

Aus dem QUEST Center for Responsible Research
des Berlin Institute of Health, Charité – Universitätsmedizin Berlin

DISSERTATION

Open Science Practices and Incentives –
a Meta-Research Investigation

Open Science Praktiken und Anreize –
eine Meta-wissenschaftliche Untersuchung

zur Erlangung des akademischen Grades
Doctor of Philosophy (PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Sophia Crüwell
aus Frankfurt am Main

Erstbetreuer: Prof. Dr. John P. A. Ioannidis

Datum der Promotion: 25.06.2023

Table of contents

List of tables	iii
List of figures	iv
Abstract	1
1 Introduction	3
1.1 The Problem: The Replication Crisis	3
1.2 The (partial) Solution? Open Science Practices and Methodological Reform	4
1.2.1 Evaluating (Open) Data Sharing Practices and Incentives as a Solution	5
1.2.2 Evaluating Replication Research as a Solution	8
1.3 Overarching Research Question	8
2 Methods	11
2.1. Open Data Badge Project	11
2.1.1. Sample	11
2.1.2. Design	11
2.1.3. Procedure	11
2.2. Citation Patterns Project	14
2.2.1. Sample	14
2.2.2. Design	14
2.2.3. Procedure	14
3. Results	15
3.1 Open Data Badge Project Results	15
3.1.1 Overall Results	15
3.1.2 Narrative Summary	17
3.2 Citation Patterns Project Results	18
4. Discussion	22
4.1 Short summary of results	22
4.2 Interpretation of results in context of the literature	22

4.3 Strengths and weaknesses of the studies	24
4.4 Implications for practice and future research	25
5. Conclusion	29
References	30
Statutory Declaration	42
Declaration of own contribution to the publications	43
Excerpt from Journal Summary List	45
Printing copies of the publications	48
Publication 1: Open Data Badge Project	48
Publication 2: Citation Patterns Project	60
Curriculum Vitae	75
Publication list	77
Acknowledgments	79

List of tables

Table 1: Overview table of provided reanalysis code

p. 13

List of figures

Figure 1: Reproducibility ratings based on the initial reproduction attempts	p. 16
Figure 2: Reproducibility ratings based on the group's reproduction attempts	p. 16
Figure 3: Standardised annual citation counts and citation valence	p. 20
Figure 4: Standardised annual citation counts and citation balance/bias	p. 21

Abstract

The replication crisis has uncovered crucial issues in sciences based on statistical inference, such as the biomedical sciences, psychology, and other social or human sciences: key findings seem to be unreplicable, and as research has been conducted largely intransparently, it is unclear how to know what future research can build on. Open Science and reproducibility-related practices have been suggested as a potential solution to this problem by the reform movement as well as metascientific researchers generally. These practices include traditional “open” practices such as openly sharing research results (Open Access), data and/or code (Open Data), and materials (Open Materials); but they also include the pre-registration of studies, and even robust and transparent replication projects.

This dissertation presents two projects aimed at evaluating Open Science practices as working towards a solution to the replication crisis, particularly in psychology. The first project, which was led by the author of this dissertation, considers the effectiveness of the Open Data badge incentive at the journal *Psychological Science* through the computational reproducibility of a full issue of this journal. We found that there is much room for improvement: the majority of articles failed to share the relevant analysis code, and only 4/14 articles were at least essentially reproducible. We make several recommendations to improve the policy of this incentive and thus its influence on the replication crisis. The second study, a collaboration led by Tom Hardwicke, investigates citation patterns after strongly contradictory replication results in four psychology studies, and finds only weak correction effects by the relevant research communities.

Overall, the studies presented in this dissertation may seem to paint a pessimistic picture of possible solutions to the replication crisis, and thus the future of fields plagued by such crises. In the text accompanying the two publications, I try to be cautiously optimistic while pointing out important challenges that need to be addressed by metaresearchers and researchers involved in the reform movement. In particular, I want to highlight the need for further theoretical work refining the underlying concepts.

Zusammenfassung

Die Replikationskrise hat in den auf statistischen Schlussfolgerungen basierenden Wissenschaften wie der Biomedizin, der Psychologie und anderen Sozial- und Humanwissenschaften entscheidende Probleme aufgedeckt: wichtige Ergebnisse scheinen nicht reproduzierbar zu sein, und da die Forschung weitgehend intransparent durchgeführt wurde, ist unklar, worauf zukünftige Forschung aufbauen kann. Sowohl die Reformbewegung als auch metawissenschaftliche ForscherInnen haben Open Science Praktiken und andere reproduzierbarkeitsbezogene Methoden als mögliche Lösung für dieses Problem vorgeschlagen. Zu diesen Praktiken gehören traditionelle "offene" Praktiken wie der offene Austausch von Forschungsergebnissen (Open Access), Daten und/oder Code (Open Data) und Materialien (Open Materials), aber auch die Präregistrierung von Studien sowie robuste Replikationsprojekte.

In dieser Dissertation werden zwei Projekte vorgestellt, die darauf abzielen, Open Science Praktiken als Beitrag zur Lösung der Replikationskrise zu evaluieren, insbesondere in der Psychologie. Das erste Projekt, das von der Autorin dieser Dissertation geleitet wurde, untersucht die Anreizmethode des Open Data Badge der Zeitschrift *Psychological Science* anhand der rechnerischen Reproduzierbarkeit einer vollständigen Ausgabe dieser Zeitschrift. Wir fanden heraus, dass es ein großes Verbesserungspotential gibt: Die Mehrheit der untersuchten Artikel hatte den relevanten Analysecode nicht geteilt, und nur 4/14 Artikel waren zumindest im Wesentlichen reproduzierbar. Die zweite Studie, eine Zusammenarbeit unter der Leitung von Tom Hardwicke, befasst sich mit den Zitationsmustern nach stark widersprüchlichen Replikationsergebnissen in vier Psychologiestudien und fand nur schwache Korrektoreffekte durch die betreffenden Forschungsgemeinschaften.

Insgesamt scheinen die in dieser Dissertation vorgestellten Studien ein pessimistisches Bild von möglichen Lösungen für die Replikationskrise und damit für die Zukunft der von solchen Krisen geplagten Fachgebiete zu zeichnen. In diesem Manteltext versuche ich, vorsichtig optimistisch zu sein und gleichzeitig auf wichtige Herausforderungen hinzuweisen, die von MetawissenschaftlerInnen und ForscherInnen, die an der Reformbewegung beteiligt sind, angegangen werden müssen. Insbesondere möchte ich auf die Notwendigkeit weiterer theoretischer Arbeiten zum besseren Verständnis der zugrunde liegenden Konzepte hinweisen.

1 Introduction

“If we as experimental psychologists are missing an opportunity to make significant contributions to natural science—if we are failing to assume leadership in an area of behavior investigation where we might be useful and effective—if these things are true, and I believe that they are, then we have no one but ourselves to blame.”
(Beach, 1950, p. 124)

What Frank Beach wrote in 1950 about experimental psychology, in particular comparative psychology, still rings true today. It rings true not just about psychological research, but also about metaresearch *evaluating* psychological and related research. In this text accompanying one first-author “top-journal” publication and a further co-authored publication, I will present research evaluating Open Science practices potential answers to the replication crisis, particularly in psychology. In the introduction, I will set the scene of the replication crisis, Open Science, and the reform movement, as well as empirical metascientists and their evaluations of existing reforms. I will discuss research on this topic and highlight gaps that may be filled, and how the two studies presented here help make these gaps somewhat smaller. In the second section, I will describe the methods used in each study. In the third section, I will briefly summarise the results of both studies. In the fourth section, I will discuss these results, also in the broader context of metascience as a currently evolving field. I will also argue that there is a lack of crucial conceptual work that needs to be filled by theoretical metascientists or philosophers of science.

1.1 The Problem: The Replication Crisis

1.1.1 History

The replication crisis is a relatively recent phenomenon in psychology, biomedicine, and the social sciences where large scale replication attempts of often key original findings fail to produce similar results. In this thesis, I am focusing on psychology in particular, where the problems surrounding the replication crisis were more fully understood from around 2011 onwards (Pashler & Wagenmakers, 2012). The proposed reasons for this crisis are varied, and include exploiting researcher degrees of freedom or the use of questionable research practices (QRPs; John et al., 2012; Simmons, Nelson,

& Simonsohn, 2011), publication bias (Nelson, 2020; Renkewitz & Keiner, 2019), and perverse incentives (Higginson & Munafò, 2016). In particular in the last decade, there has been a trend of empirical metaresearch trying to better understand and solve the problems raised by this replication crisis and its potential underlying problems, as well as evaluating proposed solutions (Hardwicke et al., 2020). Metaresearchers are both researchers focusing exclusively on metaresearch as well as researchers involved in the reform movement but working in so-called “substantive” research areas; in fact, it has been argued that not all who could be seen as doing metaresearch are aware that this is the case (Ioannidis et al., 2015).

It is important to differentiate between two closely related concepts: ‘replication’ or ‘replicability’ on the one hand and ‘reproduction’ or ‘reproducibility’ on the other (see, e.g., Nosek et al., 2022). These terms are often falsely used interchangeably, and are used by different fields in sometimes opposing meanings. In research on, or related to, psychology, a ‘replication’ usually means the repetition of an original study such that a researcher uses the original methods to collect new data, and attempts to use these new data in combination with the original analysis methods (ideally using the original analysis code or scripts) to try to get similar results. By contrast, ‘reproduction’ in this context means the repetition of an original study using that study’s original data and analysis methods or code. In this synopsis, I will use these definitions of these key terms.

1.2 The (partial) Solution? Open Science Practices and Methodological Reform

Open Science is often seen as a possible solution to the replication crisis, or at least to parts of the replication crisis. What is Open Science? Open Science is a term that is loosely and variously defined (Crüwell et al., 2019). I understand Open Science to mean science that is transparent with a view to enabling rigorous and, where applicable, reproducible research. This stance may be taken as far as viewing Open Science practices as one indicator for science that is more objective, as they can help guard against the effects of biases (van Dongen & Sikorski, 2021). It is also important to distinguish Open Science and the Open Science movement from the more specific Open Access movement. Open Science includes Open Access, but it is a broader concept that can also include specific practices such as preregistration, Registered Reports, and even replication research. More broadly, advocating for this collection of practices is often also

called “methodological reform”, meaning practices aimed at improving transparency and methodological rigour (e.g., preregistration, Registered Reports, replication research, and other open science practices).

“Open Data” describes the *openly accessible* sharing of data. It is important to distinguish this from mere data sharing, which does not necessarily happen in an openly accessible venue or format. The goal of Open Data is chiefly to share data in a publicly accessible way, but often also to enable reproducibility. Encouraging data sharing can thus mean anything from mandating any data availability statement (including “data are not available” or “data available on request”), to encouraging openly accessible data sharing through the use of Open Data badges, to mandating data sharing or openly accessible data sharing. In some cases, Open Data includes Open Code, as the relevant analysis code is as necessary as the data to assess reproducibility of the reported results.

“Open Methods”, similarly to Open Data, means openly sharing the methods used in a study. This may include stimuli, but sometimes also analysis code if this is not included in Open Data (Klein et al., 2018). Preregistration is the time-stamped registration before data collection of a study, its hypotheses and the planned analyses (Nosek et al., 2018). In the Registered Reports publishing format, studies are evaluated based on the planned methods rather than the results; the publication of the study thus does not depend on its results (Chambers et al., 2014; Hardwicke & Ioannidis, 2018).

