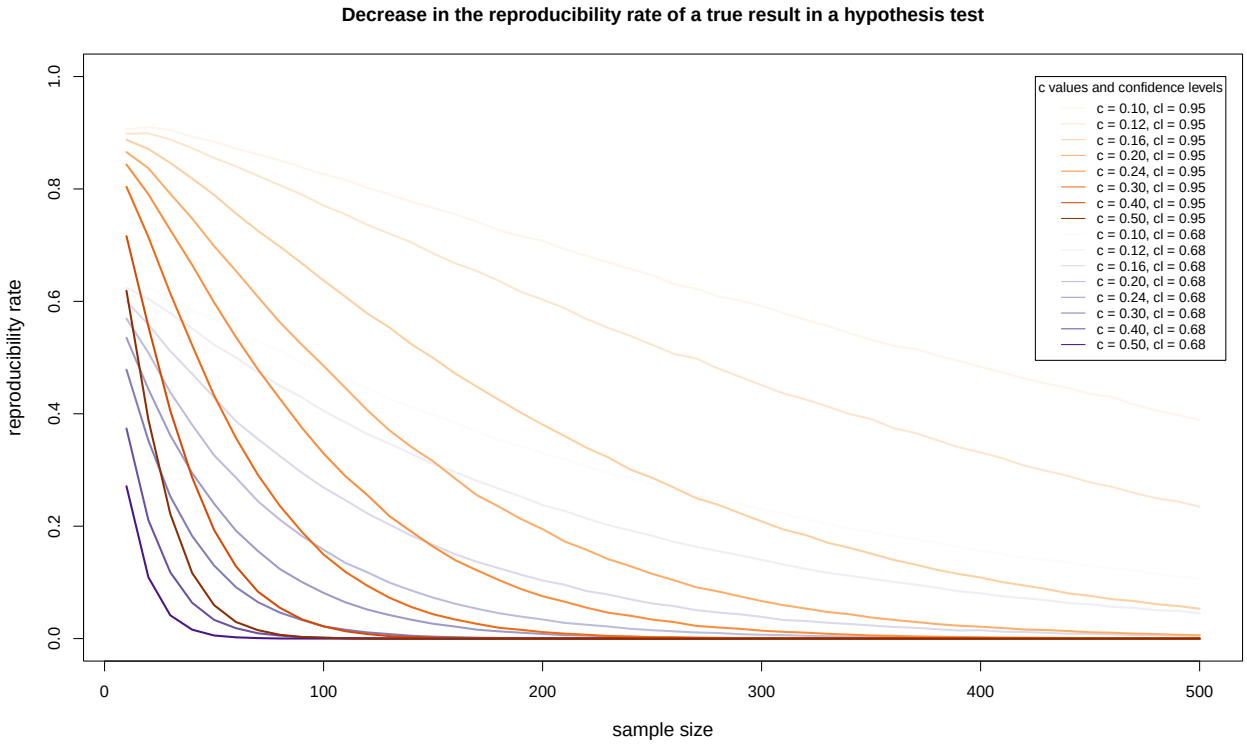

Figure 1 The reproducibility rate of a true result in a hypothesis test decreasing with increasing sample size. The (true) null hypothesis is $\beta_T = 0$, where $\beta_O = \beta_T + c$ and $c > 0$ is a constant.

Studies for testing meta-hypotheses under non-exact replications

As a more advanced example to show how the points we presented might affect the conclusions about results reproducibility in multi-site replication studies, we consider the following models for testing meta-hypotheses, in multi-site studies (e.g., author involvement effect tested in Many Labs 4, Klein et al., **2022**). For the ith observation, we define $g(u_i, v_i)$ as a function of $1 \times p$ vector of predictors such that u are variables associated with the meta-hypothesis (e.g., varying on larger experimental units) and v are variables associated with the original predictors (e.g., varying on smaller experimental units) of the study (Jones & Nachtsheim, **2009**). The matrix of predictors $\mathbf{X}_{T^*}$ has ith row as $g(u_i, v_i)$. Due to two levels of the study, stochastic error must now conform not only the errors within each replication, but also the errors associated with meta-hypothesis. For the meta-hypothesis we have $\mathbf{Z}\eta$ where $\mathbf{Z}$ is $n \times k$ of indicator functions whose k^{th} element is 1 (e.g., for a level of meta-hypothesis) and others 0, and η is $k \times 1$ vector of normally distributed errors with 0 mean and σ^2_η variance. For replications we have ϵ is $n \times 1$ normally distributed errors with 0 mean and σ^2_ϵ variance for each observation. Assuming additive errors as before, we have $\mathbf{Z}\eta + \epsilon$ for stochastic errors, and η_i and ϵ_j are assumed to be uncorrelated for all observation pairs (i, j). The true model generating the data is $M_{T^*} := \{Y|\mathbf{X}_{T^*}\beta_{T^*} = \mathbf{X}_{T^*}\beta_{T^*} + \mathbf{Z}\eta + \epsilon\}$. Parallel to the argument leading to M_O from M_T, the *feasible generalized least squares estimates* in an original study M_{O^*} is $\hat{\beta}_{O^*} = (\mathbf{X}'_{O^*}\widehat{\Sigma}^{-1}\mathbf{X}_{O^*})^{-1}\mathbf{X}'_{O^*}\widehat{\Sigma}^{-1}Y$, where $\Sigma = \sigma^2_\eta \mathbf{Z}\mathbf{Z}' + \sigma^2_\epsilon \mathbf{I}_n$ is the covariance matrix. Σ is often unknown. It is estimated by sample variances at the meta-hypothesis and replication variables levels, assuming that the replications are exact.

The estimator $\hat{\beta}_{O^*}$ clearly shows why non-exactness of replications will exacerbate errors in estimates in larger models such as those used in testing meta-hypotheses. The reason is that larger models often require nuisance parameters also to be estimated in addition to parameters of interest. Even if we take the best approach to inference, the non-exactness of the replication will be reflected on multiple estimates resulting in undesirable estimators.

For our example, to obtain the estimate of interest $\hat{\beta}_{O^*}$, we also need to obtain $\widehat{\Sigma}$. The method of estimation for $\widehat{\Sigma}$ (e.g., restricted maximum likelihood or Bayes) will inevitably use $\mathbf{X}_{O^*}$ and the properties of $\widehat{\Sigma}$ will be affected by non-exactness of replications. This leads to biased estimates. Hence, non-exactness of replications casts more doubt on the inferential results of large models such as meta-hypothesis tests.

Many Labs 4 (Klein et al., **2022**) provides a case study. The original hypothesis being tested over replications is the mortality salience effect and the experimental units are individual research participants; the meta-hypothesis being tested is the effect of original authors' involvement where experimental units are participating labs. The replications are non-exact in many ways from populations being sampled to unequal cell and sample sizes, from variations in in-house protocols to unequal block sizes. All of these factors likely bias effect size estimates in unpredictable ways. Considering the fact that inference is not made under a model that accounts for the randomization restriction and hierarchical experimental units in the actual experimental design, the results become even harder to interpret. The theory tells us what conclusions cannot be supported by the design and the analysis but it does not tell us what conclusions can be justified; that is, until that theory is advanced specifically for such cases.

Conclusion

We can clearly use mathematical statistics to advance our theoretical understanding of replications and results reproducibility under idealized conditions. This approach requires meticulous, persistent, and rigorous mathematical work that aims at theoretical clarity and precision. Nonetheless reality and scientific practice always impose new constraints on the problems at hand and as a result, oftentimes whenever theory meets reality, its reach falls short.

We know a lot about the consequences of exact replications yet we also know that exact replications are hard to achieve. In practice, we might think that even if we cannot perform exact replications, controlled lab conditions can approximate the ideal conditions assumed by statistical theory. This is a strong assumption

that often gets violated. First, the effect of sampling different populations in replications cannot be remedied by randomization. Second, an approximation is a statement about a quantity approaching to another quantity in a precise mathematical way. It is not some haphazard likeness that we do not know how to define or verify. The mathematical approximation can be measured with precision but the perceived likeness of studies cannot. For example, for the sample mean, we know that the variance of its sampling distribution decreases with the inverse of the sample size linearly and we can measure the performance of this approximation in an analysis, data point per data point if need be. We would be challenged, however, to show how much a given replication study approximates an original study with respect to a reasonable measure in a statistical sense. The theory for measuring such closeness or likeness does not yet exist. Hence, our theoretical knowledge of exact replications is of little direct use for practice and for a thorough understanding of real life non-exact replications, we lack as rigorous a theory.

Mainstream scientific reform movement has focused on procedural and bureaucratic solutions (e.g., preregistration, data and code sharing) as well as reforms in scientific policy (e.g., reducing publication bias via registered reports). As we alluded to earlier, these developments are often focused on urgent action and are not always grounded in theoretical understanding. In the same period, some progress has been made toward building theoretical foundations of metascience by our own work and others' (Bak-Coleman et al., **2022**; Fanelli et al., **2022**; Fanelli, **2022**; Smaldino & McElreath, **2016**), albeit in the margins of reform. Such theoretically guided approaches, on the other hand, are constrained by the scope of the idealizations involved and may not readily translate to scientific practice. As theoreticians our job does not end with developing rigorous theory. We also strive to reach a "post-rigorous" stage, as mathematician Terence Tao observed, where we have grown comfortable enough with the rigorous foundations of metascience so that we finally feel "ready to revisit and refine [our] pre-rigorous intuition on the subject, but this time with the intuition solidly buttressed by rigorous theory" (Tao, **2007**). Theoretical work is still in its early stages of development and needs to continue. At the same time, another major challenge arises for the next generation of reform: How do we bridge the gap between theory and practice?

Acknowledgements

The authors thank Professor Don van Ravenzwaaij for their constructive comments on an earlier draft of the paper.

Funding

Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Bak-Coleman, J., Mann, R. P., West, J., & Bergstrom, C. T. (2022). Replication does not measure scientific productivity. *SocArXiv*, *doi*, 10.31235/osf.io/rkyf7 (see p. 80).

Bernardo, M. J., & Smith, A. F. M. (2000). *Bayesian theory*. Chichester, UK: John Wiley & Sons Ltd. (See p. 75).

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., ..., & Zoltak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), Article e2203150119 (see pp. 74, 78).

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, *10*(3), 221042 (see pp. 73, 75, 78).

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), 1–23. https://doi.org/https://doi.org/10.1371/journal.pone.0216125 (see p. 75).

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 (see pp. 73, 74).

Fanelli, D. (2022). The" tau" of science-how to measure, study, and integrate quantitative and qualitative knowledge. *MetaArXiv*, *doi*, 10.31222/osf.io/67sak (see p. 80).

Fanelli, D., Tan, P. B., Amaral, O. B., & Neves, K. (2022). A metric of knowledge as information compression reflects reproducibility predictions in biomedical experiments. *MetaArXiv*, *doi*, 10.31222/osf.io/5r36g (see p. 80).

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A. J., Allen, P. J., Altay, S., Alzahawi, S., Amir, Y., Anthony, F.-V., Appiah, O. K., Atkinson, Q. D., Baimel, A., Balkaya-Ince, M., Balsamo, M., Banker, S., Bartos, F., Becerra, M., Beffara, B., … Wagenmakers, E. J. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*, 1–47 (see p. 78).

Jones, B., & Nachtsheim, C. J. (2009). Split-plot designs: What, why, and how. *Journal of quality technology*, *41*(4), 340–361 (see p. 79).

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., … Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), 35271 (see p. 79).

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490 (see p. 74).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716 (see p. 74).

Penders, B. (2022). Process and bureaucracy: Scientific reform as civilisation. *Bulletin of Science, Technology & Society*, *42*(4), 107–116 (see p. 73).

Peterson, D., & Panofsky, A. (2023). Metascience as a scientific social movement. *Minerva*, *61*, 147–174 (see p. 73).

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahníkk, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/https://doi.org/10.1177/2515245917747646 (see pp. 74, 78).

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science*, *3*(9), 160384 (see p. 80).

Tao, T. (2007). There's more to mathematics than rigour and proofs. *https://terrytao.wordpress.com/career-advice/theres-more-to-mathematics-than-rigour-and-proofs/*. https://terrytao.wordpress.com/career-advice/theres-more-to-mathematics-than-rigour-and-proofs/ (see p. 80).

Reputation Without Practice? A Dynamic Computational Model of the Unintended Consequences of Open Scientist Reputations

Maximilian Linde[1,2], Merle-Marie Pittelkow[1,3], Nina R. Schwarzbach[4], Don van Ravenzwaaij[1]

[1]Unit of Psychometrics and Statistics, Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

[2]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[3]QUEST Center for Responsible Research, Berlin Institute of Health at Charité, Berlin, Germany

[4]Unit of Clinical and Developmental Neuropsychology, Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands

Part of Special Issue
Consequences of the Science Reform Movementt - https://doi.org/10.36850/jote.i4.1

Received
December 20, 2022
Accepted
October 27, 2023
Published
March 15, 2024
Issued
May 24, 2024

Correspondence
Maximilian Linde, GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
maximilian.linde@gesis.org

Practicing open science can have benefits for the career prospects of individual researchers or labs through higher quality work and increased chances of publication. However, being an outspoken advocate of open science might also indirectly benefit individual scientific careers, in the form of status in a scientific community, decisions for tenure, and eligibility for certain kinds of funding. Therefore, it may be profitable for individual labs to appear to engage in open science practices, without actually putting in the associated effort or doing only the bare minimum. In this article, we explore two types of academic behavior through a dynamic computational model (cf. Smaldino & McElreath, **2016**) of an academic community that rewards open science: (1) practicing open science and/or (2) advocating open science. Crossing these two types of behavior leads to four different kinds of labs and we examine which of them thrive in this academic community. We found that labs that practice and advocate open science dominate in a scientific community that values open science. Implications of the model results are discussed.

Keywords *computational model, cultural evolution, metascience, open science, reform*

Initiatives to improve science often follow times of crisis. For example, the open science (OS) movement originated from a crisis in psychology, referred to as replication crisis, crisis of credibility, confidence, or reproducibility (Spellman et al., **2018**; Pashler & Wagenmakers, **2012**; Baker, **2016**; Ioannidis, **2005**; Open Science Collaboration, **2015**; Simmons et al., **2011**; Wagenmakers et al., **2012**; Fiedler, **2011**). Broadly speaking, the OS movement aims to make the scientific process more transparent, accessible, and reproducible. Practices associated with this movement (OS practices) include preregistration and the use of registered reports to reduce researcher's degrees of freedom (Munafò et al., **2017**; Chambers, **2013**), protocol, data and code sharing to improve reproducibility and replicability (National Academies of Sciences, Engineering, and Medicine, **2019**), and the use of preprints and open access publishing to increase the dissemination and accessibility of research findings (McKiernan et al., **2016**; Mikki, **2017**). Incentives such as "OS badges" rewarding openly sharing data or material (Kidwell et al., **2016**), and preregistration (https://osf.io/tvyxz/) advertise and identify "trustworthy" research (Schneider et al., **2020**). More recently, additional badges for open access publication, open code, open source, and open science grants have been proposed (Guzman-Ramirez et al., **2023**).