Given the pressing issue of the replication crisis and its effects, it is important to evaluate whether, and which, Open Science practices and incentives are helping to solve the underlying issues or the effects of this crisis. This is where empirical metaresearch comes in: there are already a number of metaresearch studies concerned with evaluating Open Science practices and incentives. In the following sub-subsections, I will go into the state of current metaresearch on evaluations of Open Science practices and incentives, focussing on Open Data and on replication research. In particular, I will point out crucial knowledge gaps that the studies presented in this thesis attempt to fill. I will begin by examining Open Data sharing practices and incentives as a solution to the replication crisis. I will then consider replication research as a further possible solution.

1.2.1 Evaluating (Open) Data Sharing Practices and Incentives as a Solution

As explained above, the aims of data sharing initiatives and incentives include increasing reusably shared data, as well as enabling reproduction and reproducibility of the reported results. Do existing initiatives and incentives lead to the desired results, that

is, to more (reusably) shared data and reproducibility of the reported results? A low-effort initiative for increasing data sharing is to encourage or mandate the use of a data availability statement. Gabelica et al. (2022) assessed the use of such a statement in articles published in Open Access journals run by the publisher BioMed Central. They categorised the data availability statements and found that the most common statement (42% of cases) was that data were available on reasonable request. When they requested this data, they received a response for 14% and data for only 6.8% of these articles (Gabelica et al., 2022). Overall, it seems clear that the mandatory data availability statements were not adhered to.

Going one step beyond mandating a data availability statement, another possibility is to mandate the post-publication sharing of data when requested by a reader of the relevant article. Such a policy, at the journal *Science* (in this case including the sharing of analysis code as well as data), was investigated by Stodden et al. (2018). They requested the data and analysis code from authors who published in *Science* after the introduction of the sharing policy, and found that they only received data from 44% of the articles under investigation. In a second step, they tried to use these data and code to reproduce the reported results, and were only able to successfully do so for 26% of the articles (Stodden et al., 2018). So, even a clear data sharing mandate from a journal does not have to make shared data more likely, including data only shared on request. There is, however, evidence that data sharing mandates can be more successful. Hardwicke et al. (2018) found mixed results on the mandatory data sharing policy at the journal *Cognition*: while there was an indication of an increase in both data availability statements and data reusability, the values of only around 31% of the articles they investigated were fully reproducible, which rose to about 63% with input from original authors. Nuijten et al. (2017) similarly found that open data sharing mandates were successful in increasing shared data, though in around a third of the cases they examined in one of their studies, there was no open data despite a statement affirming its existence.

Alternatively, rather than mandating data sharing, data availability statements, or Open Data, journals or funders can incentivise and/or reward these practices. One example for such an incentive are the Open Science badges introduced by the Center for Open Science (Blohowiak et al., 2022). The introduction of these badges and similar incentives has not always been successful. Specifically, it was found that the introduction of badges was associated with only a small increase in openly shared data at *Biostatistics* (7.6%; Rowhani-Farid & Barnett, 2018) and no such increase in *BMJ Open*, compared to

a control condition (Rowhani-Farid et al., 2020). Relevant studies in the field of Psychology in particular include Hardwicke et al. (2021b), Obels et al. (2020) and Kidwell et al. (2016). Kidwell et al. (2016) found that the introduction of Open Science badges at *Psychological Science* was associated with an increase in articles reporting that they openly shared their data, both compared to data sharing at *Psychological Science* prior to this policy and compared to other data sharing at other journals. Focusing not on studies that received a badge but on psychological research published in the Registered Reports format, Obels et al. (2020) found that 58% of articles they investigated (36 out of 62) provided both data and code. The results of 21 of these 36 articles (58%) were *computationally* reproducible by their team. Hardwicke et al. (2021b) assessed whether the 25 articles published in *Psychological Science* between 2014 and 2015 which were awarded an Open Data badge were *analytically* reproducible. They found that the main results of 9 (36%) of these articles were reproducible without involving the original authors, and a further 6 (24%) of these articles were reproducible after involving the original authors.

A crucial relevant knowledge gap in this area is that Hardwicke et al. (2021b) only reproduced a subset of the reported results. A study reproducing all reported results would be more informative regarding the exact reproducibility of results. Furthermore, the sample used in Hardwicke et al. (2021b) was published between 2014 and 2015, which was immediately after the initial introduction of the badge policy at *Psychological Science*. A more recent sample may give a better account of current data sharing practices, particularly connected to the Open Data badge policy at *Psychological Science*. Finally, these studies tend to be reproduction reports focused on numerical results, but the addition of in-depth narrative or qualitative reproduction accounts may give further insights into the issue, including into how to most usefully share data and analysis code for future reproducibility. It is important to note that the criteria for the award of an Open Data badge at *Psychological Science* include not only openly shared data, but also openly sharing the analysis code necessary for independent reproduction (Psychological Science, 2022); this has been the case since at least November 2017.¹

¹https://web.archive.org/web/20171115110444/https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#OPEN

1.2.2 Evaluating Replication Research as a Solution

Is replication research a possible solution to the replication crisis? Given that replication research is seen to be very important to science and scientific discovery (Zwaan et al., 2018), and that the label of this crisis focuses on replication, this seems plausible. However, it is unclear whether replications have enough impact in the literature and the field to be a possible solution to the crisis. Serra-Garcia and Gneezy (2021) found that nonreplicable research in economics, in psychology, and published in *Nature* or *Science* is cited more often than replicable research in the same fields or publications. Similarly, McDiarmid et al. (2021) found that while researchers and graduate students in psychology update their beliefs about an effect after an unsuccessful replication attempt, they do not do so as much as they should if they fully took into account the information provided by the replication study. These two studies are important, but do not give a full account of the impact of replication studies on the literature and more specifically the original studies or effects. Serra-Garcia and Gneezy (2021) focus mostly on the citation count, and do not fully consider the valence (i.e., whether they were positive or negative) of these citations. McDiarmid et al. (2021) investigate the beliefs of psychology researchers generally and in an artificial setting rather than the actual citation behaviour of scientists working in the relevant fields.

1.3 Overarching Research Question

It is important to note that the projects presented here have a strongly descriptive and exploratory focus. An overarching research question of the overall thesis may be summed up as: what are the implications of open and reproducible research practices and incentives? The individual projects presented in this thesis are concerned with the Open Data badge incentive, and with the implications of replication research for citation patterns. Specifically, the hypotheses or investigative foci were: 1) How effective is the Open Data badge policy at *Psychological Science* in adhering to its aim of ensuring computational or results reproducibility at this journal? 2) What is the implication of strongly contradictory replication results on the citation patterns of the original study?

These questions are addressed in two published articles. The first published article makes up my cumulative thesis project, as it is a first-author article accepted at an internationally leading peer-reviewed journal. The second article is a middle author article,

also published in a top journal. These are the two articles I am concerned with in this synopsis:

- 1) **Sophia Crüwell**, Deborah Apthorp, Bradley James Baker, Lincoln Colling, Malte Elson, Sandra J Geiger, Sebastian Lobentanzer, Jean Monéger, Alex Patterson, D Sam Schwarzkopf, Mirela Zaneva, Nicholas JL Brown (2023). "What's in a badge? A computational reproducibility investigation of the Open Data badge policy in one issue of *Psychological Science*". *Psychological Science*, 34(4), 512-522. <https://doi.org/10.1177/09567976221140828>
- 2) Tom E Hardwicke, Dénes Szűcs, Robert T Thibault, **Sophia Crüwell**, Olmo R van den Akker, Michèle B Nuijten, John PA Ioannidis (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. *Advances in Methods and Practices in Psychological Science*, 4(3), 25152459211040837. <https://doi.org/10.1177/25152459211040837>

Further to the articles presented in this thesis, I have led and been involved in a number of projects related to the topic of this thesis. These are at various stages of completion, ranging from published papers to articles in preparation for journal submission (ordered by publication year):

- 1) **Crüwell, S.**, van Doorn, J., Etz, A., Makel, M.C., Moshontz, H., Niebaum, J., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science. *Zeitschrift für Psychologie*. <https://doi.org/10.1027/2151-2604/a000387>
- 2) **Crüwell, S.**, Stefan, A.M., Evans, N.J. (2019). Robust Standards in Cognitive Science. *Computational Brain & Behaviour*. <https://doi.org/10.1007/s42113-019-00049-8>
- 3) Hardwicke, T.E., Serghiou, S., Janiaud, P., Danchev, V., **Crüwell, S.**, Goodman, S., Ioannidis, J.P.A.(2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, 7, 11-37. <https://doi.org/10.1146/annurev-statistics-031219-041104>
- 4) Hardwicke, T. E., Wallach, J. D., Kidwell, M., 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>

- 5) **Crüwell, S.**, Evans, N.J. (2021) Preregistration in Diverse Contexts: A Preregistration Template for the Application of Cognitive Models. *Royal Society Open Science*. <https://doi.org/10.1098/rsos.210155>
- 6) Kent, B. A., Holman, C., Amoako, E., Antonietti, A., Azam, J. M., Ballhausen, H., Bediako, Y., Belasen, A.M., Carneiro, C. F. D., Chen, Y., Compeer, E. B., Connor, C. A. C., **Crüwell, S.**, Debat, H., Dorris, E., Ebrahimi, H., Erlich, J. C., Fernández-Chiappe, F., Fischer, F., Gazda, M. A., Glatz, T., Grabitz, P., Heise, V., Kent, D. G., Lo, H., McDowell, G., Mehta, D., Neumann, W., Neves, K., Patterson, M., Penfold, N. C., Piper, S. K., Puebla, I., Quashie, P. K., Quezada, C. P., Riley, J. L., Rohmann, J. L, Saladi, S., Schwessinger, B., Siegerink, B., Stehlik, P., Tzilivaki, A., Umbers, K. D. L., Varma, A., Walavalkar, K., de Winde, C. M., Zaza, C., & Weissgerber, T. L. (2022). Recommendations for empowering early career researchers to improve research culture and practice. *PLoS biology*, 20(7), e3001680. <https://doi.org/10.1371/journal.pbio.3001680>
- 7) van Ravenzwaaij, D., Bakker, M., Heesen, R., Romero, F., van Dongen, N., **Crüwell, S.**, Field, S., Hartgerink, C.H.J., Held, L., Pittelkow, M.M., Tiokhin, L., Traag, V., van den Akker, O., van't Veer, A., & Wagenmakers, E. J. (2022). Perspectives on Scientific Error. [PREPRINT] <https://psyarxiv.com/wm4v6/>
- 8) **Crüwell, S.**, Hardwicke, T.E., Alsati, T., Ioannidis, J.P.A. (*in preparation for submission*). Brain Drain and Brain Deficit Estimates for Highly Cited Scientists in Their Various Career Stages.
- 9) **Crüwell, S.**, Hardwicke, T.E., Ioannidis, J.P.A. (*in preparation for stage 1 RR submission*). An empirical assessment of methodological issues in survey research on Questionable Research Practices: A replication and extension of John et al. (2012).

2 Methods

In this section, I will describe the methodology of the published projects presented in this dissertation (see also: Crüwell et al., 2023; Hardwicke et al., 2021a). This description includes a report of the sample, study design, and procedure of each study. The information presented here should enable others to understand what was studied, and further description is available in the corresponding published papers and the associated supplementary materials and OSF repositories.

2.1. Open Data Badge Project

2.1.1. Sample

We examined all 14 research articles of the April 2019 issue of *Psychological Science* (Bae & Luck, 2019; Dorfman et al., 2019; Garcia & Rimé, 2019; Geniole et al., 2019; Hakim et al., 2019; Hilgard et al., 2019; Johnson & Wilson, 2019; Lindsay et al., 2019; Obaidi et al., 2019; Olsson-Collentine et al., 2019; Vardy & Atkinson, 2019; Wójcik et al., 2019; Woolley & Fishbach, 2019; Yousif & Keil, 2019).

2.1.2. Design

The study design can best be described as an “observational, descriptive one-group study” (Crüwell et al., 2023, p. 8).

2.1.3. Procedure

After dealing with initial issues with the assignment of reproducers to the 14 articles (for more information, see Crüwell et al., 2023), 12 reproducers were assigned to reproduce the published results of three to five of the articles under examination. At least three researchers per article reported on a reproduction attempt on the results published in said article. We produced 46 individual reproduction reports in total (Crüwell et al., 2023).

The process of reproducing the results reported in the target articles was divided into two phases (Crüwell et al., 2023). The first phase consisted of individual reproduction attempts by each researcher of the results of the articles assigned to them. This was done entirely independently of the other reproducers on the team and entirely independently of the authors of the original articles (NJLB had contacted the authors of two articles for help

before the start of this project, as a result of which he was not involved in reproducing or discussing these articles). In the second phase, the individual reports for each article were combined into a summary report per article (Crüwell et al., 2023). The individual and summary reports are narrative and in-depth accounts of the reproduction process. The individual reports are also accompanied by analysis code or scripts in every case but one (BJB's reproduction of article 9), unless a) the original article was accompanied by analysis code which was adapted as described in the reports, or b) reproduction was not possible (see Table 1). To further quantify reproduction success, each researcher rated the reproducibility of the results of each article they attempted to reproduce. Reproducibility was rated as either "Not at all" (i.e., no consistency between reported results and reproduction), "Mostly not" (i.e., major deviations and few numerically consistent results), "Partially" (i.e., more than minor deviations, but mostly numerically consistent results), "Essentially" (i.e., minor deviations in the decimals, or obvious typographical errors), and "Exactly" (i.e., no deviations from the reported results; Crüwell et al., 2023). Each reproducer rated the reproducibility of their assigned articles based on their individual reproduction attempts and based on all reproduction attempts for the relevant articles.

Table 1: Overview Table of reanalysis code provided by the reproducers or original author code used (own representation: Sophia Crüwell).

	Reproduction 1	Initials	Reproduction 2	Initials	Reproduction 3	Initials	Reproduction 4	Initials	Reproduction 5	Initials
Article 1	Code provided	DA	Code provided	LJC	No reproduction possible	DSS				
Article 2	Code provided	DA	Code provided	SC	Code provided	DSS				
Article 3	No reproduction possible	AP	No reproduction possible	ME	No reproduction possible	NJLB				
Article 4	Code provided	LJC	Original author code	MZ	Original author code	ME	Code provided	NJLB		
Article 5	Code provided	DA	Code provided	MZ	Code provided	DSS				
Article 6	Code provided	LJC	Original author code	JM	Original author code	SL				
Article 7	Original author code	BJB	Original author code	AP	Original author code	JM	Original author code / Code in report	ME	Original author code	SL
Article 8	Original author code	MZ	Original author code	ME	Original author code	NJLB				
Article 9	Code provided	LJC	Code missing	BJB	Code provided	SJG				
Article 10	Original author code	SL	Original author code	JM	Original author code	AP				
Article 11	Code provided	MZ	Code provided in pdf	AP	Code provided	SJG				
Article 12	Code provided	DA	Code provided	NJLB	Code provided	SC				
Article 13	Code provided / Original author code / Code in report	SC	Original author code	BJB	Code provided / Original author code / Code in report	SJG				
Article 14	Code provided	DA	Code provided	LJC	Code provided	JM				

2.2. Citation Patterns Project

These methods are described and published in Hardwicke et al. (2021a).

2.2.1. Sample

We examined four cases of original studies that failed to replicate in preregistered, multisite replication studies (Hardwicke et al., 2021a). Specifically, these studies were Baumeister et al. (1998) and Sripada et al. (2014)², concerned with ego depletion and not replicated by Hagger et al. (2016), Strack et al. (1988), concerned with the facial feedback hypothesis and not replicated by Wagenmakers et al. (2016), Caruso et al. (2013), concerned with money priming and not replicated by Klein et al. (2014), and Carter et al. (2011), concerned with flag priming and not replicated by Klein et al. (2014).

2.2.2. Design

The study design can best be described as a “retrospective observational study” (Hardwicke et al., 2021a, p. 4).

2.2.3. Procedure

The study was preregistered on the OSF (<https://osf.io/eh5qd>). We extracted annual citation counts for the original studies as well as a reference class of articles published in the same journal and year as the original study. Following this, we qualitatively coded citation valence for these citations as favourable, equivocal, unfavourable, or unclassifiable, in order to investigate whether the citation patterns showed that existing beliefs were corrected in light of the failed replication or perpetuated in spite of the failed replication. We further coded co-occurring citations of the original and replication study (to investigate citation balance or bias), as well as the frequency and type of counterarguments to investigate whether these articles included an explicit defense of the original study or whether this was absent. Specific step by step information for this process is available in Hardwicke et al. (2021a) as well as on the OSF (<https://osf.io/w8h2q/>).

² These studies were both concerned with the “ego-depletion effect”. Although Baumeister et al. (1998) is the original study of ego depletion, the relevant replication study (Hagger et al., 2016) replicated the computer-based version introduced by Sripada et al. (2014) instead. We therefore investigated the citation patterns of both Baumeister et al. (1998) and Sripada et al. (2014) for the purposes of our Citation Patterns project (Hardwicke et al., 2021a).

3. Results

In this section, I will summarise and present the important new results of the research carried out as part of this thesis project. I will go through the articles in turn, starting with the project on the Open Data badge.

3.1 Open Data Badge Project Results

3.1.1 Overall Results

The main results of this project are that the Open Data badge policy at the journal *Psychological Science* was not effective in bringing about the desired change, at least for the April 2019 issue under investigation. In total, we produced 46 individual reproduction reports and 14 summary reproduction reports – at least three individual reports and one summary report for each article. We found that there was at least some data available for all fourteen articles examined, but only six articles were accompanied by the relevant analysis code or scripts. Following our more in-depth investigation into the reproducibility of the reported results, we only rated one out of the fourteen articles to be exactly reproducible. We rated a further three articles as essentially reproducible with minor deviations (Crüwell et al., 2023).

More specifically, the reproducibility ratings can be seen in Figure 1 and Figure 2. Figure 1 shows each rater's ratings for each article they attempted to reproduce based on their single, individual reproduction attempt. Figure 2 shows each rater's ratings for each article they attempted to reproduce based on all reproduction attempts of this article and the resultant summary report. As you can see, the median rating of the article changes from individual to group reports in four cases, namely articles 114, 110, 101, and 109. In the latter three articles, the change is in favour of the article's reproducibility, moving from "mostly not" to "partially" and from "partially" to "essentially". The rating for article 114, however, changes from "essentially" to "partially" – though please note that the group ratings for article 114 are split exactly between essentially and partially reproducible.

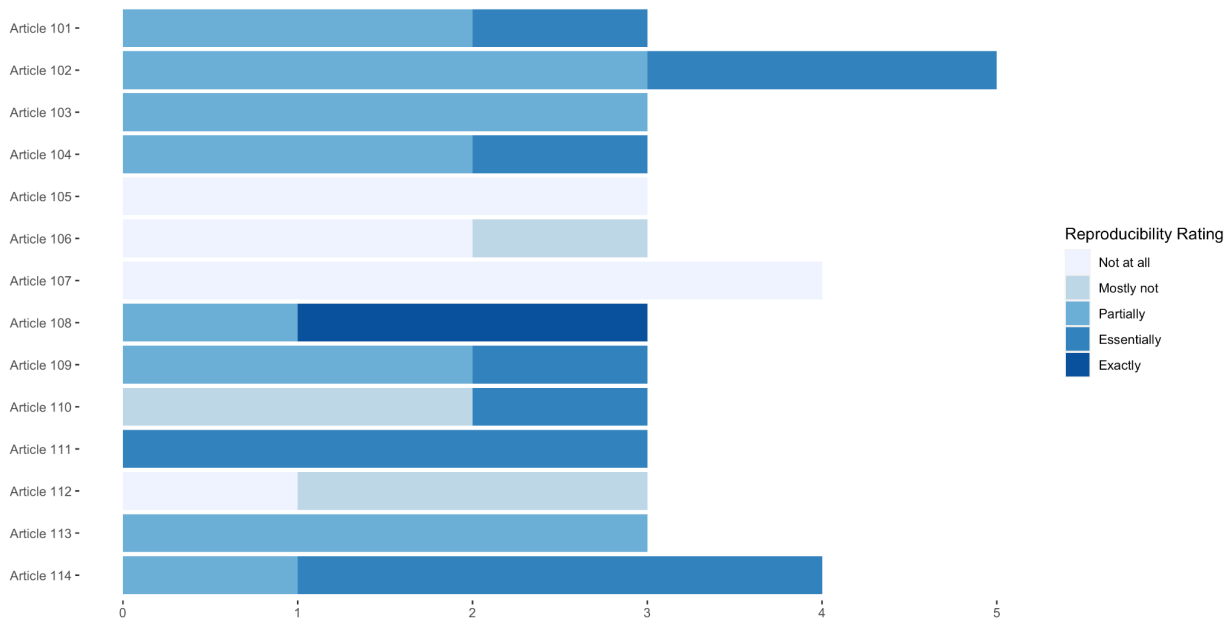


Figure 1: Reproducibility ratings (Not at all, Mostly not, Partially, Essentially, Exactly) by the relevant reproducers for each article after the initial reproduction attempts. The x-axis is the number of reviewers. (own representation, modified from own representation as shared in the initial preprint version of Crüwell et al., 2023: Sophia Crüwell)