There is wide-spread agreement that adopting OS practices has advantages for both science at large as well as the individual researcher (Allen & Mehler, **2019**; McKiernan et al., **2016**; Markowetz, **2015**). Researchers are encouraged to use OS practices to advance their career by increasing their citation

Take-home Message

Labs that practice open science (e.g., preregistration, registered reports, sharing data, materials and codes, and open access publishing) and advocate open science (e.g., through social media) thrive in a scientific community that values open science. At the same time, "quick-and-dirty" science is still prevalent, as evidenced by high false positive and false discovery rates. Based on the specific assumptions of our model, our results suggest that labs that practice and advocate open science are dominating in a scientific community that values open science. These results are encouraging to those who feel practicing open science "is not worth it": in addition to benefits to science at large, our results suggest engaging with open science can benefit individual researchers if open science is sufficiently rewarded.

count, generating media attention, attracting potential collaborators, and getting job and funding opportunities (McKiernan et al., **2016**). Moreover, policy decisions aim to recognize and reward the use of open science (see, e.g., https://www.nwo.nl/en/recognition-and-rewards). However, incentivising OS practices might bring along secondary, unintended problems. As the traditional publish-or-perish culture may have inspired questionable research practices like p-hacking and hypothesizing after the results are known, so too may the elevated status and increased publication chances of practitioners of OS inspire advocating OS without actually engaging in OS practices.

The Present Study

In this article, we explore the benefit of practicing and advocating OS in a scientific community that rewards OS. To this end, we extended a computational model by Smaldino and McElreath (**2016**) who demonstrated that the current incentive structure in science, that rewards many and highly cited publications, could lead to low quality studies. Their results imply that in order to be successful, labs should favor a "quick-and-dirty" approach to conducting studies even though that would lead to a high false positive rate and a high false discovery rate. We extended this computational model by including four different lab types that are a factorial combination of practicing OS (yes/no) and advocating OS (yes/no). In this work, we (1) examined which lab type(s) dominate(s) in a scientific community that values OS and (2) investigated the dynamics of several characteristics in a scientific culture.

We highlight that this work is exploratory and meant to be a proof of principle. While we ground our operationalizations and the selection of parameter values in the existing literature and our personal experiences as OS researchers, we do not claim that our results fully capture the complexity and the individuality of OS labs. Rather they are a simplification of reality and aim to illustrate how the landscape of science might change under different conditions.

Evolution of Bad Science

The original methods to build the evolutionary model are reported in Smaldino and McElreath (**2016**), and a detailed explanation is provided in Box 1. In short, the model starts with a population of $N = 100$ labs that conduct research, publish papers, and gain rewards based on the number of publications and their associated value. Each lab is characterized by: (1) power W, the ability/probability of a lab to positively identify a true effect; (2) replication rate r, the probability of conducting a replication; (3) effort e, the amount of time a lab spends on conducting a study; and (4) false positive rate α, the probability that a lab incorrectly claims an effect (in statistical testing referred to as the significance level).

Variation in these four characteristics leads to variation in fitness of the labs, which determines which labs "die" (e.g., a principal investigator no longer has any students or funding and as a result decides to leave academia) and which labs "reproduce" (e.g., a prolific PhD-student from a successful lab starts a lab of their own) to create offspring labs. Survival of the labs depends on payoffs that they receive for publishing research projects. At each time step, each lab either initiates a new investigation or not. The new investigation can be either a replication study or not. Results of a new investigation can be negative (−) or positive (+),

Box 1: Evolution in Smalidino and Mcelreath (2016)

Evolution Characteristics

Evolution takes place over many time steps in the model (i.e., $100,000$ in Figure 3 and $1,000,000$ in Figures 4 and 5 of Smaldino & McElreath, **2016**). At time step 1, the simulations are initialized with lab characteristics $W = 0.8$, $r = 0.01$, and $e = 75$ for each lab.

Probability and Type of Investigation

The probability that a lab launches a new investigation (h) at a given point in time depends on η (the influence of effort on productivity) and e of the corresponding lab:

$$h(e) = 1 - \eta \log_{10}[e]. \quad (1)$$

If the lab initiates a new study at a given time step, it is a replication study with probability r, which varies across labs; it is a novel study with probability $1 - r$.

Probability of Obtaining a Positive Result

If the new study is a novel study, the underlying hypothesis is true with probability b, which is fixed at $b = 0.1$ for all time steps and labs; the underlying hypothesis is false with probability $1 - b$. If the underlying hypothesis is true, the lab observes a positive novel finding with probability W, which varies across labs; if the underlying hypothesis is false, the lab observes a positive novel finding with probability α, which varies across labs.

Probability of Publishing and Payoff

A positive novel finding will be published with probability $C_{N+} = 1$. If the positive novel finding is published, the corresponding lab receives a payoff of $V_{N+} = 1$ that is added to the already accumulated payoff. A negative novel finding will be published with probability $C_{N-} = 0$. If the negative novel finding is published, the corresponding lab receives a payoff of $V_{N-} = 1$ that is added to the already accumulated payoff. Any published novel finding (i.e., both positive and negative) will be added to the literature and is therefore available as a target for replication by other labs.

If the new study is a replication study, a hypothesis is randomly chosen from the literature (i.e., from the collection of studies that were already conducted by other labs). If the underlying hypothesis of the original study is true, the lab observes a positive replication finding with probability W, which varies across labs; if the underlying hypothesis of the original study is false, the lab observes a positive replication finding with probability α, which varies across labs.

A positive replication finding will be published with probability $C_{R+} = 1$. If the positive replication finding is published, the corresponding lab receives a payoff of $V_{R+} = 0.5$ that is added to the already accumulated payoff. Moreover, the lab that originally investigated the hypothesis receives a payoff of $V_{0+} = 0.1$ that is added to the already accumulated payoff. A negative replication finding will be published with probability $C_{R-} = 1$. If the negative replication finding is published, the corresponding lab receives a payoff of $V_{R-} = 0.5$ that is added to the already accumulated payoff. Moreover, the lab that originally investigated the hypothesis receives a payoff of $V_{0-} = -100$ (a penalty) that is added to the already accumulated payoff.

Evolution Dynamics

At each time step, the mean of W, α, and e across labs, and the false discovery rate (FDR) are calculated. FDR corresponds to the proportion of false positive findings among all positive findings across labs at a given time step. However, data is only collected (written to a file) every $2,000$ time steps.

After every time step, an evolution step takes place in which one lab ``dies'' and one lab ``is born''. To determine the dying lab, $d = 10$ labs are randomly selected, of which the lab with the highest number of active time steps dies. If multiple labs tie, one is chosen at random. To determine the lab that procreates, $d = 10$ labs are randomly selected, of which the lab with the highest payoff reproduces. If multiple labs tie, one is chosen at random.

The offspring lab inherits the characteristics from the reproducing lab. However, the inherited characteristics are allowed to mutate. All characteristics (r, e, W) mutate with a probability of $\mu_r = \mu_e = \mu_W = 0.01$. If characteristics do mutate, the new value is $\mathcal{N}(x, y)$, where x corresponds to the old value of the characteristic and y to either 0.01 for r and W or 1 for e. If this mutation process exceeds a boundary of the allowed range ($[0, 1]$ for r and W; $[1, 100]$ for e), the corresponding boundary is used as the new characteristic value.

Lastly, if labs publish novel studies, they are added to the literature. The size of the literature is limited to $1,000,000$ hypotheses. If the number of hypotheses in the literature exceeds $1,000,000$, the oldest hypotheses are removed until the number of hypotheses in the literature is $1,000,000$ again.

which determines their probability to be published C. The payoff for a published result V depends on whether it is a novel (N) or a replication (R) study and whether its outcome is negative ($-$) or positive ($+$).

Extension

We extended the model of Smaldino and McElreath (**2016**) by differentiating between labs that do or do not practice OS and between labs that do or do not advocate OS, yielding four types of labs (see Table 1).

The four lab types were initially represented in equal proportions (i.e., at time step 1). When a lab reproduced, the offspring lab automatically inherited the lab type from the reproducing lab. To have enough labs of each category, we increased the number of labs from $N = 100$ to $N = 400$.

We made the following assumptions about the impact of practicing OS on the survival of labs:

1. **Practicing OS leads to higher workload (e).**
   The practice of OS requires more work and time compared to closed science (Hostler, **2023**). New skills and knowledge need to be acquired and the research process involves additional steps, such as pre-registration, data and code cleaning, and additional administration (e.g., drafting openness agreements Hostler, **2023**). Indeed, practicing OS is associated with an increase in workload, work-related stress, and longer time to completion of a research project (Sarafoglou et al., **2022**; Toth et al., **2021**). Seen as "increasing effort decreases the productivity of a lab, because it takes longer to perform rigorous research" (Smaldino & McElreath, **2016**, p. 6), we reasoned that labs that practice OS should have a higher e than labs that do not practice OS. If, for instance, a traditional study were to take 400 work hours to be completed, we assumed that practicing OS would add 20 hours. This translates into a 5% increase in effort for OS studies (i.e., $(400 + 20)/400 = 1.05$). We believe this to be a conservative estimate of the increase in e.
2. **Practicing OS increases the probability of publishing negative novel findings (C_{N-}).**

Table 1 The four different types of labs

		Practicing OS	
		yes	no
Advocating OS	yes	Practice; advocate	Practice; not advocate
	no	Not practice; advocate	Not practice; not advocate

The proportion of published findings with statistically non-significant results is higher for registered reports (60.5%; Allen & Mehler, **2019**) or preregistered studies (52%; Toth et al., **2021**) compared to traditional research, with estimates ranging from 0% to 20% (e.g., Allen & Mehler, **2019**; Fanelli, **2012**). In what follows, we take the liberal estimates: 60% of non-significant OS studies get published and 20% of non-significant traditional studies get published. Assuming the same absolute number of published significant studies in both fields, this means that for every eight statistically significant traditional studies two statistically non-significant traditional studies get published (80% vs 20%); and for every eight statistically significant OS studies twelve statistically non-significant OS studies get published (40% vs 60%). Taking the ratio of statistically non-significant OS studies to statistically non-significant traditional studies, we find that for every non-significant traditional study that gets published, six non-significant OS studies get published. In the original model, the probability of publishing a novel non-significant finding was $C_{N-} = 0$. In our extension, we increased this to $C_{N-} = 0.05$ for traditional studies. To incorporate the six-to-one ratio of non-significant studies between traditional and OS, we set $C_{N-} = 0.3$ for OS studies.

3. **Practicing OS leads to papers that are rewarded more (V_{N+}, V_{N-}, V_{R+}, V_{R-}).**
Citation advantages have been observed for several OS practices. In a systematic review, Langham-Putrow et al. (**2021**) identified 64 studies that claim a citation advantage, 37 studies that do not claim a citation advantage, 32 studies that claim a citation advantage in some subfields, and one inconclusive study (see Table 1 in their article). We used these numbers to approximate a value for the citation advantage. We only considered the 64 studies that claim an effect and the 37 studies that do not claim an effect. We assumed that the number of citations for OS papers and non-OS papers come from two Normal distributions. Let X be a random variable of $NOS \sim \mathcal{N}(1, 0.1)$ and let Y be a random variable of $OS \sim \mathcal{N}(c, 0.1)$. Through numerical optimization, we found c such that $P(Y < X) = 64/(37 + 64)$. We found an optimal value of $c = 1.0483$, which led to a citation advantage of:

$$\frac{\frac{c - \mu_{NOS}}{\sigma} + 1}{\mu_{NOS}} = \frac{\frac{1.0483 - 1}{0.1} + 1}{1} = 1.483 \quad (2)$$

We found a 48.3% (i.e., 1.483) citation advantage and used this value in our extended model. Note that this is a non-parametric approach to converting the 64 studies that claim an advantage and 37 studies that claim no advantage into a numeric value. Note also that in this approach, the 37 no-advantage studies are operationalized as OS studies being disadvantaged in terms of citation rate.

We made the following assumptions about the impact of advocating OS on the survival of labs:

1. **Advocating OS leads to spending more time advocating (e.g., on Twitter) and less time doing research (η).**
Advocating OS might lead to less available time for doing research because some proportion of the work time is spent on profiling oneself (e.g., posting on Twitter). Therefore, labs that advocate OS had a higher η than labs that do not. We assumed that labs that advocate OS spend two hours of their work time per week (40 hours) on social media.

Table 2 Parameters for the four lab types.

Par.	Value			
	Practice; advocate	Practice; not advocate	Not practice; advocate	Not practice; not advocate
η	{0.205, **0.210**, 0.215}	0.200	{0.205, **0.210**, 0.215}	0.200
e	{77, **79**, 81}	{77, **79**, 81}	75	75
C_{N-}	{0.175, **0.300**, 0.425}	{0.175, **0.300**, 0.425}	0.050	0.050
V_{N+}, V_{N-}	{1.543, 1.841, 2.142, **2.199**, 2.558, 2.976}	{1.242, **1.483**, 1.725}	{1.242, **1.483**, 1.725}	1.000
V_{R+}, V_{R-}	{0.771, 0.921, 1.071, **1.100**, 1.279, 1.488}	{0.621, **0.742**, 0.863}	{0.621, **0.742**, 0.863}	0.500

Parameter values for main analyses are shown in bold font; parameter values for sensitivity analyses are shown in regular font. η is the influence of effort on productivity; e is effort; C_{N-} is the probability of publishing a negative novel finding; V_{N+} is the payoff for publishing a positive novel finding; V_{N-} is the payoff for publishing a negative novel finding; V_{R+} is the payoff for publishing a positive replication; and V_{R-} is the payoff for publishing a negative replication. Par. = Parameter.

It does not matter to the model how much additional time they spend on social media in their free time. We therefore believe that η increases by 5% when advocating OS (i.e., $40/(40-2) = 1.05$).

2. **Advocating OS leads to papers that are rewarded more (V_{N+}, V_{N-}, V_{R+}, V_{R-}).** We assume that publications from labs that advocate OS are rewarded more because they might be read and cited more often. For example, papers that are shared on Twitter (as done by many OS advocates) have a citation advantage over papers that are not shared (Ladeiras-Lopes et al., **2020**; Luc et al., **2021**). We did not find studies specifically focusing on the citation advantage for sharing OS papers on Twitter or other platforms, so we decided to use a heuristic of equating the payoff advantage for labs that advocate OS with the payoff advantage for labs that practice OS (see above; i.e., 48.3%).

The parameter values of our model extension are summarized in Table 2. To make sure our results are robust and not contingent on specific choices for parameter values, we included two additional parameter values for each parameter and ran the factorial combination of each of these as sensitivity analyses (see Table 2). The parameter values of the sensitivity analyses are always 50% and 150% of the difference between the main parameter values and the reference parameter values.

As a primary result, we collected data on which lab type(s) survive over time and which die out in a world where everyone "plays the game". Specifically, we investigated in what proportions the lab types are present over time: Which lab type(s) is/are most successful within the academic community? As a secondary result, we collected similar data as Smaldino and McElreath (**2016**) about the mean e, mean r, mean α, FDR, and mean W across all lab types. For our simulations, we kept e fixed within lab type (see Table 2).

To reduce the computation time of the simulations, we used a maximum literature size of 100,000 instead of 1,000,000. Moreover, we simulated over 1,000,000 time steps. We sampled every time step for iterations between 1 and 1,000 iterations, every 10th time step for iterations between 1,000 and 10,000, every 100th time step for iterations between 10,000 and 100,000, and every 1,000th time step for iterations between 100,000 and 1,000,000.

Further Exploration

We ran an additional set of simulations that incorporated a few more changes to the computational model. First, we reasoned that the payoff for negative novel studies (i.e., V_{N-}) should probably not be as high as the payoff for positive novel studies (i.e., V_{N+}). In an attempt to

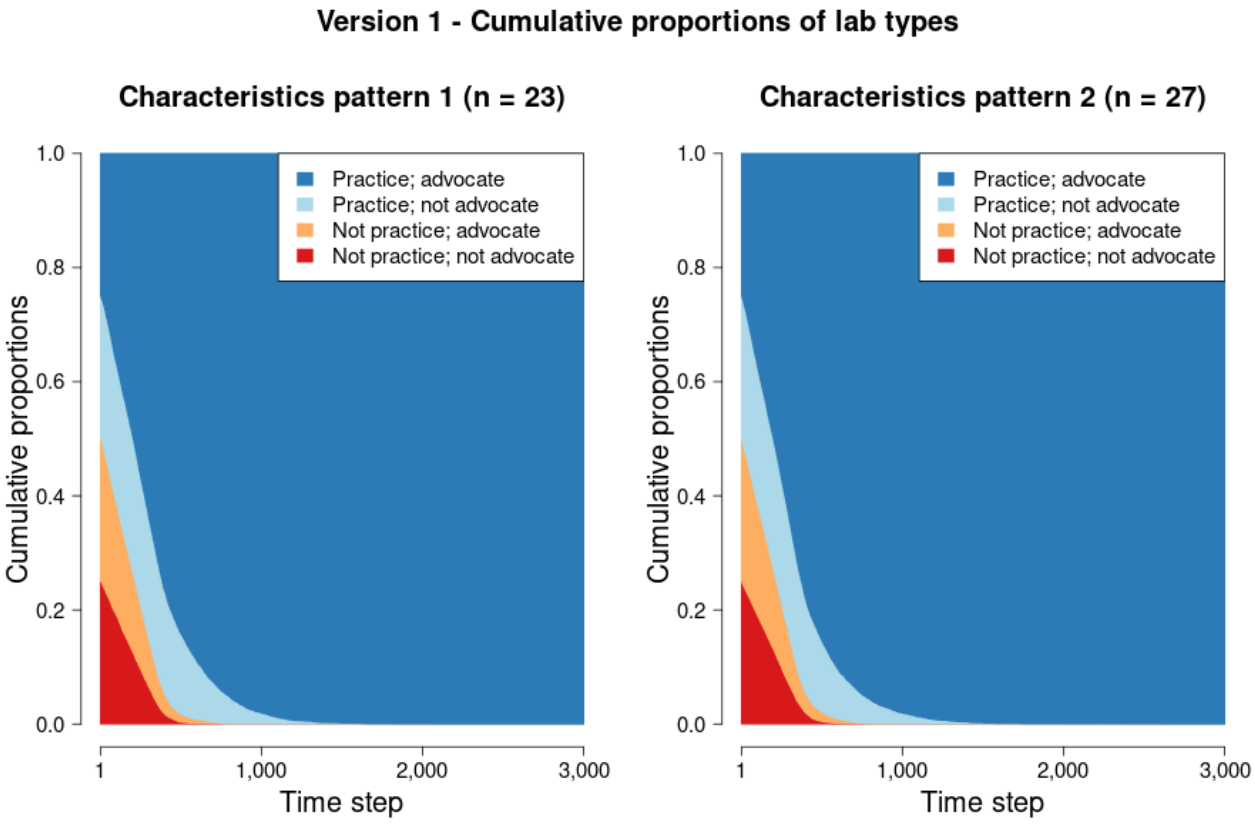

Figure 1 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the first set of simulations. The two panels represent simulation runs with two qualitatively different patterns of characteristics.

get a more informed estimate, we quantified publication advantage using articles published in the New England Journal of Medicine in 2015. Data was previously extracted by Hoekstra et al. (**2018**). In determining whether a study was considered positive or negative, we focused on statistical inference for the primary outcome. We excluded case studies, descriptive studies, non-inferiority trials, and single-arm studies. Next, we counted all citations (as counted through Google Scholar on date December 9, 2022) for the 120 positive results papers and the 42 null result papers. In the past seven years, null result papers were cited a median of 601.5 times and positive trials were cited a median of 826 times. The ratio of these two medians is 0.728 (i.e., positive novel results have a citation advantage of 1-to-0.728). Therefore, we set $V_{N+} = 1$ and $V_{N-} = 0.728$ for labs that do not practice and do not advocate OS.

Second, the results of the previous simulations suggest that the characteristics of the scientific community have not yet reached a steady state after 1,000,000 time steps (see Figure 2). To investigate this further, we decided to increase the mutation probabilities for r and W from $\mu_r = \mu_W = 0.01$ to $\mu_r = \mu_W = 0.1$, so that lab characteristics change more quickly.

Results

Looking at the development of the lab characteristics in the 50 simulation runs separately, we noticed that individual simulations resulted in one of two qualitatively different patterns of characteristics (i.e., FDR, W, α, and r), which are described below. To acknowledge this dichotomy, instead of averaging all 50 simulation

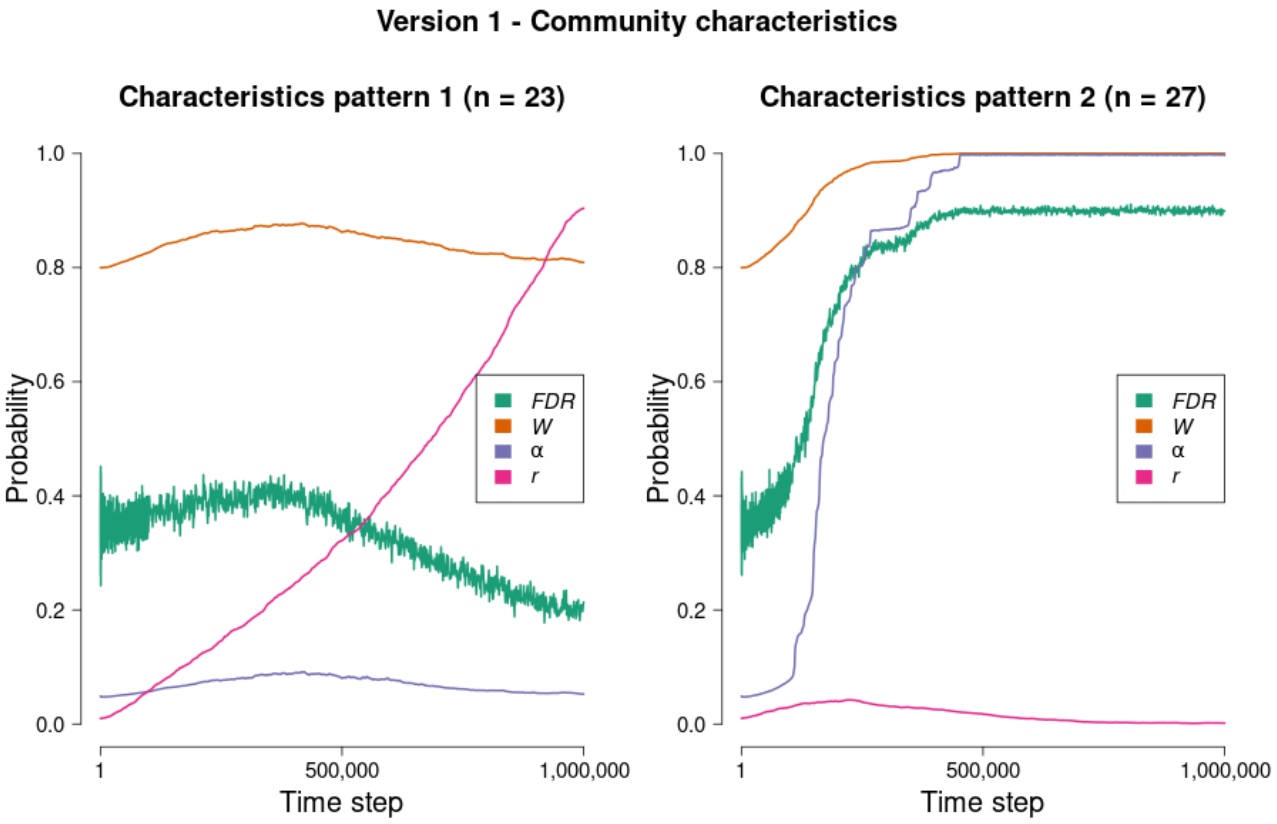

Figure 2 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over labs and simulation runs over all $1,000,000$ time steps of the first set of simulations. The two panels differentiate between simulation runs with two qualitatively different patterns of characteristics.

runs, we split them according to the pattern of characteristics observed and averaged them separately. This applies to both the proportions of lab types as well as the community characteristics. Accordingly, for our results, we always provide one plot for simulations that exhibited characteristics pattern 1 and one plot for simulations that exhibited characteristics pattern 2.

Who Dominates Science?

Figure 1 shows the proportions of the four lab types over the first $3,000$ time steps. The two panels differentiate between simulation runs that displayed two qualitatively different patterns of characteristics, explained in the next section (see also Figure 2). Labs that practice and advocate OS reach a proportion of 1 very quickly; at the same time, the other three lab types vanish. Of those, labs that do not practice OS and do not advocate OS disappear most quickly, followed by labs that do not practice OS but advocate OS and labs that practice OS but do not advocate OS.

The observed behavior indicates that practicing OS is more important than advocating OS, but that doing both is most advantageous. This advantage of practicing over advocating holds across the entire range of parameter values we investigated (see Figures 5, 6, and 7 in Appendix A.1). The explanation for this is that there is an additional advantage for labs that practice OS, which is the higher probability of publishing a negative novel finding (i.e., C_{N-}, 0.3 versus 0.05). Furthermore, the same behavior is observed for various values of V_{0-} (-5, -25, -50, -100; see Figure 17 in Ap-

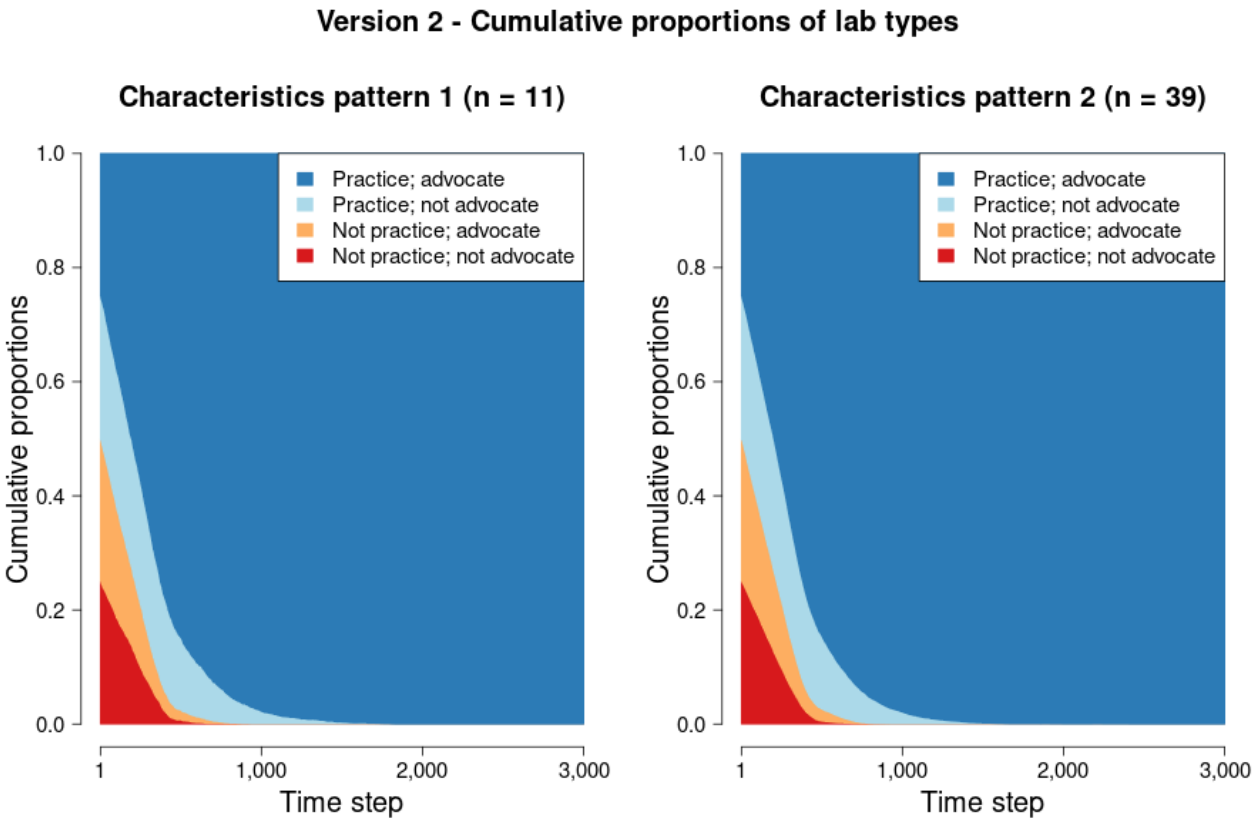

Figure 3 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the first set of simulations. The two panels represent simulation runs with two qualitatively different patterns of characteristics.

pendix A.2). In sum, in our evolutionary model the additional cost in terms of increased effort (practicing) and a reduction of work time spent on the actual research (advocating) is easily outweighed by the increased payoff when the work gets published.