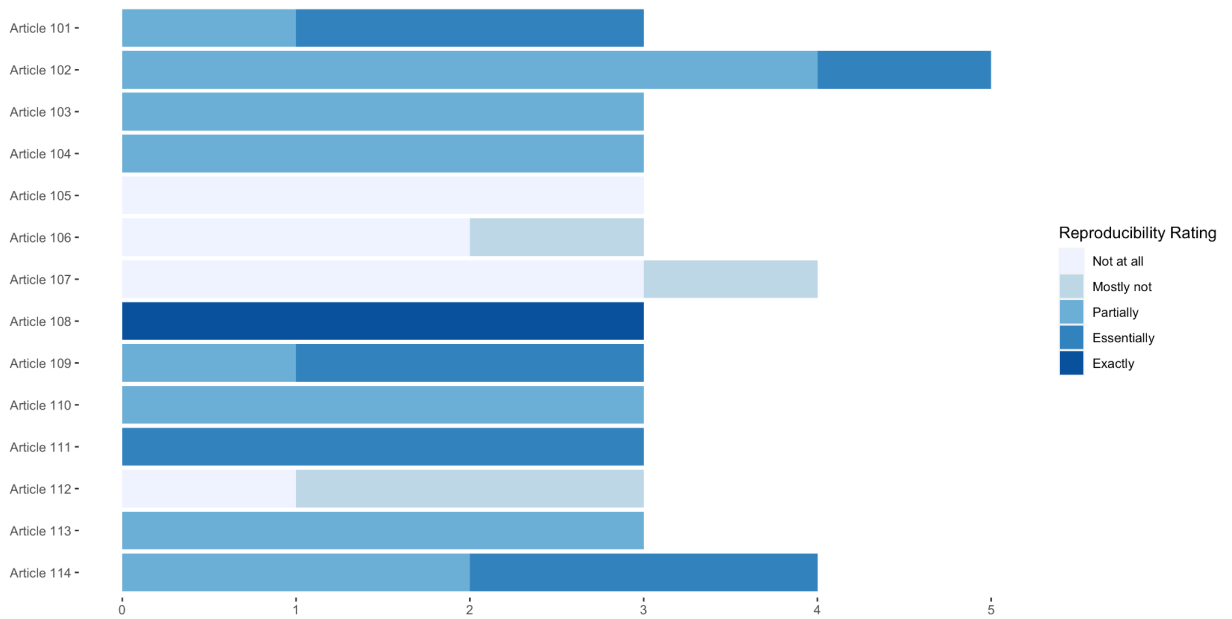


Figure 2: Reproducibility ratings (Not at all, Mostly not, Partially, Essentially, Exactly) by the relevant reproducers for each article based on the group discussions and the summary reports. The x-axis is the number of reviewers. (own representation, modified from own representation as shared in the initial preprint version of Crüwell et al., 2023: Sophia Crüwell)

3.1.2 Narrative Summary

The following narrative report summarises the key results of this study, namely the in-depth and rich descriptions of 46 individual reproduction attempts and 14 group reproduction summaries across the 14 articles under investigation. A similar narrative summary was initially shared online as part of the initial *preprint* (Version 1) for Crüwell et al. (2023), but it is not part of the final published article and is not published elsewhere. Both the version of this narrative summary initially shared as part of the preprint and the current restructured version of this narrative summary were written in full by the author of this thesis, Sophia Crüwell.

Only the results of Article 108 were exactly reproducible. We commend the authors of this article for their efforts in making their data and code openly available.

Articles 101, 109, and 11 were essentially reproducible by our team. The results of Article 101 and Article 111 were essentially reproducible even without the provision of analysis code. This was only possible, with some minor deviations, because the analyses were both simple and sufficiently described. In the case of Article 101, however, only summarised data was shared. Article 109 was essentially reproducible with minor deviations thanks to the analysis syntax provided as well as the descriptions of the analyses in the article.

The results reported in Articles 102, 103, 104, 110, 113, and 114 were partially reproducible overall. The data and code shared alongside Article 102 were sufficient to reproduce most of the results with several major deviations but no change in direction of the results. In the case of Article 103, Table 1 was reproducible, while Table 2 was not due to issues with the code and lack of a cleaned version of the data. There were problems with the analysis code provided alongside Article 104, which contributed to reproduction difficulties. Article 110 did not share any statistical analysis code, and some of the shared data could not be downloaded by the reproducers. The reproduction experiences differed: one reproducer achieved more than partial reproduction with considerable effort and due to fortuitous guesswork, while the other two reproducers partially reproduced the results. Similarly, the results of Article 113 could not be successfully reproduced without analysis code, particularly as the description in the article of the logistic regression variables was unclear. Article 114 provided code and most of the data, and most of the results in the main article were reproducible with some deviations. The results in the Supplementary Materials were not all reproducible.

Article 112 was the only article that was rated to be mostly not reproducible. There were issues with the provided modelling code, and the statistical analysis code was missing. Nevertheless, a few t-values, p-values, and degrees of freedom were reproducible.

It was not at all possible to reproduce Articles 105, 106, and 107. In these cases, no analysis code was shared and there were issues with the shared data. In the case of Article 105, the analyses were too complex to reproduce without the original code, and a lack of raw data meant that important results are not in principle reproducible. Article 106 provided only raw data, without code or instructions for processing, which made reproduction without the original code infeasible. Similarly, Article 107 provided no analysis code and shared raw data without instructions or code for further processing. One reproducer was nevertheless able to reproduce some similar but not identical results after three days of work.

3.2 Citation Patterns Project Results

We found 2,829 articles citing any of the original studies under examination at any point. Of these articles³, 632 articles were published in the relevant preregistered time period for qualitative assessment: 1 year prior to publication of the replication until 31st December 2019. Further, 28 articles were excluded from the qualitative analysis due to lack of access (n=22), and lack of relevant citation in the main text (n=5) or at all (n=1), resulting in a sample of 604 articles (Hardwicke et al., 2021a).

The key result of this study is that we were able to find only weak correction effects in the citation patterns of the investigated studies after a strongly contradictory replication result (Figure 3; Hardwicke et al., 2021a). After the publication of the contradictory replication study, in the case of the Baumeister study, we found a small initial decline in citations to the original study after the publication of the replication study, followed by a small increase two years later, increasing from 191 to 199 in the relevant time period (2015-2019; Hardwicke et al., 2021a). The citations to the Strack study declined somewhat from 56 to 41 citations in the years 2015 to 2019. There was no considerable overall change in citations for Carter, Caruso, and Sripada, although there was an initial citation increase for Carter and Caruso and an initial citation decrease for Sripada. Regarding citation valence, most citations of the original study were favourable pre-

³ We took a random sample of 40% of articles citing the Baumeister et al. study.

replication. This changed somewhat after the publication of the strongly contradictory replication result, as we found somewhat fewer favourable citations and somewhat more unfavourable citations of the original study in the cases of Carter, Caruso, Sripada, and Strack (no change for Baumeister). There were few unfavourable citations overall, however, and the general valence of the citations of the original studies remained positive and favourable across all studies. Taken altogether, this suggests that we found a largely *unchallenged belief perpetuation* across the examined studies (Hardwicke et al., 2021a).

We also found indications of *citation bias*, i.e. studies citing the original study after the replication study was published without citing the replication study, for all articles except for Sripada, the citing articles of which showed *citation balance* (see Figure 4). For Baumeister, Carter, Caruso, and Strack, fewer than half of all articles citing the original study also cited the replication study for the time points we examined. In the case of Sripada, more than 88% of citing studies also cited the replication study (Hardwicke et al., 2021a).

Furthermore, we found that of the 127 articles (across all studies) citing both the original and the replication studies, 51 articles explicitly defended the original study or effect by providing a counterargument to the findings of the relevant replication study. 60 of the studies citing both original and replication study cited the original study in a favourable way, and of those studies, 31 provided a counterargument (Hardwicke et al., 2021a). Overall, this indicates neither a clearly *explicit defense* pattern nor a clearly *absent defense* pattern.

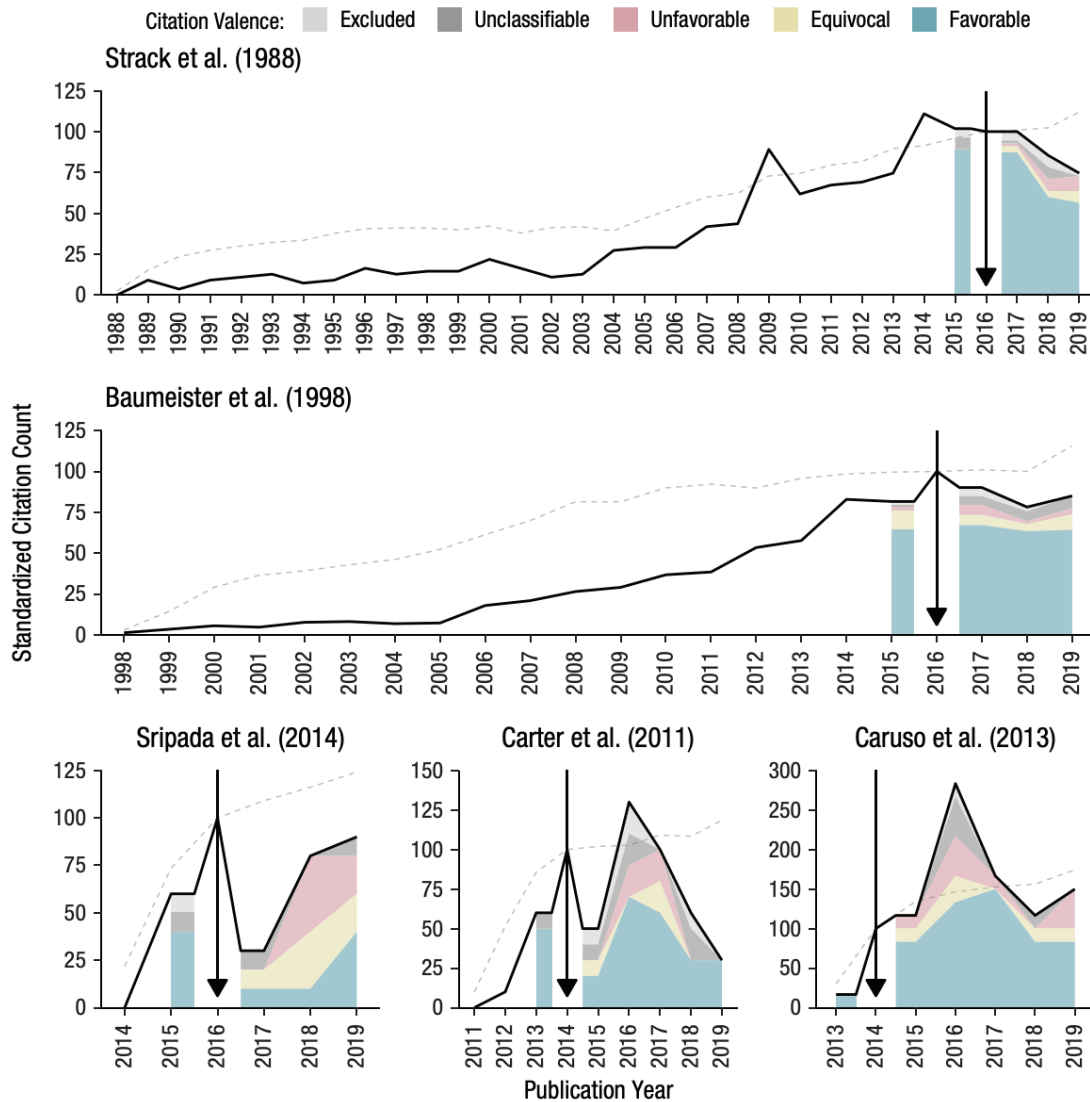


Figure 3: “Standardized annual citation counts (solid line) for the five original studies with citation valence (favorable, equivocal, unfavorable, unclassifiable) illustrated by colored areas in prereplication and postreplication assessment periods. The dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Citation valence classifications for the Baumeister case are extrapolated to all articles in the assessment period according to a 40% random sample.” (Hardwicke et al., 2021a, p. 6). This figure is reproduced from Hardwicke et al. (2021a; Figure 1), which was shared under the CC BY-NC 4.0 license (<https://creativecommons.org/licenses/by-nc/4.0/>).

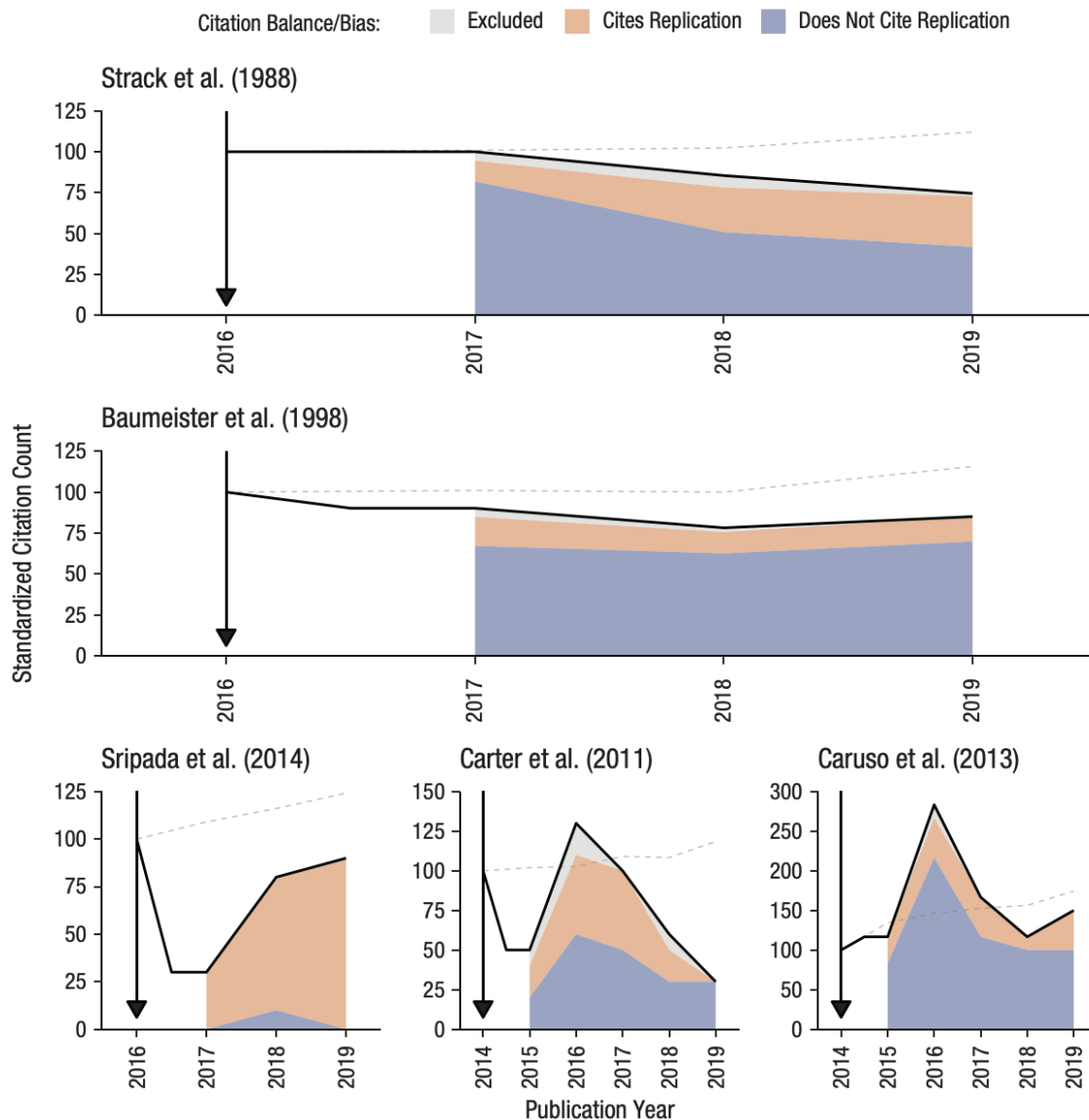


Figure 4: “Standardized annual citation counts (solid line) for the five original studies with citation balance/bias (i.e., whether the replication is cited) illustrated by colored areas in the postreplication assessment period. The dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Replication citation proportions for the Baumeister case are extrapolated to all articles in the assessment period according to a 40% random sample.” (Hardwicke et al., 2021a, p. 8). This figure is reproduced from Hardwicke et al. (2021a; Figure 2), which was shared under the CC BY-NC 4.0 license (<https://creativecommons.org/licenses/by-nc/4.0/>).

4. Discussion

4.1 Short summary of results

The main result presented in this thesis is that there is room for improvement of Open Science practices, incentives for such practices, and the way the field adjusts to the results of these practices.

The research aim of the current thesis was to investigate Open Science practices broadly construed, that is, practices aimed at making science more transparent, rigorous, and reproducible. The projects presented in this thesis focussed on the Open Data badge incentive, replication research, and overall markers of transparent and/or reproducible practices, respectively. The work was descriptive in nature, and the broad hypotheses can be described as: 1) How successful is the Open Data badge policy at *Psychological Science* in adhering to its aim of ensuring computational/results reproducibility at this journal? 2) What is the implication of strongly contradictory replication results on the citation patterns of the relevant original study?

A more detailed summary of the specific results of the projects that make up this thesis is provided in the results section above, but broadly we found that 1) the Open Data badge policy at *Psychological Science* does not seem to have been successful in ensuring reproducibility of the published results, at least in the April 2019 issue of the journal (Crüwell et al., 2023) and 2) strongly contradictory replication results do not seem to have much influence on the citation patterns of the relevant original study, at least in the case of the five original and four replication studies we considered (Hardwicke et al., 2021a).

4.2 Interpretation of results in context of the literature

The results of the Open Data badge study (Crüwell et al., 2023) are in line with those found by Obels et al. (2020) and Hardwicke et al. (2021b), if you take into account that we looked at all the published results whereas Hardwicke et al. (2021b) only looked at the “substantive” part of the results and dichotomously categorised these results as either reproducible or not reproducible (as well as with or without author assistance, for each category). The Open Data badge study both investigated all the reported results and had used more nuanced categorisation of reproducibility. Hardwicke et al. (2021b) found that only 36% of the *main* results of the articles they investigated were reproducible

without involving the original authors, and an additional 24% were reproducible with help from the authors. Laurinavichyute et al. (2022) found comparable results in an interdisciplinary but psychology-related journal (*Journal of Memory and Language*) which implemented a mandatory open data and code policy: 34% of the papers they investigated were reproducible given a strict understanding of reproducibility, and 56% given a more lenient understanding. Obels et al. (2020), a study that investigated open data sharing but not badges in particular, found a similar rate of computational reproducibility: around 33% (21 out of 62) of the articles they examined were reproducible by their team. Similarly, we found that 4/14, or about 28% of the articles we examined were at least essentially reproducible. We are certain that many of the articles rated as partially, mostly not, or even not at all reproducible would have been reproducible if we had asked the authors for their input and assistance. However, the goal of our study was not to investigate whether the relevant articles could be made computationally reproducible, but whether the Open Data badge was effective in achieving its aim, meaning that the information provided and signalled by the badge was sufficient for independent reproduction. We were able to test this in a relatively recent issue of *Psychological Science*, and provided 46 in-depth reproduction reports, giving further insights into opportunities for better sharing Open Data.

The results of the Citation Patterns study (Hardwicke et al., 2021a) are also broadly in line with those found by Serra-Garcia and Gneezy (2021) and McDiarmid et al. (2021). The main difference to McDiarmid et al. (2021) is that they found that psychology researchers and graduate students do seem to update their belief, but not as much as should be expected. It seems that, if researchers are shown both the original and the replication study next to each other, they might better take replication studies into account. Whether or not this would make a better citation balance more likely is another question, particularly as the study by McDiarmid et al. (2021) is concerned with researcher's beliefs rather than their behaviours. Our study investigated the actual citation behaviour as found in published research in the field, which arguably gives a more accurate account of the extent to which replication research is received and successfully integrated (or, as might be said is the case, not integrated).

It is important to note that both open data sharing and replication research is far from the norm in psychological research, and that any move towards increases in either of these practices is valuable. Houtkoop et al. (2018) surveyed researchers in psychology and found that most respondents openly shared data for less than 10% of their projects,

and that researchers were reluctant to share data because this is uncommon in the field and takes additional effort, among other perceived barriers. Hardwicke et al. (2022) found that only 2% of the psychological research they examined was accompanied by openly shared raw data. Similarly, despite an increased prevalence of and attention for large-scale, multi-site replication studies, replication research remains rare in psychological research. Makel et al. (2012) found that 1.6% of research in 100 psychology journals mention “replication”, only 68% of which were replications. A more recent picture of the number of replications for a sample of psychological research is given by Hardwicke et al. (2022), who found that replication was still rare, with only 5% of articles investigated reporting a replication study. Furthermore, a quick search on Scopus revealed that there were 1,898,227 articles published in the subject area psychology to date, of which 123,543 or about 6% even mention “replication” (using “replicat*” as in Makel et al., 2012; Scopus accessed via Charité-Medical University on 26th August 2022). Therefore, although both the Open Data badge project and the Citation Patterns project reveal problems with the implementation and/or incentivisation of Open Science practices, the fact that studies like this are possible at all is a sign of positive change.

4.3 Strengths and weaknesses of the studies

The main limitation of the studies presented in this thesis is that they are observational studies, which means that we have to be particularly careful when trying to draw causal inferences. Nevertheless, these kinds of studies give important insights into the issues at hand. Regarding external validity or generalisability, both the Open Data badge study and the Citation Patterns study are based on a limited sample, meaning that it is important to be careful when generalising beyond this sample. However, in the case of the Open Data badge project, there are no clear reasons to believe that results would be different for an extended or more current sample. Neither the criteria nor the badge awarding process substantially changed since the April 2019 issue of *Psychological Science*, and our results do not differ considerably from those from 2014-2015 reported in Hardwicke et al. (2021b), so we would expect studies of issues published since April 2019 to find similar results (Crüwell et al., 2023). Regarding the Citation Patterns study, it is important to note that the original studies as well as the replications were particularly prominent, and the replication results relatively unambiguous. Therefore, it is not unlikely that a study of less prominent replications and/or original studies would find very different

results, as would a study including less straightforward replication results (Hardwicke et al., 2021a).

The key strength of both studies is that the research focus is narrow, allowing for a deeper understanding of the issues, instead of a broader focus (e.g. via larger or more varied samples) which only allow for a much more shallow understanding (cf. Hardwicke et al., 2021a). For example, in the case of the Open Data badge project, we created 46 individual reproduction reports and 12 summary reports, which are filled with detailed information that could be used to better understand why it is important to share one's data and code in a reusable and reproducible way.

One difficulty we had to contend with in the Open Data badge project was the tension between anonymisation and Open Science. As the focus of our study was on the Open Data badge and its implementation at *Psychological Science*, rather than on the specific reproducibility of this particular set of articles, it was important to ensure that the authors should not be identified in the main body of the text. We therefore tried to de-identify the articles in the manuscript as much as possible: in particular, we randomly assigned numerical labels to the articles. However, complete anonymisation would be at odds with data sharing: if the aim was to fully de-identify the articles, potentially going as far as the supplementary materials, then we would not have been able to share any reanalysis code at all (or individual and summary reports; or even the specific issue of *Psychological Science* under investigation), as this clearly re-identifies the articles in any case. We therefore decided to openly share both our reproduction attempts and summary reports, as well as any analysis code we used in the reproduction process, while being as cautious as possible regarding anonymisation in the main body of the text. We also removed direct references to the original authors' names from 1) the file names of the summary reports and 2) the folder names of the individual reports.