Characteristics of the Scientific Community

Figure 2 shows the development of characteristics across lab types over the whole range of 1,000,000 time steps. As all lab types except for the "practice; advocate" lab type ceased to exist within around 3,000 time steps, the development of characteristics shown in Figure 2 almost exclusively reflects the "practice; advocate" lab type. Here, simulations can be differentiated by two qualitatively distinct patterns of characteristics. In the left panel, it can be seen that W increases slightly and then decreases very slowly to the initial value. Similarly, α remains fairly constant. FDR rises a bit and then declines over time. Lastly, r increases strongly to almost 1 in an almost linear fashion. An explanation for this is that r increases to a critical value, at which point the values of W and α do not matter (enough) anymore. Recall that for a replication, the type of result does not matter in terms of payoff. As such, r grows to 1. In this variant, mutations are such that the certain payoff of replications was higher than the variable payoff of novel results, with a relatively high occurrence of null results. The reader may notice that W does not grow quite so strongly in the first time steps, giving r enough time to gain momentum.

The right panel displays entirely different characteristics. Here, both W and α increase rapidly to 1 and remain constant. FDR also

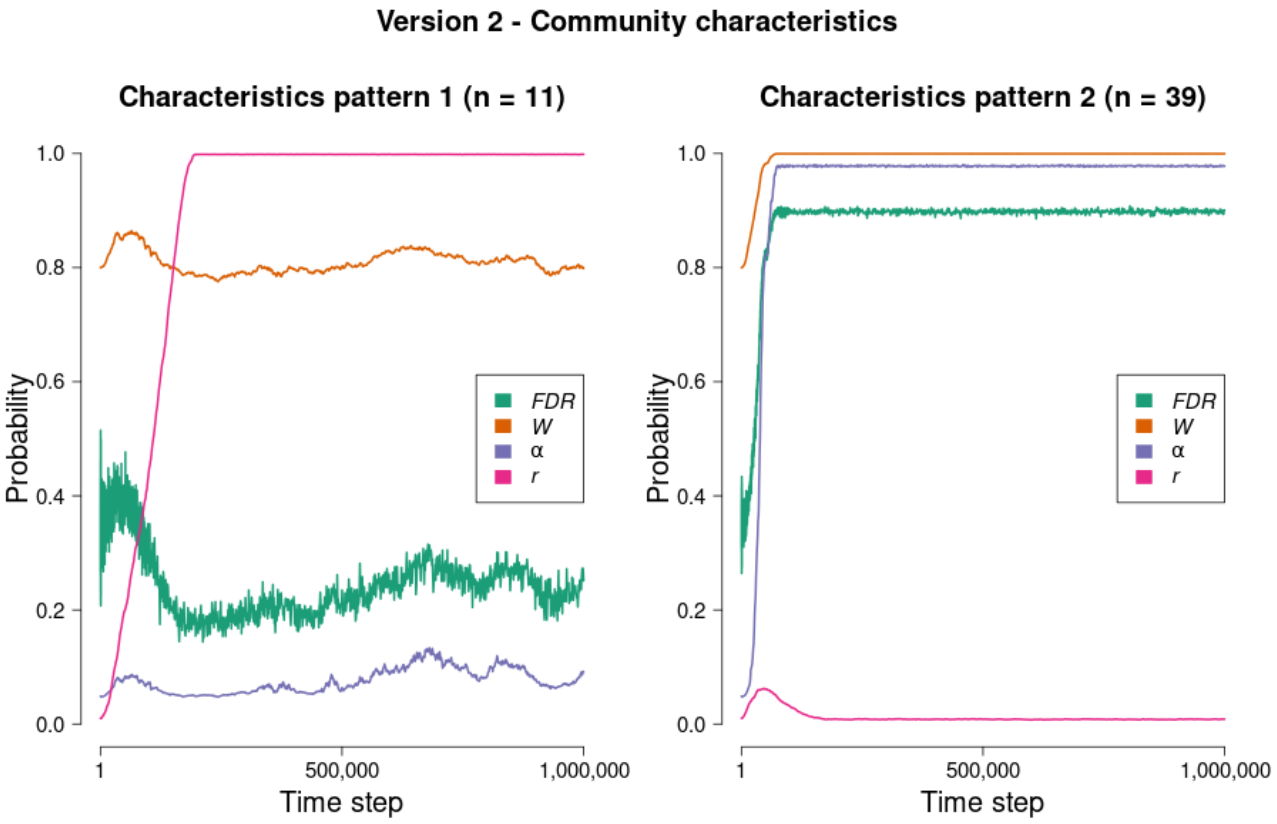

Figure 4 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over labs and simulation runs over all $1,000,000$ time steps of the second set of simulations. The two panels differentiate between simulation runs with two qualitatively different patterns of characteristics.

increases strongly but reaches a plateau at around 0.85. In contrast, r remains very low at almost 0. The explanation is straightforward: if W and α are 1, all results are positive results by default, meaning that every lab should conduct novel studies over replications. This is the case because the payoff is double, and there is no drawback. Note that this second pattern of results was not obtained in Kohrt et al. (**2022**) as they fix power in their simulations.

Further Exploration

We further explored the proportions of lab types and the characteristics of the scientific community with some slight parameter modifications: Changing μ_W and μ_r from 0.01 to 0.1 and changing V_{N-} from 1 to 0.728. Figure 3 shows the proportions of lab types over time. The behavior is very similar to the previous simulations in Figure 1, such that labs that practice and advocate OS dominate very quickly while the other lab types die out. Once again, these results are robust to different choices of parameter values (see Figures 11, 12, and 13 in Appendix A.1).

Figure 4 shows the characteristics of the scientific community. As in the first set of simulations (see Figure 2), the characteristics reflect those of the labs that practice and advocate OS through most of the time. As in the first set of simulations, we observed two qualitatively different patterns of characteristics of the labs. The explanation for this is the same as for the first set of simulations (see previous section).

Discussion

Science is not just about the academic work– it is ultimately a joint enterprise by people. People who depend on their academic position for their livelihood. As such, doing well, or at least doing better than others, on whatever metric is used to evaluate one's success becomes important to people. Smaldino and McElreath (**2016**) demonstrate that if all people do is "play the game", the scientific work their field produces over time gradually degenerates to low-effort, quick-and-dirty work with a high proportion of false positives.

The OS movement should restrict the feasibility of some of the quick-and-dirty strategies to which researchers might, inadvertently, fall prey to. For instance, preregistering one's work makes it very difficult to employ *p*-hacking or to hypothesize after the results are known. That said, practicing OS brings with it its own set of success indicators, such as prestige in the field, exclusive funding opportunities, and increased visibility of the work; it may therefore well be optimal to continue to "play the game", just with a slightly adjusted set of rules. In other words, exploiting the incentive structure for OS practices might lead to receiving the same advantages as actually practicing OS. In this study, we explored how different types of academics would thrive in a scientific system that values OS practices. Namely, we compared labs that practice or do not practice OS and labs that advocate or do not advocate OS.

In an incentive structure that values OS practices, practicing OS while also advocating OS is most advantageous. Our simulation results suggest that labs that follow OS practices and engage as "OS advocates" on Twitter or related social media platforms have a survival advantage. The cost associated with both practicing and preaching in terms of a slower "rate of completion" of research projects gets outweighed by the increase in payoff for publications. Within the simulation, only labs that both practiced and advocated OS persisted, all other types quickly vanished from the scientific landscape (i.e., they were less successful in terms of attracting attention and gathering citations).

In our model, advocating OS did not have the same advantages as practicing OS. This was true even in the condition where advocating was worth more than practicing in terms of publication payoff (73% versus 48%, respectively). The likely reason for this lies in the probability of being able to publish negative or null results: in our parameterization, this probability was six times higher for OS work. Many journals that align with OS principles vouch to judge the publishability of a work based on the soundness of the research question, the methodology, and the study protocol. More traditional journals, in addition to these criteria, tend to lean heavily on the (statistical) significance of the results.

Practicing and incentivizing OS practices did not eliminate quick-and-dirty science in our simulation. While we observed slightly lower values for the false discovery rate and false positive rate compared to Smaldino and McElreath (**2016**) and the replication (Kohrt et al., **2022**), the values were still considerably high (0.7 compared to > 0.8; but see Appendix A.1). Without changing the incentive structure that is currently valuing quantity over quality, OS cannot prevent the rise of quick-and-dirty science. As Simine Vazire once put it (Vazire, **2020**), OS does not act as quality control itself but enables quality control.

Limitations

Although in our model, practicing and advocating OS practices translated to career advantages, some factors cast doubt on the extent to which these findings translate to the real world. Our operationalization of practicing OS involved (slightly) more work per project and a substantial increase in pay-off per publication. Although our parameter settings were grounded to some extent on previous literature, the exact size will be no more than a rough estimate. Perhaps an increase in workload of, say, 50% would be more realistic than an increase of 5%, making the practicing of OS far less attractive for purposes of furthering one's career.

For our operationalization of advocating OS, we assumed that scientists spend two hours of their working weeks on their social media of choice in lieu of working to build and maintain their OS profile. Perhaps two hours is unrealistic and ten hours gets closer to the truth. Or perhaps there is no difference at all in hours spent working between active social media scientists and those that are not active on social

media: time on social media could be spent entirely during free time.

Similarly, it can be argued that the payoff advantage for labs that practice and/or advocate OS is too high and that it is unrealistic that the payoff advantage remains constant throughout evolution. It is possible that it is more realistic for the payoff advantage to diminish over time as an increasing amount of conducted studies are OS studies.

An additional limitation is the omission of consideration regarding the potential consequences of disclosing specific research data. Opening up access to such data may facilitate the identification of inaccuracies and provide a basis for heightened scrutiny from the broader research community (Allen & Mehler, **2019**). Theoretically, this increased transparency could adversely affect the sustainability of a research laboratory, particularly if (unintentional) errors are unveiled and subject to public discourse.

Another, more general, limitation of our setup was the generic classification of scientists as OS practitioners versus non-practitioners and advocators versus non-advocators. In the real world, these categories (and the payoffs they entail) will rarely be so black-and-white. In addition, there will be individual differences in academic success that are tangential to the four categories specified here due to field of interest, background, social network, and even luck. In our simulation, these natural sources of variation were all completely equated. As such, the results of this study should be thought of more as a proof of concept in a drastically simplified representation of what in reality is a very complicated academic ecosphere.

Lastly, our modeling approach to investigate what lab types survive in a community that values open science is only one of many. Different approaches may shed light on the conditions under which labs with different strategies flourish. For instance, using a game theory approach would model the individual labs as rational agents with (potentially) different strategies (e.g., affinity to OS, payoff, etc.). In such an approach, the individual labs would play the "science game", which would allow the computation of an equilibrium distribution of different types of labs.

Conclusion

In our simulation, labs that practice and advocate OS thrive in a scientific community that values OS. At the same time, "quick-and-dirty" science is still prevalent, as evident by high false positive and false discovery rates. These results are encouraging to those who feel practicing open science "is not worth it": in addition to benefits to science at large, our results suggest engaging with OS benefits the individual researcher as well.

Acknowledgements

We are grateful to Joyce M. Hoek, Jasmine Muradchanian, and Ymkje Anna de Vries for interesting and inspiring discussions.

Protocol, Code, and Data Availability

A transparency documentation of our research process, the code for the simulations, and the data of the simulations can be found online at https://osf.io/h5tfv/.

References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), 1–14. https://doi.org/10.1371/journal.pbio.3000246 (see pp. 82, 85, 92).

Baker, M. (2016). Dutch agency launches first grants programme dedicated to replication. https://doi.org/10.1038/nature.2016.20287 (see p. 82).

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016 (see p. 82).

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7 (see p. 85).

Fiedler, K. (2011). Voodoo correlations are everywhere - not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. https://doi.org/10.1177/1745691611400237 (see p. 82).

Guzman-Ramirez, L., Schettino, A., Sweeney, J., & Sunami, N. (2023). Badges to reward open & responsible research practices [Publisher: Zenodo]. https://doi.org/10.5281/zenodo.8278785 (see p. 82).

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & Wagenmakers, E.-J. (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PLoS ONE*, *13*(4), e0195474. https://doi.org/10.1371/journal.pone.0195474 (see p. 87).

Hostler, T. J. (2023). The invisible workload of open research [Publisher: JOTE Publishers]. *Journal of Trial & Error*. https://doi.org/10.36850/mr5 (see p. 84).

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https:/ / doi.org/ 10.1371/ journal.pmed.0020124 (see p. 82).

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), e1002456. https://doi.org/10.1371/journal.pbio.1002456 (see p. 82).

Kohrt, F., Smaldino, P. E., McElreath, R., & Schönbrodt, F. D. (2022). Replication of the natural selection of bad science. https://doi.org/10.31222/osf.io/sjyp3 (see pp. 90, 91).

Ladeiras-Lopes, R., Clarke, S., Vidal-Perez, R., Alexander, M., Lüscher, T. F., & On behalf of the ESC (European Society of Cariology) Media Committee and European Heart Journal. (2020). Twitter promotion predicts citation rates of cardiovascular articles: A preliminary analysis from the ESC journals randomized study. *European Heart Journal*, *41*(34), 3222–3225. https://doi.org/10.1093/eurheartj/ehaa211 (see p. 86).

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PLOS ONE*, *16*(6), 1–20. https://doi.org/10.1371/journal.pone.0253129 (see p. 85).

Luc, J. G. Y., Archer, M. A., Arora, R. C., Bender, E. M., Blitz, A., Cooke, D. T., Hlci, T. N., Kidane, B., Ouzounian, M., Varghese, T. K., & Antonoff, M. B. (2021). Does tweeting improve citations? one-year results from the TSSMN prospective randomized trial. *The Annals of Thoracic Surgery*, *111*(1), 296–300. https://doi.org/10.1016/j.athoracsur.2020.04.065 (see p. 86).

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, *16*(1), 274. https:/ / doi.org/ 10.1186/ s13059-015-0850-7 (see p. 82).

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, *5*, e16800. https://doi.org/10.7554/eLife.16800 (see pp. 82, 83).

Mikki, S. (2017). Scholarly publications beyond paywalls: Increased citation advantage for open publishing. *Scientometrics*, *113*(3), 1529–1538. https:/ / doi.org/ 10.1007/ s11192-017-2554-0 (see p. 82).