4.4 Implications for practice and future research

The investigations presented here touch on issues that are often seen as central to science: reproducibility and replicability, and the perception of both. These considerations are of course relevant not just to psychology, but to other fields affected by the replication crisis and similar crises, such as in biomedicine (Ioannidis, 2005) or economics (Christensen & Miguel, 2018; Page et al., 2021). Future research could replicate the studies presented in this thesis in fields other than psychology. Another

avenue for further research is the expansion of the studies presented here within psychological research. Specifically, a study building on the Open Data badge project could examine the qualitative data provided by our reproduction reports in order to better understand how to best share Open Data for reuse and easy reproducibility. Another possibility is expanding the sample for an in-depth study of several issues of *Psychological Science* over a period of time, as well as investigating other journals in the field which offer badges. Similarly, a study building on the Citation Patterns project could expand its sample to more and more varied replication studies (as well as original studies), including in particular more those with more ambiguous results.

A key issue with implications for practice that we encountered in the reproduction attempts of the Open Data badge project is that of ‘software rot’ (Hinsen, 2019), across the examined articles. As an example, take Article 104: the authors shared analysis code in R using different versions of a specific package, and this enabled us to at least partially reproduce the reported results. It is likely that the provision of this code was at least somewhat coincidental as the package was changed close to when the authors submitted their manuscript. It is laudable that the authors made this further step to provide additional code, but as we mention in Crüwell et al. (2023), it would be even better to try to avoid or alleviate the effects of software rot altogether and in a more sustainable way, for example by using containers (e.g. using a platform such as Docker or Code Ocean; see for an example the Citation Patterns project container: <https://doi.org/10.24433/CO.4225975.v3>; Nüst et al., 2020; Wiebels & Moreau, 2021). Another option would be for journals to assume further responsibility by committing to ensuring that articles with an Open Data badge are fully reproducible for a certain period of time, or supporting authors in making this more likely. While this might not straightforwardly be feasible logistically or financially, a service similar to this could give journals a reason to continue to exist beyond mere prestige considerations.

A more abstract question raised by the Open Data badge project is whether the use of Open Science practices, as specifically signalled by a badge such as the Open Data badge, should be taken as an indication for research quality, credibility, or reproducibility. It has previously been argued that using such practices does signal something beyond mere accessibility of data, materials, or preregistration: for example, that it may be an indicator of more objective science (van Dongen & Sikorski, 2021). Our findings in the Open Data badge project give at least some indication that the use or signalling of Open Science practices by itself is not necessarily an indicator for

computational reproducibility—despite the fact that this is a key aim of the badge. Schneider et al. (2022) examined whether Open Science badges affect trust in the results of published articles, and found that badges increase such trust, but only in undergraduates and researchers; the public was not affected by the inclusion of badges. While the specific vignettes used distinguished between Open Data and Open Code badges, they were also accompanied by an explanation of what the badges mean, including a reference to reproducibility for the Open Data badge (Schneider et al., 2022). If the award of such a badge increases trust in a paper but does not necessarily correspond to what, presumably, gives rise to this increased trust (namely, at least in principle reproducibility), the field might need to reconsider whether the use of badges is beneficial overall. Further research on this is needed, both into the extent to which badges correspond to what they signal and into whether this signal is believed. As discussed in subsection 4.2, however, any improvement in open data sharing is arguably an improvement on the status quo.

Finally, there are some important concepts surrounding the replication crisis that need further clarification. What is a successful replication? What is a “failed” replication? Does it make sense to distinguish between “direct” and “conceptual” replications, and if so, what does that mean (cf. Machery, 2020; Nosek & Errington, 2020)? What is a preregistration, and what exactly does it do (Lakens, 2019)? Relatedly, is the dichotomy of exploratory and confirmatory research useful or intelligible at all (Scheel et al., 2021; Szollosi & Donkin, 2021)? And what are questionable research practices, really? Are they clearly pernicious, epistemically or otherwise (Erasmus, 2021; Hitzig & Stegenga, 2020)? While there are proposed answers to some aspects of these questions, there is clearly much conceptual work that needs to be done surrounding the replication crisis and Open Science. Take the first concept above as an example: the concept of replication, in the context of the current replication crisis. Machery (2020) proposes an interesting “Resampling” account of replication, but qualifies this by explaining that his work is conceptual engineering. Nosek and Errington (2020) use a similar definition as Machery (2020), which is so broad that either everything or nothing is a replication. There are further proposals for and discussions of how to understand the concept of replication (e.g., Feest, 2019; Fletcher, 2021), but overall little debate or engagement with these suggestions. Even more than a few papers on a concept so central to such a key issue in current science are not enough – there is a clear need for an ongoing and innovative theoretical debate about these topics, and we should make space for such discussions

in empirical research areas, particularly in metaresearch. It is likely that the results of the Citation Patterns project can be at least partially explained by very different and underexplained understandings of the concept of a replication.

5. Conclusion

The results presented in this thesis suggest that the practices and incentives proposed by Metaresearchers and the reform movement should continue to be monitored and evaluated to ensure their utility. What is the use of Open Science badges if they do not stand for what they signal? What is the use of carrying out replication studies if these are not taken into account on a field-wide level? Both replication research and the expansion of Open Science practices are important endeavours, but how exactly and which of these practices we (continue to) implement in order to achieve the greatest benefit is a question for extensive further research. Nevertheless, it is a sign of great progress that it was possible for either of the studies presented in this thesis to be carried out at all: Psychology as a field seems to be working towards being a more transparent, reproducible, and replicable science.

References

- Bae, G., & Luck, S. J. (2019). Reactivation of previous experiences in a working memory task. *Psychological Science*, *30*(4), 587–595.
<https://doi.org/10.1177/0956797619830398>
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. <https://doi.org/10.1037/0022-3514.74.5.1252>
- Beach, F. A. (1950). The Snark was a Boojum. *American Psychologist*, *5*(4), 115–124.
<https://doi.org/10.1037/h0056510>
- Blohowski, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., Hasselman, F., Holcombe, A. O., Humphreys, M., Lewis, M., Nosek, B. A., Peirce, J., Spies, J. R., Seto, C., Bowman, S., Green, D., Nilsson, G., Grahe, J., Wykstra, S., Hofelich Mohr, A., Sallans, A., Giner-Sorolla, R., Parker, T.H., Forstmeier, W., Nakagawa, S., Kidwell, M.C., Mellor, D.T., DeHaven, A.C., Riss, C., Lowrey, O. (2022, February 4). Badges to acknowledge open practices. <https://osf.io/tvyxz>
- Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward republicanism up to 8 months later. *Psychological Science*, *22*(8), 1011–1018.
<https://doi.org/10.1177/0956797611414726>
- Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, *142*(2), 301–306.
<https://doi.org/10.1037/a0029288>
- Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports

- at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(1), 4-17.
<https://doi.org/10.3934/Neuroscience.2014.1.4>
- Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920-80.
<https://doi.org/10.1257/jel.20171350>
- Crüwell, S., Apthorp, D., Baker, B. J., Colling, L. J., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a badge? A computational reproducibility investigation of the Open Data badge policy in one issue of *Psychological Science*. *Psychological Science*, 34(4), 512-522.
<https://doi.org/10.1177/09567976221140828>
- 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.
<https://doi.org/10.1027/2151-2604/a000387>
- Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, 30(4), 516–525.
<https://doi.org/10.1177/0956797619828724>
- Erasmus, A. (2021). *P-hacking: its costs and when it is warranted*. [Manuscript submitted for publication]. Department of Philosophy, University of Alabama.
- Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, 86(5), 895-905.
<https://doi.org/10.1086/705451>
- Fletcher, S. C. (2021). The role of replication in psychological science. *European Journal for Philosophy of Science*, 11(1), 1-19. <https://doi.org/10.1007/s13194-020-00329-2>

- Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33–41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
- Garcia, D., & Rimé, B. (2019). Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, 30(4), 617–628. <https://doi.org/10.1177/0956797619831964>
- Geniole, S. N., Procyshyn, T. L., Marley, N., Ortiz, T. L., Bird, B. M., Marcellus, A. L., Welker, K. M., Bonin, P. L., Goldfarb, B., Watson, N. V., & Carré, J. M. (2019). Using a psychopharmacogenetic approach to identify the pathways through which—and the people for whom—testosterone promotes aggression. *Psychological Science*, 30(4), 481–494. <https://doi.org/10.1177/0956797619826970>
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., Elson, J.R., J. R. Evans, B. A. Fay, B. M. Fennis, A. Finley, Z. Francis, E. Heise, H. Hoemann, M. Inzlicht, S. L. Koole, L. Koppel, F. Kroese, F. Lange, K. Lau, B. P. Lynch, C. Martijn, H. Merckelbach, N. V. Mills, A. Michirev, A. Miyake, A. E. Mosser, M. Muise, D. Muller, M. Muzi, D. Nalis, R. Nurwanti, H. Otgaar, M. C. Philipp, P. Primoceri, K. Rentzsch, L. Ringos, C. Schlinkert, B. J. Schmeichel, S. F. Schoch, M. Schrama, A. Schütz, A. Stamos, G. Tinghög, J. Ullrich, M. vanDellen, S. Wimbari, W. Wolff, C. Yusainy, O. Zerhouni, & Zwieneberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573. <https://doi.org/10.1177/1745691616652873>

- Hakim, N., Adam, K. C. S., Gunseli, E., Awh, E., & Vogel, E. K. (2019). Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. *Psychological Science*, 30(4), 526–540. <https://doi.org/10.1177/0956797619830384>
- Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, 17(1), 239-251. <https://doi.org/10.1177/1745691620979806>
- Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021a). Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/25152459211040837>
- Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021b). Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: An observational study. *Royal Society Open Science*, 8(1), 201494. <https://doi.org/10.1098/rsos.201494>
- Hardwicke, T. E., & Ioannidis, J. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, 2(11), 793-796. <https://doi.org/10.1038/s41562-018-0444-y>
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and

- analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society open science*, 5(8), 180448.
<https://doi.org/10.1098/rsos.180448>
- Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S., & Ioannidis, J. P. A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, 7, 11-37.
<https://doi.org/10.1146/annurev-statistics-031219-041104>
- Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, 14, e2000995.
<https://doi.org/10.1371/journal.pbio.2000995>
- Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L., & Bartholow, B. D. (2019). Null Effects of Game Violence, Game Difficulty, and 2D:4D Digit Ratio on Aggressive Behavior. *Psychological Science*, 30(4), 606–616.
<https://doi.org/10.1177/0956797619829688>
- Hinsen, K. (2019). Dealing with software collapse. *Computing in Science & Engineering*, 21(3), 104–108. <https://doi.org/10.1109/MCSE.2019.2900945>
- Hitzig, Z., & Stegenga, J. (2020). The problem of new evidence: P-hacking and pre-analysis plans. *Diametros*, 17(66). <https://doi.org/10.33392/diam.1587>
- Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E. J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1, 70–85. <https://doi.org/10.1177/2515245917751886>
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>

- Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLoS biology*, 13(10), e1002264. <https://doi.org/10.1371/journal.pbio.1002264>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5), 524-532. <https://doi.org/10.1177/0956797611430953>
- Johnson, D. J., & Wilson, J. P. (2019). Racial bias in perceptions of size and strength: The impact of stereotypes and group differences. *Psychological Science*, 30(4), 553–562. <https://doi.org/10.1177/0956797619827529>
- 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., 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>
- Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsson, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, 4(1). <https://doi.org/10.1525/collabra.158>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J.R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J.A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J.