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 Behavior*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021 (see p. 82).

National Academies of Sciences, Engineering, and Medicine. (2019). Improving reproducibility and replicability. In *Reproducibility and replicability in science* (pp. 105–142). National Academies Press (US). (See p. 82).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716 (see p. 82).

Pashler, H., & Wagenmakers, E.-J. (2012). Editor's introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253 (see p. 82).

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research worklow: Better science but more work. *Royal Society Open Science*, *9*, 211997. https://doi.org/10.1098/rsos.211997 (see p. 84).

Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2020). (Re)Building trust? journals' open science badges influence trust in scientists. https://doi.org/10.23668/PSYCHARCHIVES.3364 (see p. 82).

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. 82).

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384. https://doi.org/10.1098/rsos.160384 (see pp. 82, 83, 84, 86, 91).

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. In *Stevens' handbook of experimen-*

tal psychology and cognitive neuroscience (pp. 1–47). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119170174.epcn519 (see p. 82).

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*, 553–571. https://doi.org/10.1007/s10869-020-09695-3 (see pp. 84, 85).

Vazire, S. (2020). Open scholarship: Where are the self-correcting mechanisms of science? Retrieved December 12, 2022, from https://www.google.com/search?q=simine+vazire+opening+the+hood&oq=simine+vazire+opening+the+hood+&aqs=chrome..69i57j33i160l2.4891j0j7&sourceid=chrome&ie=UTF-8#fpstate=ive&scso=_ouyWY8G6NKWR9u8P5fSGqAo_31:0&vld=cid:4a209471,vid:Vfc98WDfDJE,st:752 (see p. 91).

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. 82).

Appendices

A. Sensitivity Analyses for Lab Proportions and Characteristics

For all sensitivity analyses, we did not differentiate between simulation runs with two qualitatively different patterns of characteristics (see Results section). Instead, we averaged over all simulation runs. Each of the following Figures contains various parameter combinations. One additional parameter (i.e., the payoff advantage for advocating OS γ) differentiates between Figures. Figures 5, 6, and 7 show the lab proportions for the first set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively; Figures 8, 9, and 10 show the community characteristics for the first set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively. Similarly, Figures 11, 12, and 13 show the lab proportions for the second set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively; Figures 14, 15, and 16 show the community characteristics for the second set of simulations with payoff advantages for advocating OS of $\gamma = \{1.242, 1.483, 1.725\}$, respectively.

Figures 5, 6, 7, 11, 12, and 13 clearly demonstrate that the lab proportions are robust against specific choices of parameter combinations. In all cases, the "practice; advocate" lab type wins and suppresses the other lab types. Although there is more variation in the community characteristics for different parameter combinations (see Figures 8, 9, 10, 14, 15, and 16), the overall trends are still quite robust.

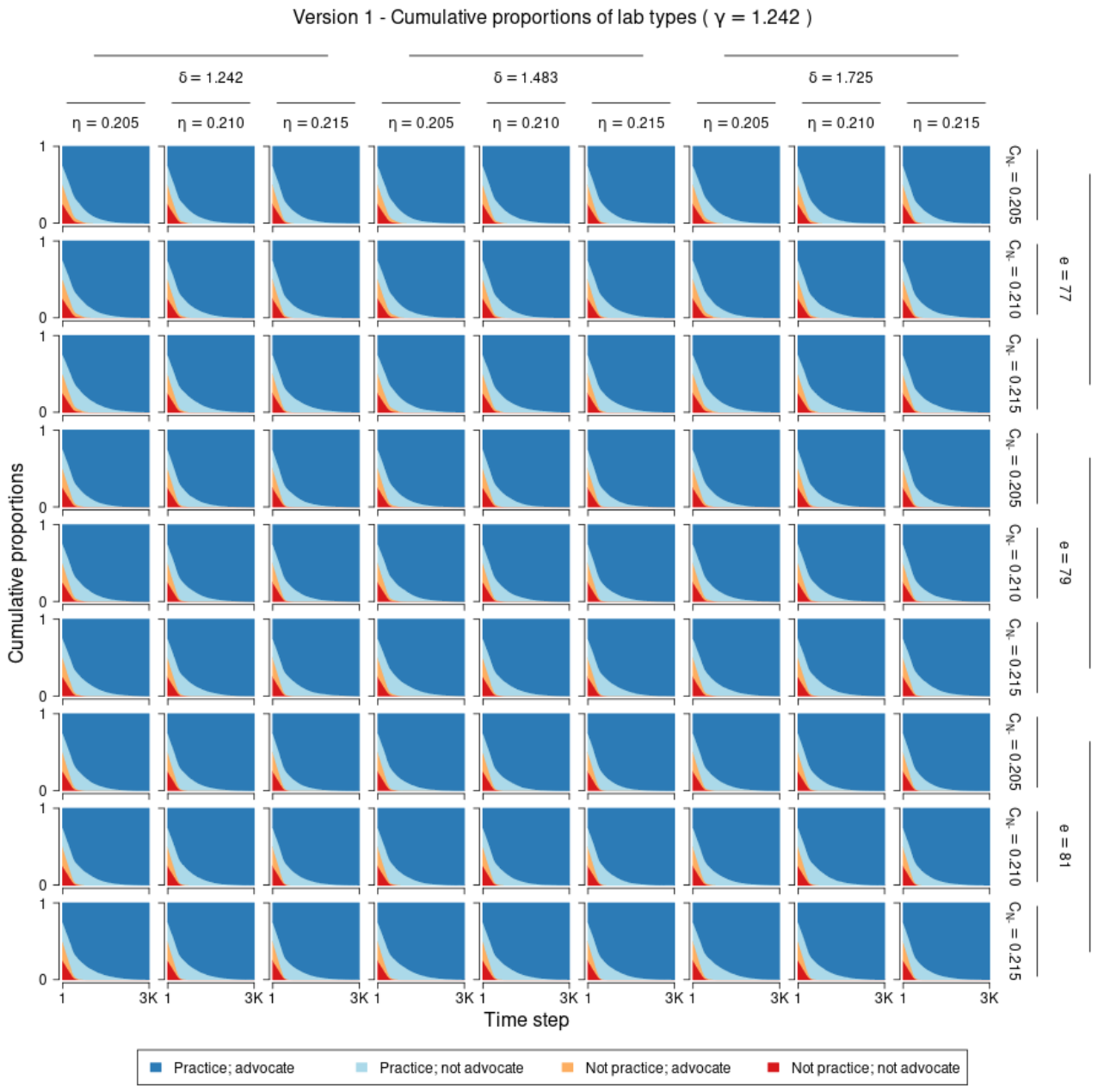

Figure 5 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the first set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

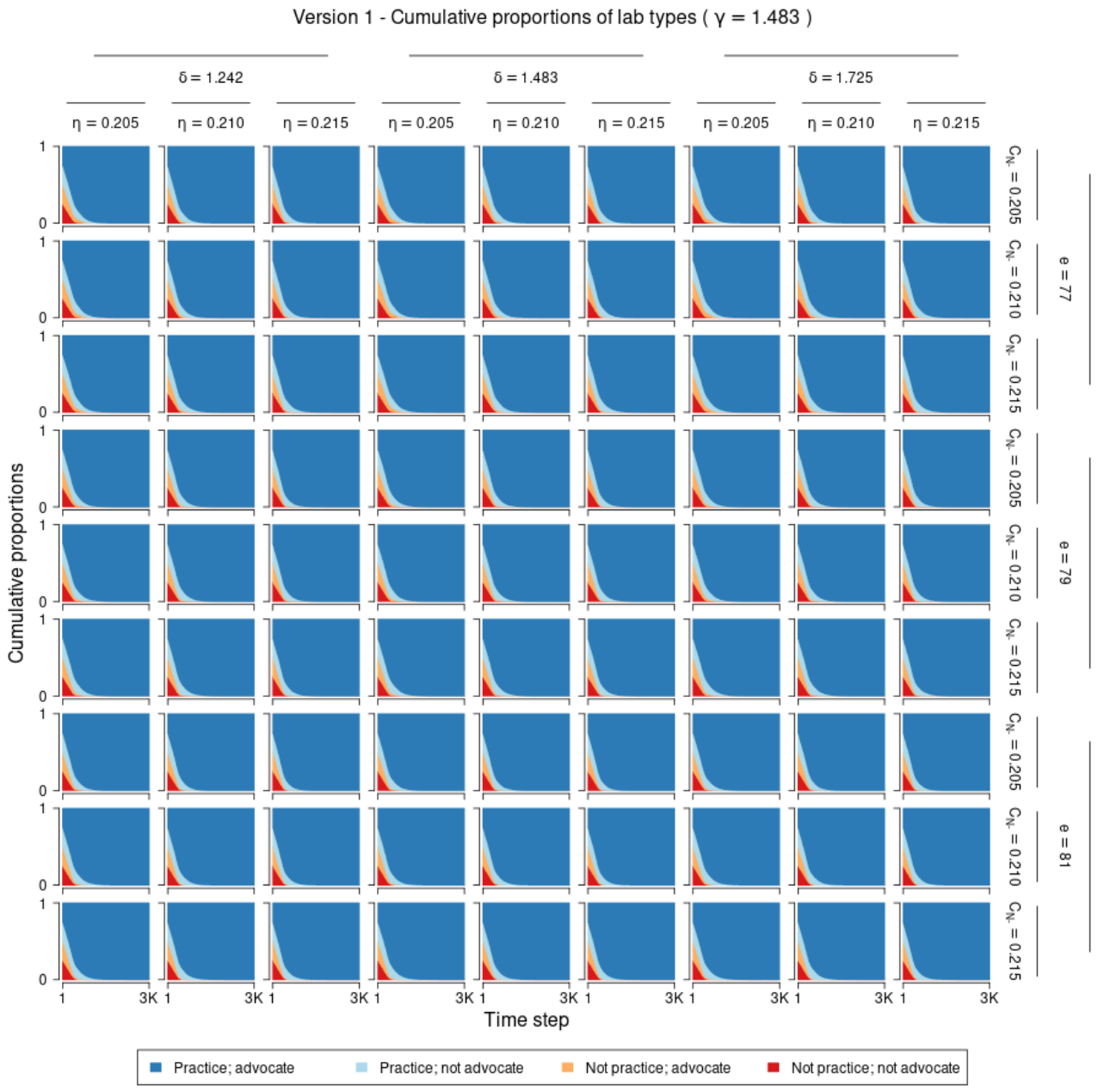

Figure 6 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the first set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

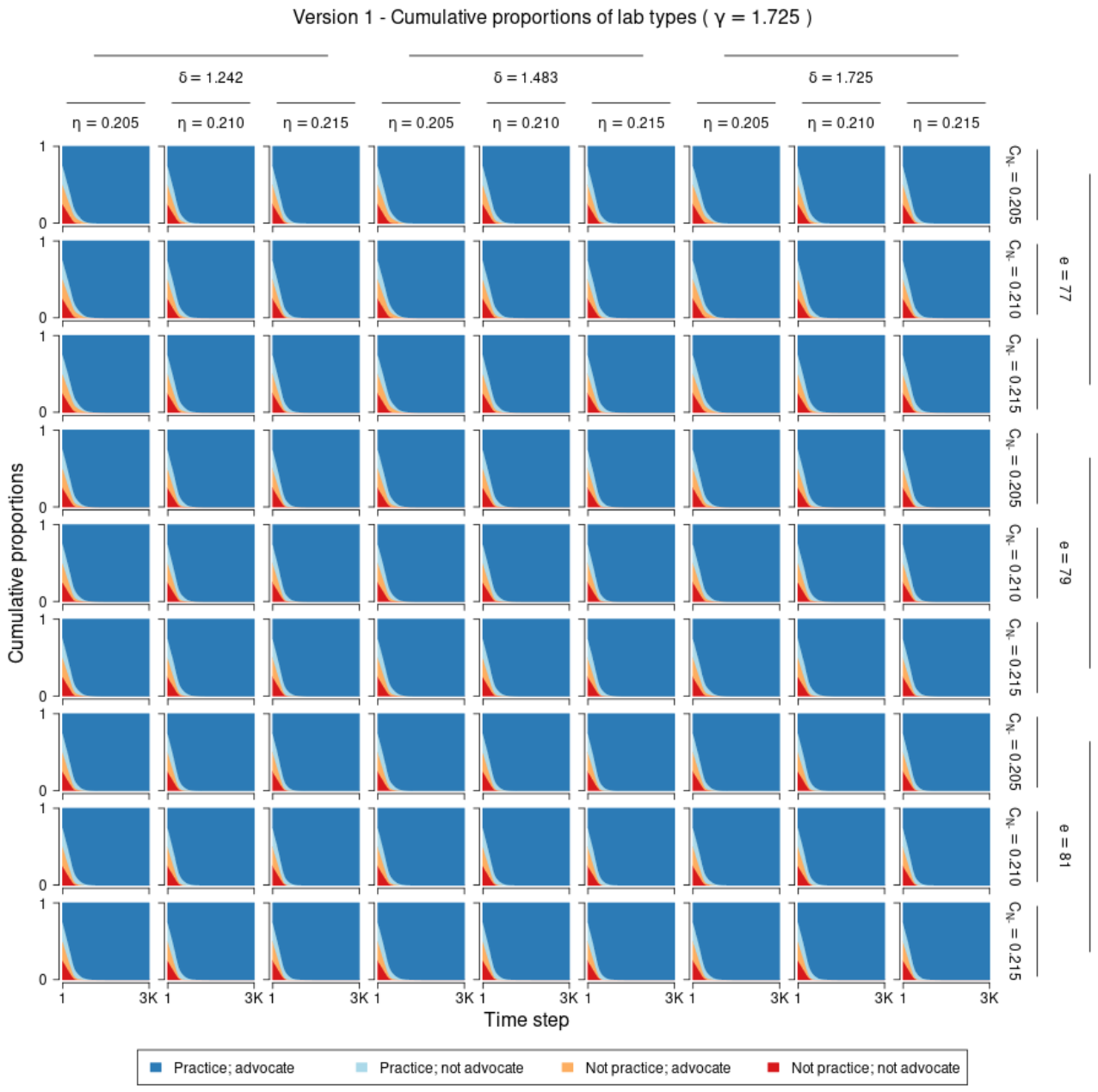

Figure 7 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the first set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

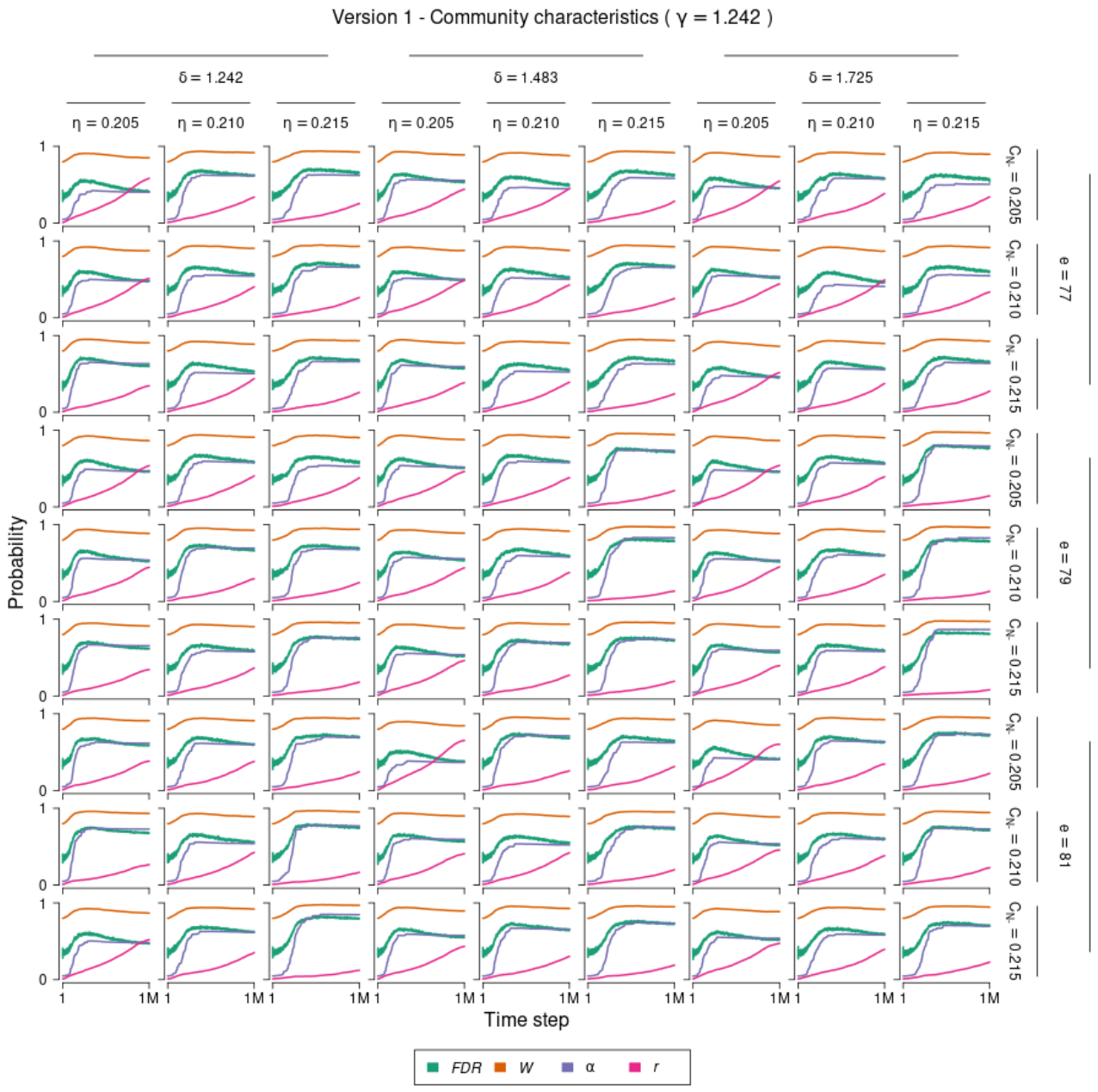

Figure 8 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

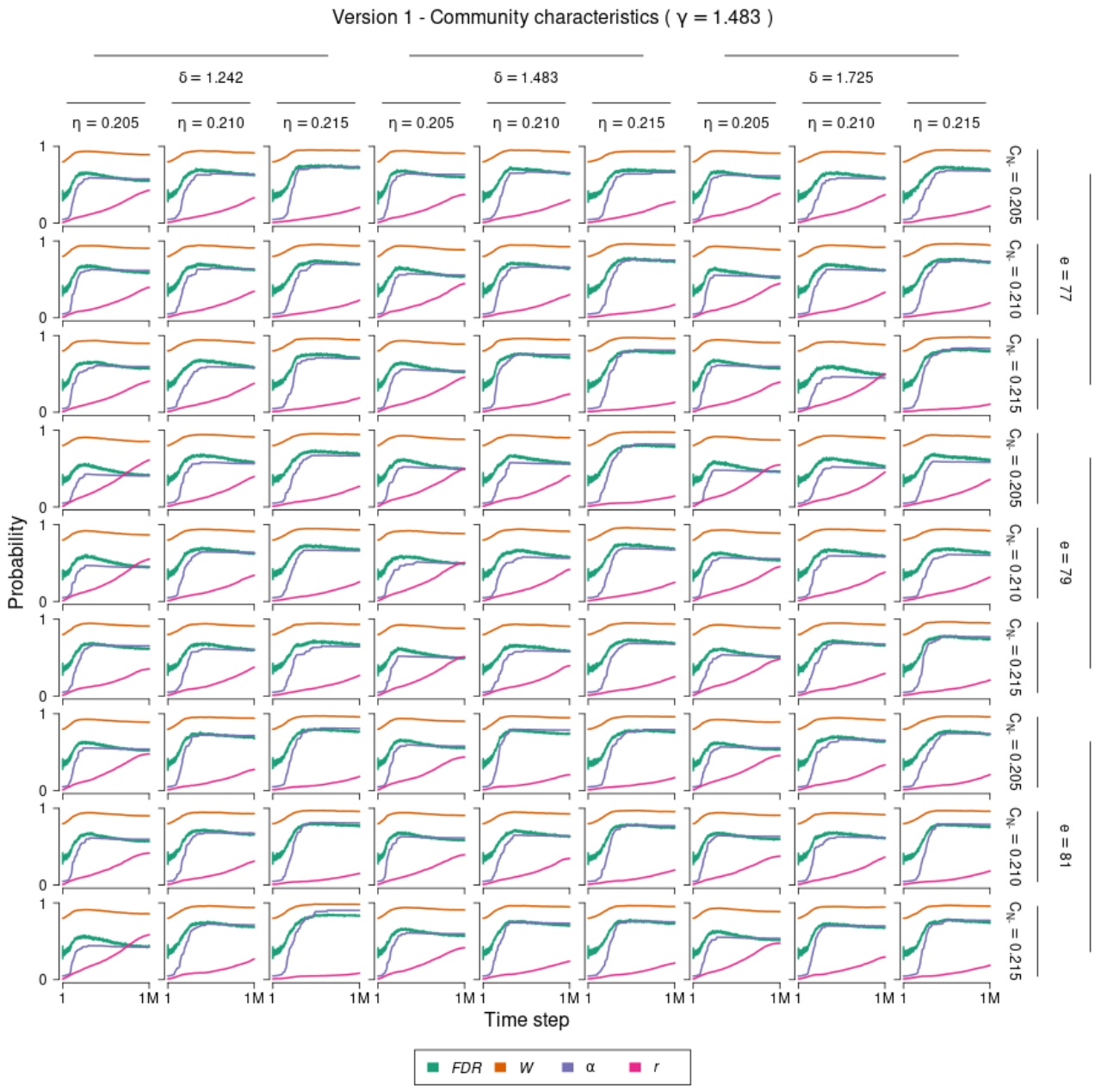

Figure 9 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

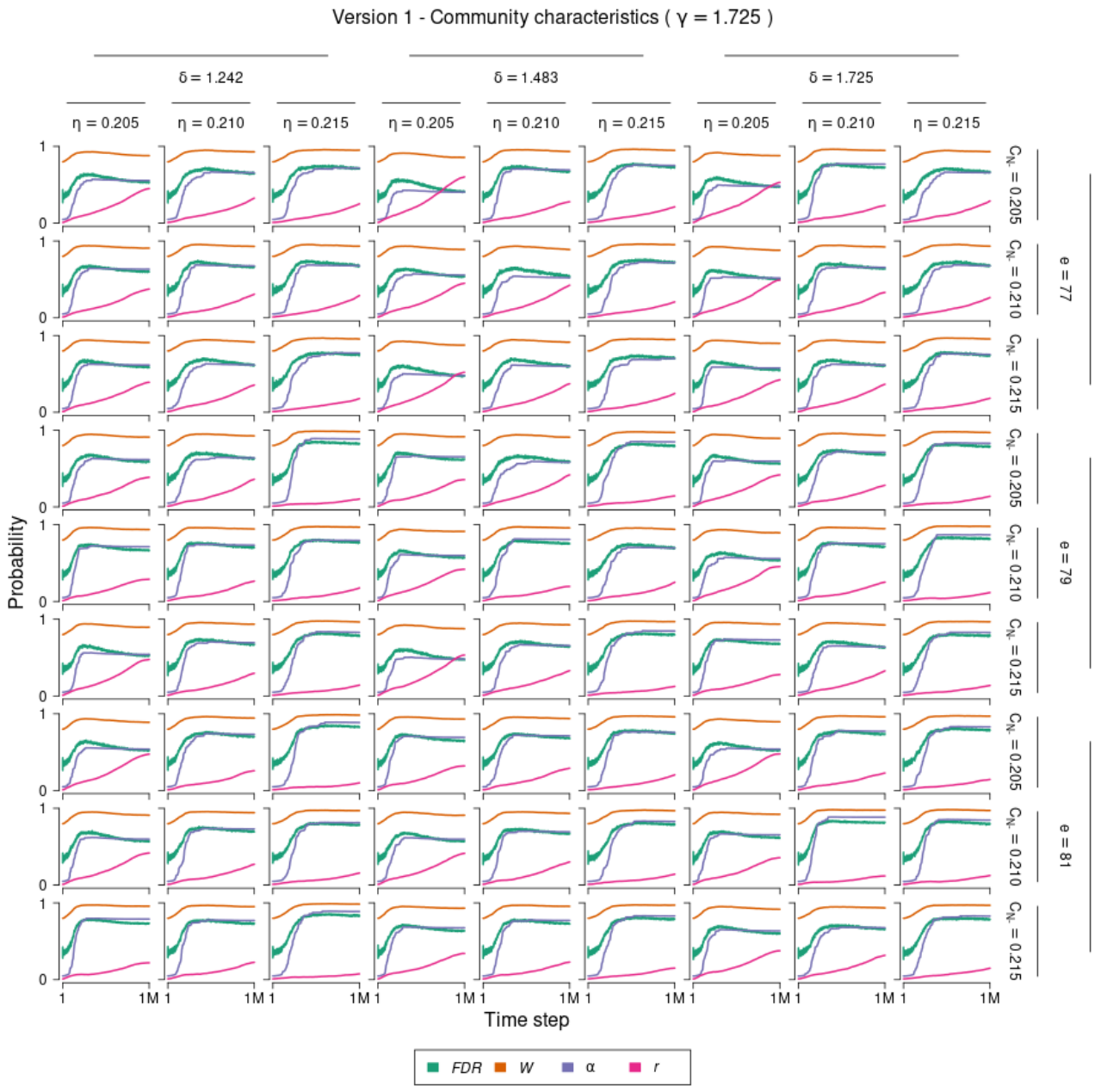

Figure 10 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the first set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

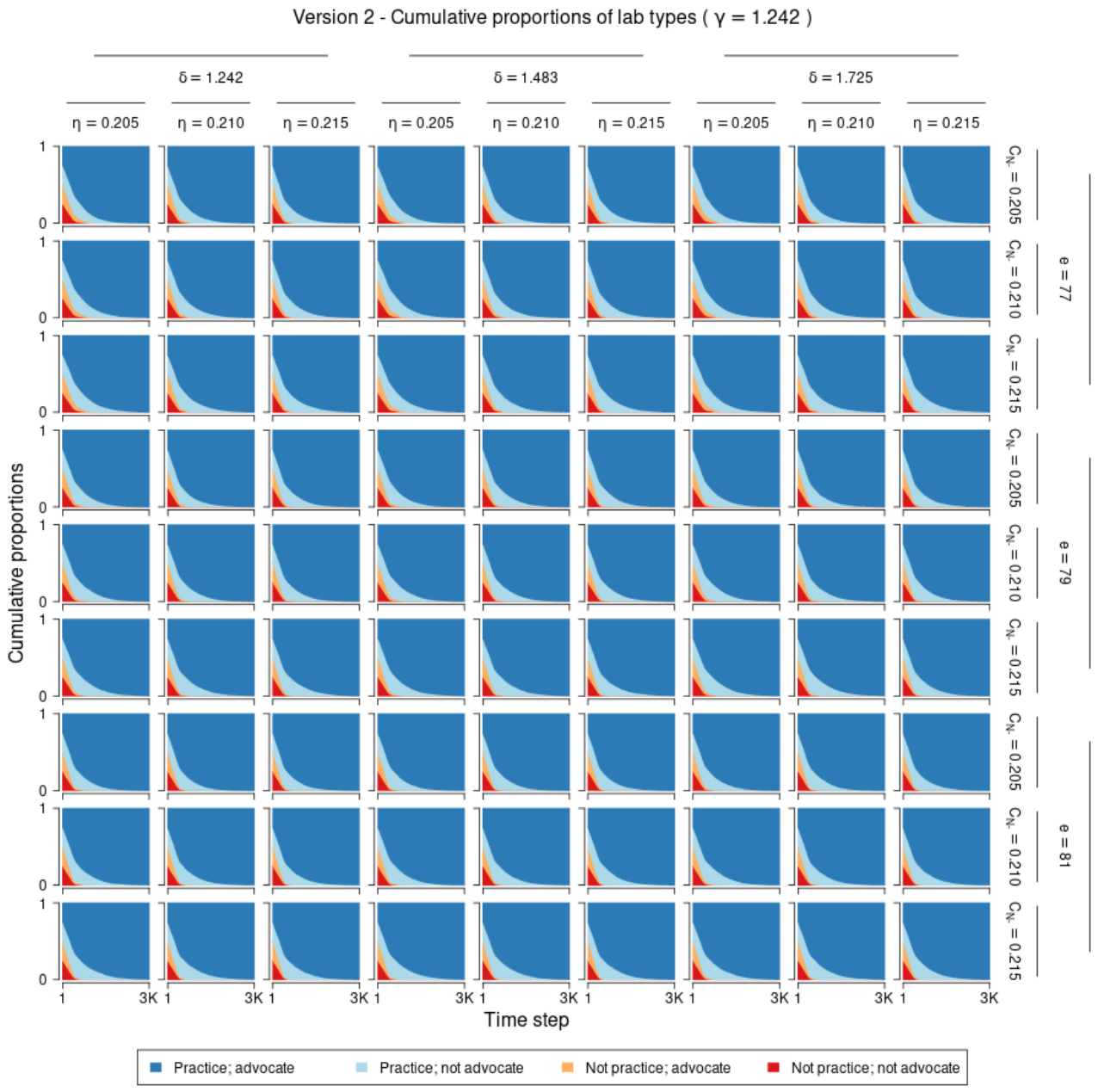

Figure 11 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the second set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