- L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A., & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221-230. https://doi.org/10.24602/sjpr.62.3_221
- Laurinavichyute, A., Yadav, H., & Vasisht, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the *Journal of Memory and Language* under the open data policy. *Journal of Memory and Language*, 125, 104332. <https://doi.org/10.1016/j.jml.2022.104332>
- Lindsay, L., Gambi, C., & Rabagliati, H. (2019). Preschoolers optimize the timing of their conversational turns through flexible coordination of language comprehension and production. *Psychological Science*, 30(4), 504–515. <https://doi.org/10.1177/0956797618822802>
- Machery, E. (2020). What Is a Replication? *Philosophy of Science*, 87(4), 545-567. <https://doi.org/10.1086/709701>
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, 7(6), 537–542. <https://doi.org/10.1177/1745691612460688>
- McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, J. E. (2021). Psychologists update their beliefs about effect sizes after replication studies. *Nature human behaviour*, 5(12), 1663-1673. <https://doi.org/10.1038/s41562-021-01220-7>

- Nelson, N. (2020). Towards an Expanded Conception of Publication Bias. *Journal of Trial & Error*, 1(1). <https://doi.org/10.36850/mr2>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73, 719-748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Errington, T. M. (2020). What is replication?. *PLoS Biology*, 18(3), e3000691. <https://doi.org/10.1371/journal.pbio.3000691>
- Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra: Psychology*, 3(1). <https://doi.org/10.1525/collabra.102>
- Nüst, D., Sochat, V., Marwick, B., Eglen, S. J., Head, T., Hirst, T., & Evans, B. D. (2020). Ten simple rules for writing Dockerfiles for reproducible data science. *PLoS Computational Biology*, 16(11), e1008316. <https://doi.org/10.1371/journal.pcbi.1008316>
- Obaidi, M., Bergh, R., Akrami, N., & Anjum, G. (2019). Group-based relative deprivation explains endorsement of extremism among Western-born Muslims. *Psychological Science*, 30(4), 596–605. <https://doi.org/10.1177/0956797619834879>

- 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>
- Olsson-Collentine, A., van Assen, M. A. L. M., & Hartgerink, C. H. J. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, 30(4), 576–586. <https://doi.org/10.1177/0956797619830326>
- Page, L., Noussair, C. N., & Slonim, R. (2021). The replication crisis, the rise of new research practices and what it means for experimental economics. *Journal of the Economic Science Association*, 7(2), 210-225. <https://doi.org/10.1007/s40881-021-00107-7>
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science. *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- Psychological Science. (2022, February 1). *Psychological Science Submission Guidelines*. Retrieved March 16, 2022, from https://www.psychologicalscience.org/publications/psychological_science/ps-submissions
- Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research: A comparative evaluation of six statistical methods. *Zeitschrift für Psychologie*, 227(4), 261. <https://doi.org/10.1027/2151-2604/a000386>
- Rowhani-Farid, A., & Barnett, A. G. (2018). Badges for sharing data and code at *Biostatistics*: An observational study. *F1000Research*, 7, 90. <https://doi.org/10.12688/f1000research.13477.2>

- Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in *BMJ Open*? A randomized controlled trial. *Royal Society Open Science*, 7(3), 191818. <https://doi.org/10.1098/rsos.191818>
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. <https://doi.org/10.1177/1745691620966795>
- Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2022) Do Open Science Badges increase trust in scientists among undergraduates, scientists, and the public? *Psychological Science*. <https://doi.org/10.1177%2F09567976221097499>
- Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science advances*, 7(21), eabd1705. <https://doi.org/10.1126/sciadv.abd1705>
- 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>
- Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science*, 25(6), 1227–1234. <https://doi.org/10.1177/0956797614526415>
- Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584-2589. <https://doi.org/10.1073/pnas.1708290115>
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the Facial Feedback Hypothesis. *Journal of*

Personality and Social Psychology, 54(5), 768–777. <https://doi.org/10.1037/0022-3514.54.5.768>

Szollósi, 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>

van Dongen, N., & Sikorski, M. (2021). Objectivity for the research worker. *European Journal for Philosophy of Science*, 11(3), 93. <https://doi.org/10.1007/s13194-021-00400-6>

Vardy, T., & Atkinson, Q. D. (2019). Property damage and exposure to other people in distress differentially predict prosocial behavior after a natural disaster. *Psychological Science*, 30(4), 563–575. <https://doi.org/10.1177/0956797619826972>

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., K. Dijkstra, A. H. Fischer, F. Foroni, U. Hess, K. J. Holmes, J. L. H. Jones, O. Klein, C. Koch, S. Korb, P. Lewinski, J. D. Liao, S. Lund, J. Lupianez, D. Lynott, C. N. Nance, S. Oosterwijk, A. A. Ozdoğru, A. P. Pacheco-Unguetti, B. Pearson, C. Powis, S. Riding, T.-A. Roberts, R. I. Rumiati, M. Senden, N. B. Shea-Shumsky, K. Sobocko, J. A. Soto, T. G. Steiner, J. M. Talarico, Z. M. van Allen, M. Vandekerckhove, B. Wainwright, J. F. Wayand, R. Zeelenberg, Zetzer, E. E., & Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on*

Psychological Science, 11(6), 917–928.

<https://doi.org/10.1177/1745691616674458>

Wiebels, K., & Moreau, D. (2021). Leveraging containers for reproducible psychological research. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211017853. <https://doi.org/10.1177/25152459211017853>

Wójcik, M. J., Nowicka, M. M., Bola, M., & Nowicka, A. (2019). Unconscious detection of one's own image. *Psychological Science*, 30(4), 471–480. <https://doi.org/10.1177/0956797618822971>

Woolley, K., & Fishbach, A. (2019). Shared plates, shared minds: Consuming from a shared plate promotes cooperation. *Psychological Science*, 30(4), 541–552. <https://doi.org/10.1177/0956797619830633>

Yousif, S. R., & Keil, F. C. (2019). The additive-area heuristic: An efficient but illusory means of visual area approximation. *Psychological Science*, 30(4), 495–503. <https://doi.org/10.1177/0956797619831617>

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41. <https://doi.org/10.1017/S0140525X17001972>

Statutory Declaration

"I, Sophia Crüwell, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic "Open Science Practices and Incentives – a Meta-Research Investigation" or "Open Science Praktiken und Anreize – eine Meta-wissenschaftliche Untersuchung", independently and without the support of third parties, and that I used no other sources and aids than those stated.

All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; <http://www.icmje.org>) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me."

Date

Signature

Declaration of own contribution to the publications

Sophia Crüwell contributed the following to the below listed publications:

Top-Journal Publication: **Sophia Crüwell**, Deborah Apthorp, Bradley James Baker, Lincoln Colling, Malte Elson, Sandra J Geiger, Sebastian Lobentanzer, Jean Monéger, Alex Patterson, D Sam Schwarzkopf, Mirela Zaneva, Nicholas JL Brown, Investigating the Effectiveness of the Open Data Badge Policy at Psychological Science Through Computational Reproducibility, *Psychological Science*, 2023.

Impact Factor (2021): 10.172; Impact Factor (2020): 7.029

Contribution: Sophia Crüwell took over the organisation and conceptualisation of the project from the senior author before all the individual reports were written, and before the majority of the summary reports were finished. This means that she largely coordinated and pushed forward the summary reports, for example by creating a template summary report. Sophia Crüwell also conceptualised the framing, the specific wording and the survey for the reproducibility ratings, in order to more straightforwardly represent each author's opinions about reproducibility of the individual papers, as well as to show any changes and variations of reproducibility ratings. The tables of the reproducibility ratings (Tables 1 and 2) were also produced by Sophia Crüwell. She wrote the first full draft of the manuscript. Sophia Crüwell contributed individual reproduction attempts for articles 101, 109, and 112, which involved different statistical analyses of existing data. She also created initial drafts of summary reports for articles 101 and 109. Table 3 is partially based on another summary table she created (<https://osf.io/3h9j5/>). Sophia Crüwell was responsible for the revisions of the article during the extensive peer review process, and did most of this revision work herself.

Publication 2: Tom E Hardwicke, Dénes Szűcs, Robert T Thibault, **Sophia Crüwell**, Olmo R van den Akker, Michèle B Nuijten, John PA Ioannidis, Citation patterns following a strongly contradictory replication result: Four case studies from psychology, *Advances in Methods and Practices in Psychological Science*, 2021.

Impact Factor (2021; no prior IF available): 15.817

Contribution: As stated in the "Author Contributions" section of this published article, Sophia Crüwell performed the data collection alongside T. E. Hardwicke, D. Szűcs, R. T. Thibault, O. R. van den Akker, and M. B. Nuijten, and she performed the data analysis with T. E. Hardwicke (see Hardwicke et al., 2021). Specifically, as is noted on the relevant data sheets, Sophia Crüwell first coded more than 100 articles and second coded more than 40 articles, and together with O. R. van den Akker did the supplementary coding of 63 articles to examine whether there was an explicit or absent defense pattern. This latter coding was done following conceptual discussions between T.E.Hardwicke, O. R. van den Akker, and Sophia Crüwell. Regarding the analysis work, T.E. Hardwicke created an initial report of the analyses based on his and Sophia Crüwell's initial analysis efforts. Specifically, she independently did a first run through of the major analyses and created first versions of figure 2 and table 4. She also gave feedback on the manuscript and approved its final version for submission.

Signature, date and stamp of first supervising university professor / lecturer

Signature of doctoral candidate

Excerpt from Journal Summary List

Note on the AMPPS Impact Factor used

The journal *Advances in Methods and Practices in Psychological Science* (AMPPS) is a relatively new Association for Psychological Science (APS) journal: its first article appeared in February 2018. Therefore, until the 2021 Journal Summary List, it does not appear to have had an impact factor in the ISI Web of Science framework. We submitted the article "Citation patterns following a strongly contradictory replication result: four case studies from psychology" (Hardwicke et al., 2021) in February 2021, meaning that the 2019 lists would have been current. However, AMPPS did not yet appear on a 2019 list, as the journal was only launched in 2018. In the 2021 lists, AMPPS has an impact factor in the ISI Web of Science system for the first time. This 2021 impact factor is 15.817, and AMPPS thus ranks 5/148 on the "PSYCHOLOGY, MULTIDISCIPLINARY" list, as well as 5/77 on the "PSYCHOLOGY" list.