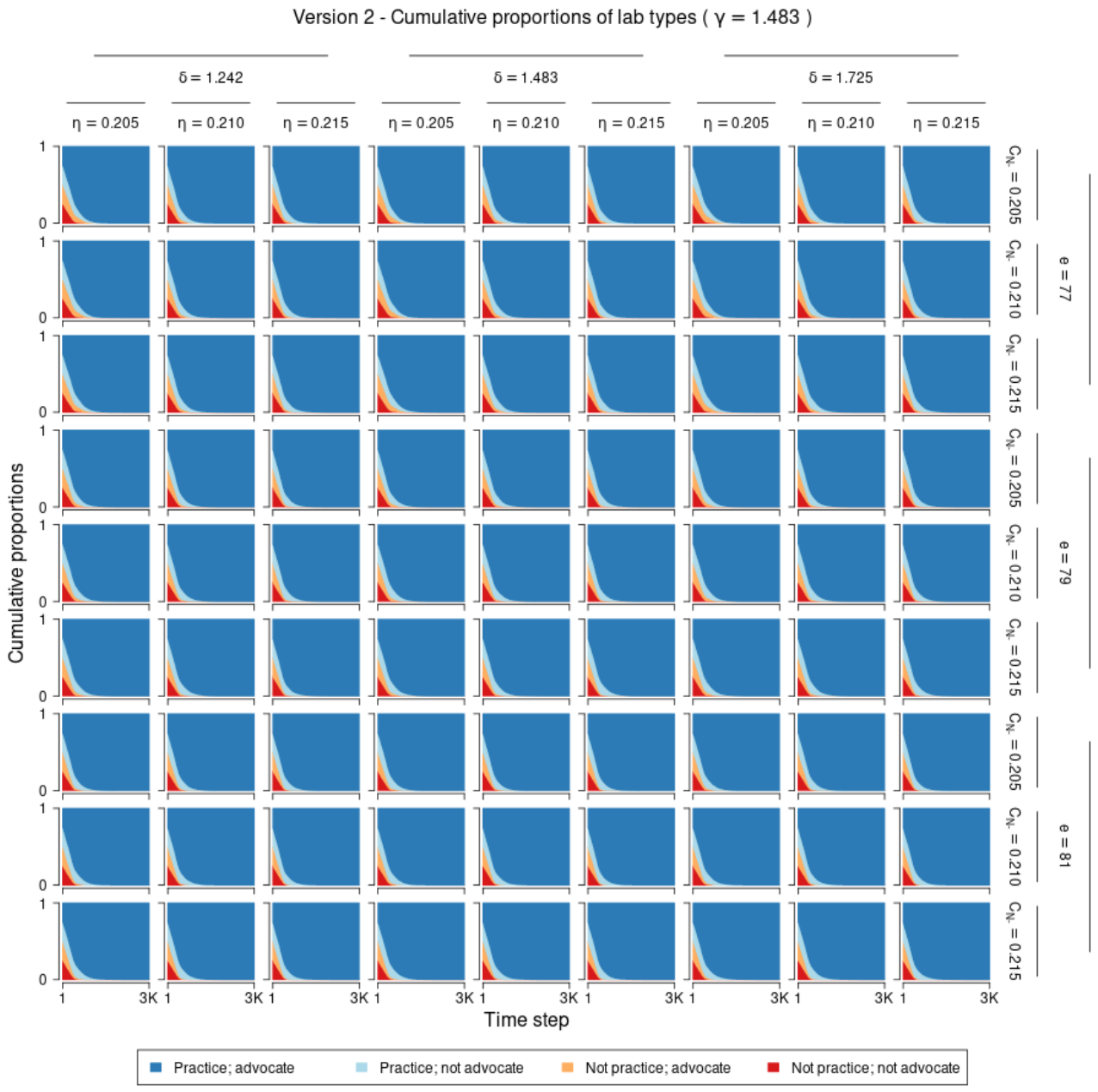

Figure 12 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the second set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

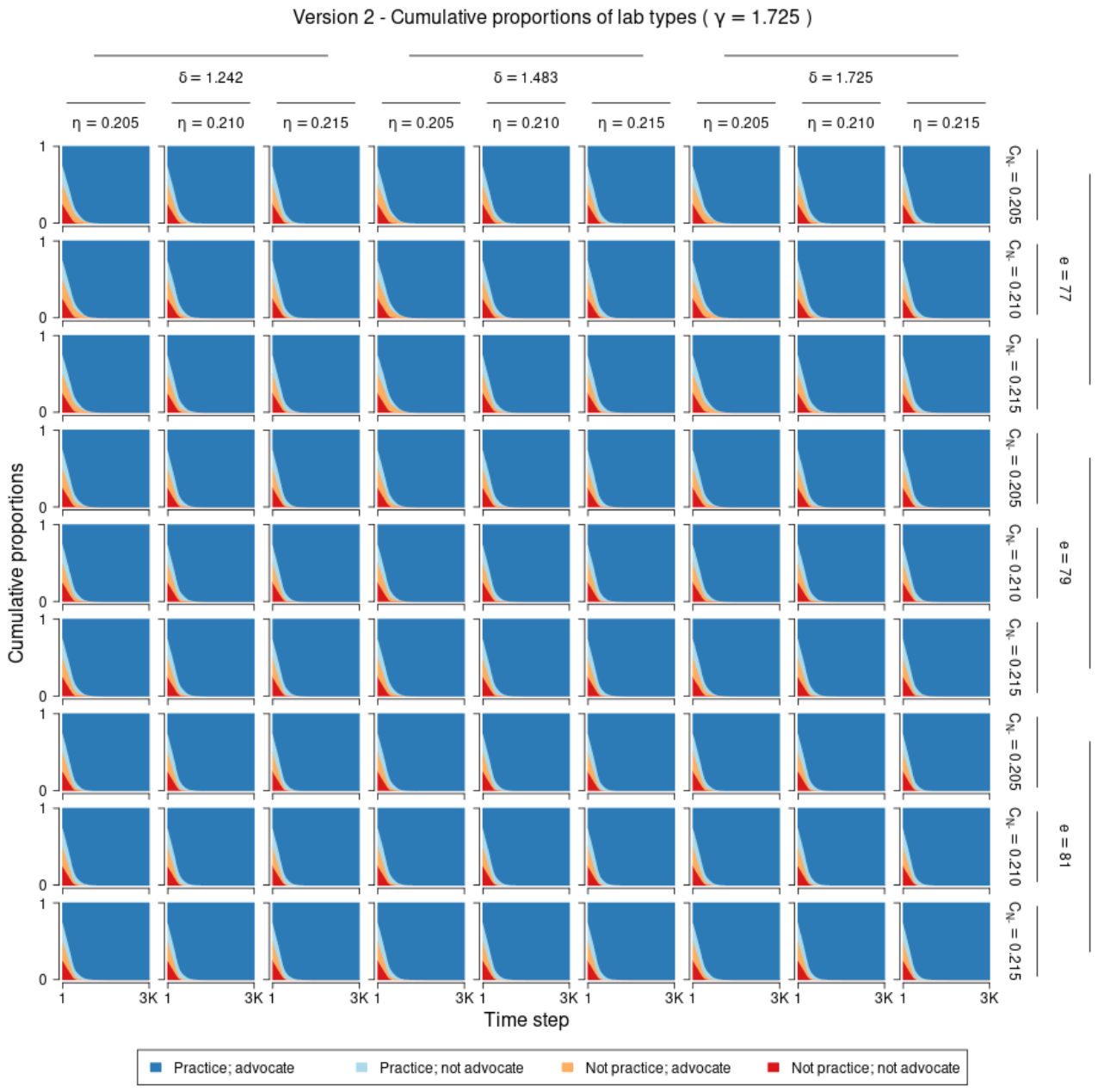

Figure 13 Cumulative proportions of lab types during the first 3,000 out of 1,000,000 time steps of the second set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

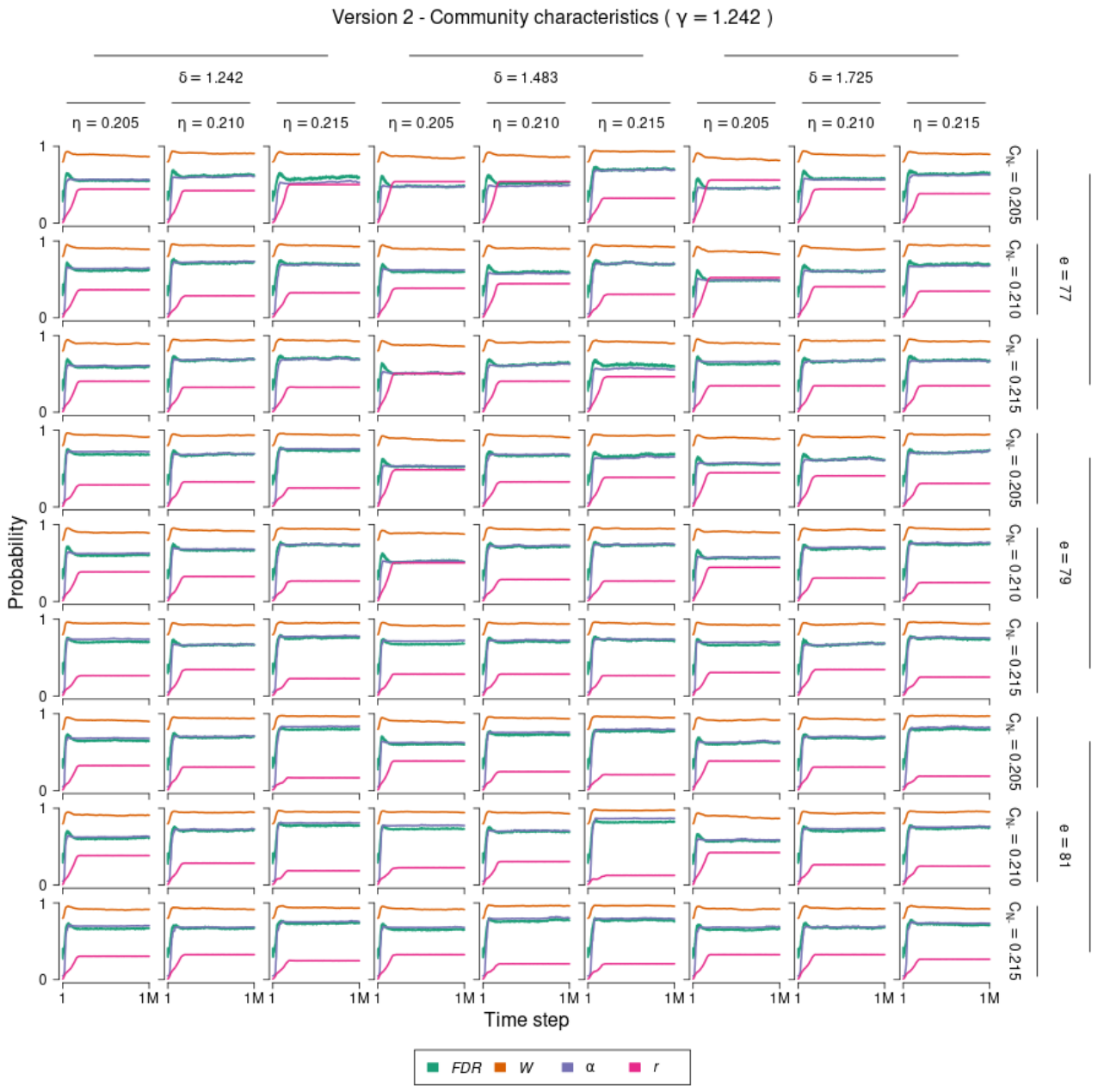

Figure 14 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all $1,000,000$ time steps of the second set of simulations with $\gamma = 1.242$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

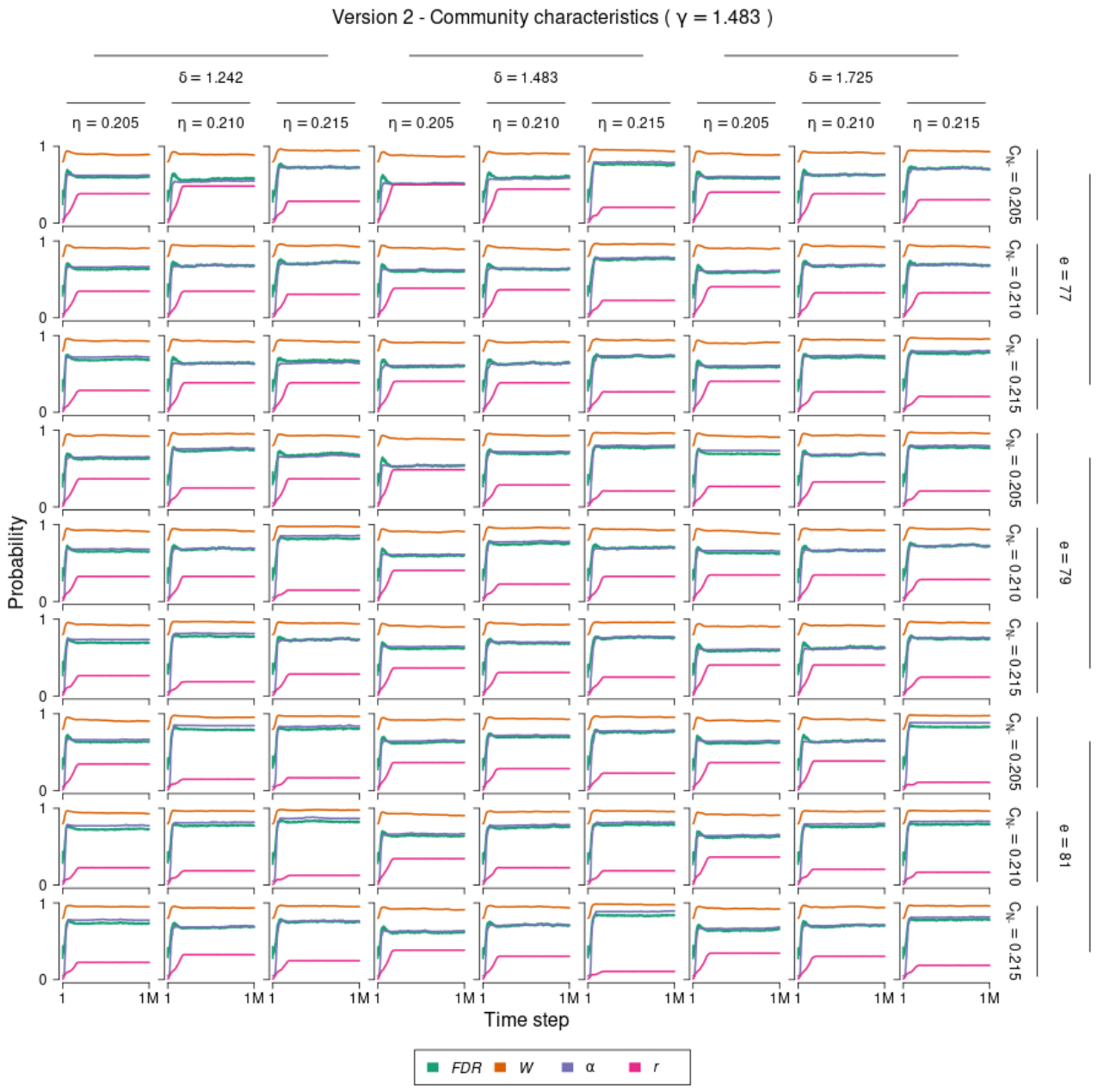

Figure 15 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all 1,000,000 time steps of the second set of simulations with $\gamma = 1.483$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

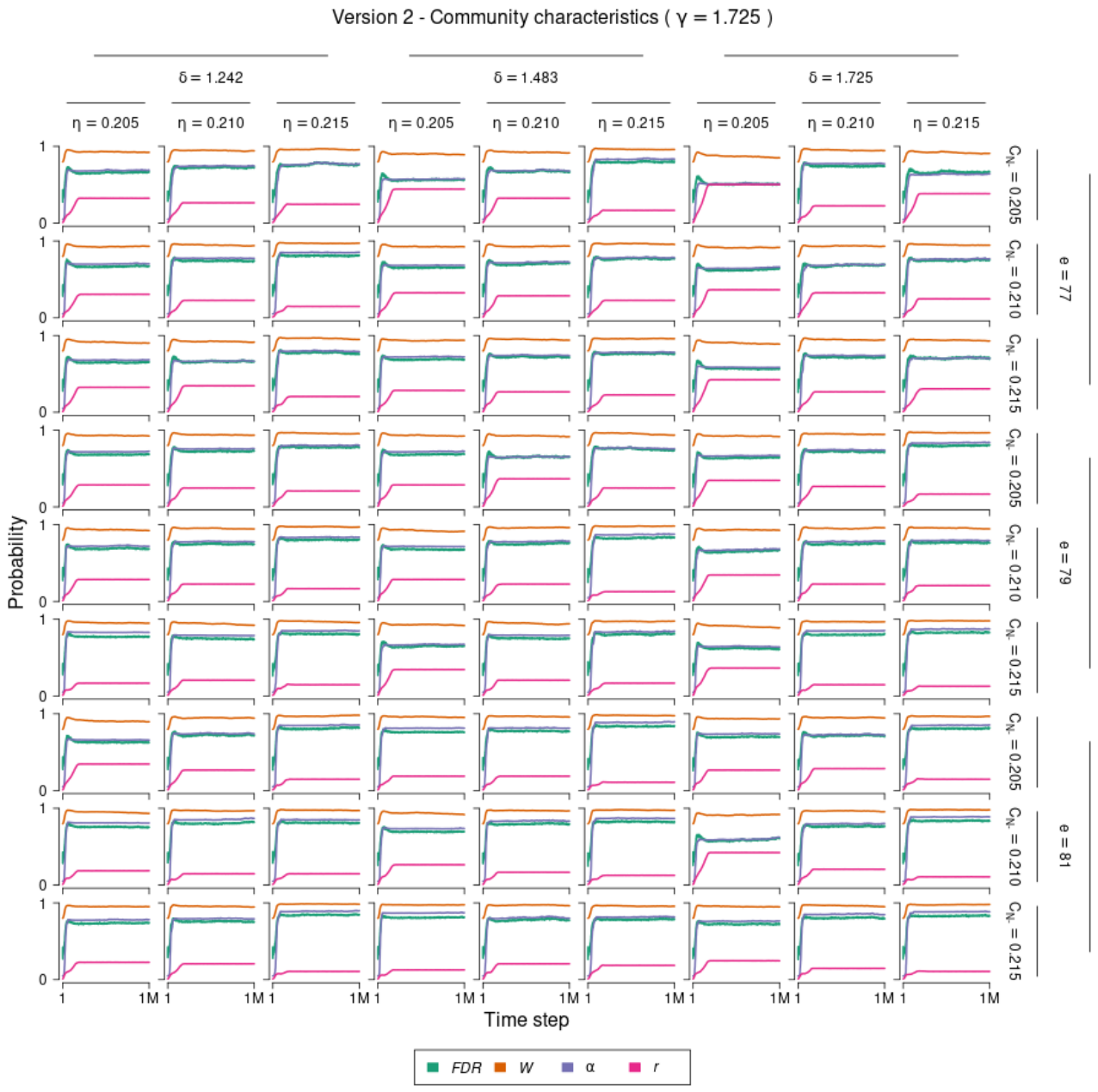

Figure 16 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs over all 1,000,000 time steps of the second set of simulations with $\gamma = 1.725$ and different parameter combinations. Note that panels show an average of all 50 simulation runs. γ and δ are the payoff advantages for advocating and practicing OS, respectively.

B. Lab proportions and characteristics as a function of V_{0-}

For these additional sensitivity analyses, we did not differentiate between simulation runs with two qualitatively different patterns of characteristics (see Results section). Instead, we averaged over all simulation runs. Figure 17 shows the lab proportions for the first set of simulations as a function of V_{0-} and Figure 18 shows the community characteristics for the first set of simulations as a function of V_{0-}. All parameter values, except for V_{0-} are fixed at the values that were used in the main analyses (see Table 2). Figure 17 clearly demonstrates that the lab proportions are robust against specific choice of V_{0-}. Figure 18 shows that the development of FDR, W, and α is also robust against variations of V_{0-}; all of them increase quickly and remain at a high level. However, it can be seen that r increases more strongly the higher the value for V_{0-}.

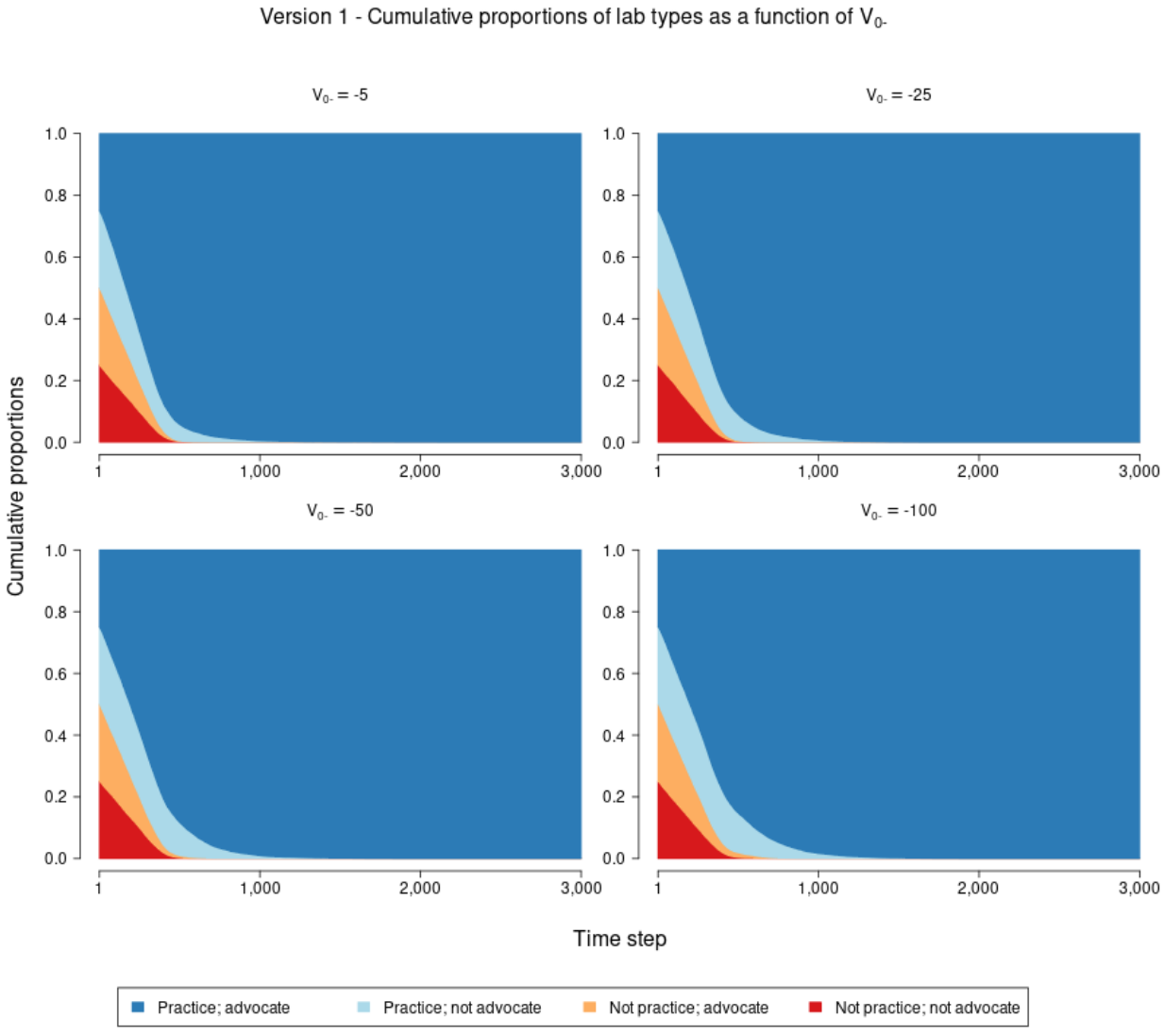

Figure 17 Cumulative proportions of lab types as a function of V_{0-} during the first 3,000 out of 1,000,000 time steps of the first set of simulations. The panels correspond to different values of V_{0-}. All other parameter values are fixed at the values that were used in the main analyses (see Table 2). Note that panels show an average of all 50 simulation runs.

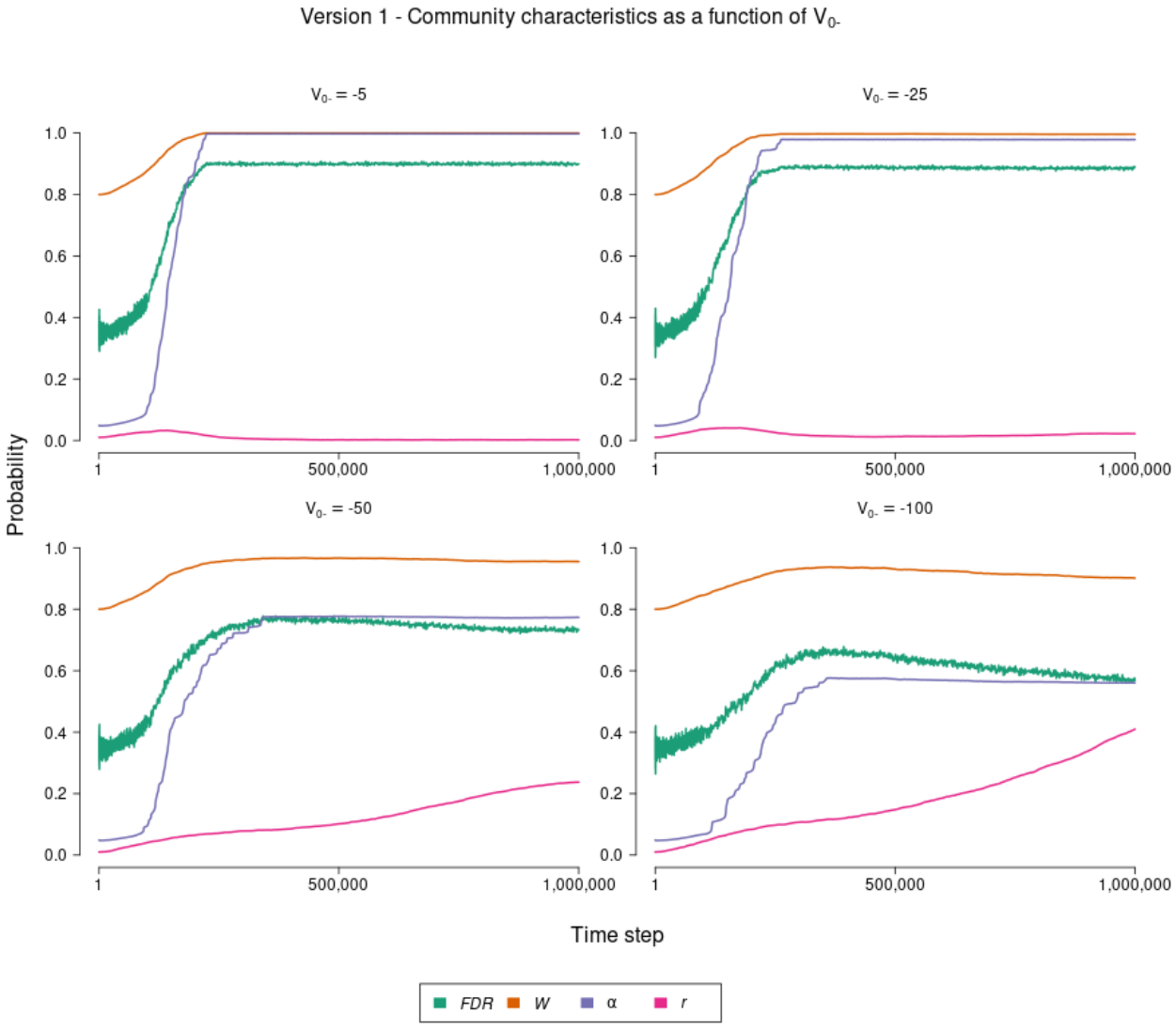

Figure 18 Power W, false discovery rate FDR, false positive rate α, and replication probability r averaged over all labs as a function of V_{0-} over all $1,000,000$ time steps of the first set of simulations. The panels correspond to different values of V_{0-}. All other parameter values are fixed at the values that were used in the main analyses (see Table 2). Note that panels show an average of all 50 simulation runs.