Journal Data Filtered By: **Selected JCR Year: 2020** Selected Editions: SCIE, SSCI
 Selected Categories: **"PSYCHOLOGY, MULTIDISCIPLINARY"** Selected Category
 Scheme: WoS

Gesamtanzahl: 140 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	Psychological Science in the Public Interest	2,337	37.857	0.003510
2	Annual Review of Psychology	28,065	24.137	0.018970
3	PSYCHOLOGICAL BULLETIN	66,720	17.737	0.023450
4	PSYCHOLOGICAL METHODS	18,605	11.302	0.015170
5	AMERICAN PSYCHOLOGIST	36,450	10.885	0.014740
6	Perspectives on Psychological Science	14,496	9.837	0.023420
7	European Journal of Psychology Applied to Legal Context	454	9.300	0.000680
8	PSYCHOLOGICAL REVIEW	35,190	8.934	0.008600
9	PSYCHOLOGICAL SCIENCE	40,033	7.029	0.034290
10	COMPUTERS IN HUMAN BEHAVIOR	45,035	6.829	0.059730
11	CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE	16,454	6.811	0.015610
12	Emotion Review	3,305	6.469	0.005090
13	Body Image	6,377	6.406	0.006410
14	ENVIRONMENT AND BEHAVIOR	8,885	6.222	0.004000
15	Current Opinion in Psychology	5,732	5.717	0.020510
16	EUROPEAN PSYCHOLOGIST	2,372	5.569	0.001880
17	JOURNAL OF ENVIRONMENTAL PSYCHOLOGY	15,100	5.192	0.008290
18	Psychosocial Intervention	786	5.083	0.000770
19	ANNALS OF BEHAVIORAL MEDICINE	9,314	4.908	0.007840

Journal Data Filtered By: **Selected JCR Year: 2021** Selected Editions: SCIE,SSCI
 Selected Categories: **"PSYCHOLOGY, MULTIDISCIPLINARY"** Selected Category
 Scheme: WoS

Gesamtanzahl: 148 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfaktor
1	Psychological Science in the Public Interest	2,781	56.200	0.00370
2	Annual Review of Psychology	30,374	27.782	0.01502
3	PSYCHOLOGICAL BULLETIN	68,919	23.027	0.02074
4	AMERICAN PSYCHOLOGIST	40,383	16.358	0.01771
5	Advances in Methods and Practices in Psychological Science	2,016	15.817	0.00816
6	Perspectives on Psychological Science	17,025	11.621	0.02197
7	PSYCHOLOGICAL METHODS	20,525	10.929	0.01423
8	Qualitative Research in Psychology	14,928	10.568	0.00314
9	PSYCHOLOGICAL SCIENCE	42,641	10.172	0.02762
10	European Journal of Psychology Applied to Legal Context	526	9.850	0.00068
11	COMPUTERS IN HUMAN BEHAVIOR	53,712	8.957	0.05434
12	PSYCHOLOGICAL REVIEW	36,958	8.247	0.00854
13	CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE	17,694	7.867	0.01421
14	JOURNAL OF ENVIRONMENTAL PSYCHOLOGY	17,441	7.649	0.00849
15	Emotion Review	3,923	7.345	0.00495
16	Current Opinion in Psychology	8,090	6.813	0.02039
17	ENVIRONMENT AND BEHAVIOR	9,422	6.548	0.00376
18	FEMINISM & PSYCHOLOGY	1,674	5.833	0.00180
19	PSYCHOLOGICAL INQUIRY	7,168	5.581	0.00237
20	Body Image	6,788	5.580	0.00574

Printing copies of the publications

Publication 1: Open Data Badge Project

Crüwell, S., Apthorp, D., Baker, B. J., Colling, L. J., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., & Brown, N. J. L. (2023). What's in a badge? A computational reproducibility investigation of the Open Data badge policy in one issue of *Psychological Science*. *Psychological Science*, 34(4), 512-522.

DOI: <https://doi.org/10.1177/09567976221140828>



Editor's Note

In the article that follows this Editor's Note, Crüwell and colleagues report the results of an audit of the computational reproducibility of the 14 research articles published in the April 2019 issue of *Psychological Science* (Vol. 30, Issue 4). The audit was author-initiated—it was not by invitation of the journal. Crüwell and colleagues defined computational reproducibility as “the ability to recreate results using the original data and code (or at least a detailed description of the analyses)” (p. 514). They selected Volume 30, Issue 4 because it was the first in which all of the research articles were awarded the Open Data badge. Of the 14 research articles in the issue, Crüwell et al. assessed only one as meeting the requirements for the Open Data badge.

In their assessment, Crüwell and colleagues relied on the criteria provided in the Submission Guidelines of the journal at the time the 2019 authors submitted their articles (*Psychological Science*, Submission Guidelines, Open Science Badges section). The guidelines state that authors may receive an “Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported result. This includes annotated copies of the code or syntax used for all exploratory and principal analyses.” In their judgments regarding Open Data badge eligibility, Crüwell and colleagues emphasized the availability of analysis code or syntax. Importantly, neither the Open Science Framework (OSF) criteria that guide badge eligibility nor the Open Practices Disclosure (OPD) form completed by the 2019 authors makes explicit reference to analysis code or syntax. The OSF criteria state that “The Open Data badge is awarded when digitally-shareable data necessary to reproduce the reported results are publicly available,” and that “A data dictionary (for example, a codebook or metadata describing the data) is included with sufficient description for an independent researcher to reproduce the reported analyses and results” (<https://osf.io/tyvzx/wiki/1.%20View%20the%20Badges/>). Similarly, the OPD form that the authors completed required them to “Confirm that there is sufficient information for an independent researcher to reproduce **all of the reported results**, including codebook if relevant” (emphasis in original). Neither set of criteria specifies sharing of analysis code or syntax.

The difference between the Submission Guidelines and the OPD form that authors completed is important. The Submission Guidelines provide advice, but they are not the rule of law. The rule of law is established in the OPD form, which at the time the 2019 authors completed it, made no mention of analysis code. I emphasize this point because it establishes that the 2019 authors did not openly flaunt explicit criteria when they applied for the Open Data badge, and nor did eligibility for the badge turn on provision of analysis code, as established either by *Psychological Science* or OSF.

On behalf of *Psychological Science*, I apologize for the discrepancy between the Open Data badge elements listed in the Submission Guidelines and the less explicit requirement in the OPD form. The criteria outlined in the OPD form were not sufficiently explicit regarding the elements that should be included in an open-access registry in order to ensure independent reproducibility. We have changed the wording of the OPD form such that it now provides better guidance to authors in their efforts to make their science open by making their data publicly available.











Setting aside for the moment the vagueness of the requirements of the previous version of the OPD form, it is clear that it instructed authors to provide “sufficient information for an independent researcher to reproduce all of the reported results.” The OSF eligibility criteria give the same charge. By their report, for several of the articles published in the April 2019 issue, the audit team of Crüwell and colleagues was not successful in achieving the goal of independent reproduction of all of the reported results based on the information in the registry alone, with the methods they employed. Importantly, ensuring that analyses can be reproduced is only one of several possible motivations for authors to make their data openly accessible. Other possible motivations include reducing the need for duplicative data-collection efforts; facilitating collaborations; and even enabling analysis of data in different ways, thus helping to ensure findings are robust to different analytic approaches, to name a few. I venture to guess that it was goals such as these, not independent reproducibility alone, that were paramount in the minds of many of the 2019 authors as they made their data publicly available.

Critically, transparency and scientific community building are not mutually exclusive goals. In this regard, it is my pleasure to report that upon learning of the work of Crüwell and colleagues, several of the author groups with articles in Volume 30, Issue 4 of *Psychological Science* appended their registries to include elements identified in the audit as missing or insufficient. I appreciate the positive response of these author groups and their ongoing contributions to open science.

Patricia J. Bauer
Editor in Chief

What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in One Issue of *Psychological Science*



Sophia Crüwell^{1,2} , Deborah Apthorp^{3,4} , Bradley J. Baker⁵ ,
Lincoln Colling⁶ , Malte Elson^{7,8} , Sandra J. Geiger⁹ ,
Sebastian Lobentanzer¹⁰, Jean Monéger^{11,12} , Alex Patterson¹³ ,
D. Samuel Schwarzkopf^{14,15}, Mirela Zaneva¹⁶ , and
Nicholas J. L. Brown¹⁷ 

¹Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité – Universitätsmedizin Berlin; ²Department of History and Philosophy of Science, University of Cambridge; ³School of Psychology, University of New England; ⁴School of Computing, Australian National University; ⁵Department of Sport and Recreation Management, Temple University; ⁶School of Psychology, University of Sussex; ⁷Faculty of Psychology, Ruhr University Bochum; ⁸Horst Görtz Institute for IT Security, Ruhr University Bochum; ⁹Environmental Psychology, Department of Cognition, Emotion, and Methods, Faculty of Psychology, University of Vienna; ¹⁰Institute for Computational Biomedicine, University Hospital Heidelberg, Germany; ¹¹Department of Psychology, University of Poitiers; ¹²Research Center on Cognition and Learning, Centre National de la Recherche Scientifique (CNRS) 7295; ¹³Sheffield Methods Institute, The University of Sheffield; ¹⁴School of Optometry and Vision Science, University of Auckland; ¹⁵Experimental Psychology, University College London; ¹⁶Department of Experimental Psychology, University of Oxford; and ¹⁷Department of Psychology, Linnaeus University

Psychological Science
2023, Vol. 34(4) 513–522
© The Author(s) 2023



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976221140828
www.psychologicalscience.org/PS



Abstract

In April 2019, *Psychological Science* published its first issue in which all Research Articles received the Open Data badge. We used that issue to investigate the effectiveness of this badge, focusing on the adherence to its aim at *Psychological Science*: sharing both data and code to ensure reproducibility of results. Twelve researchers of varying experience levels attempted to reproduce the results of the empirical articles in the target issue (at least three researchers per article). We found that all 14 articles provided at least some data and six provided analysis code, but only one article was rated to be exactly reproducible, and three were rated as essentially reproducible with minor deviations. We suggest that researchers should be encouraged to adhere to the higher standard in force at *Psychological Science*. Moreover, a check of reproducibility during peer review may be preferable to the disclosure method of awarding badges.

Keywords

open data, data sharing, open badges, reproducibility, journal policy, open data

Received 4/1/22; Revision accepted 9/30/22

Open science badges are incentives for researchers to participate in open science practices such as preregistration and sharing of data and materials. Sharing data is encouraged in order to increase transparency, reuse

or reproducibility, and citations (Colavizza et al., 2020; Piwowar & Vision, 2013). *Psychological Science* adopted

Corresponding Author:

Sophia Crüwell, Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Charité – Universitätsmedizin Berlin
Email: sophia.cruwell@bih-charite.de

Correction (March 2023): This article has been updated with the Open Data badge.

the badges in 2014 (Eich, 2014), and, in April 2019, published its first issue in which all 14 Research Articles received Open Data badges (Volume 30, Issue 4). The aim of this badge is to incentivize authors to share online the data necessary to reproduce the reported results (Blohowiak et al., 2022). *Psychological Science's* submission guidelines state that articles may receive this badge “for making publicly available the digitally shareable data necessary to reproduce the reported result. This includes annotated copies of the code or syntax used for all exploratory and principal analyses” (Psychological Science, 2022, Open Practices Badges section; these eligibility criteria were operative in 2019).¹ The corresponding Open Practices Disclosure form uses somewhat more permissive language, requiring confirmation of “sufficient information for an independent researcher to reproduce all of the reported results.” This equates to provision of analysis code or syntax for all but the simplest analyses and data sets. We understand *reproducibility* to mean computational reproducibility: the ability to recreate results using the original data and code (or at least a detailed description of the analyses). *Psychological Science* awards badges based on the *disclosure method*: Authors complete an Open Practices Disclosure form, and the journal may confirm the existence of data, materials, or a preregistration (Blohowiak et al., 2022; Psychological Science, 2022).

Kidwell et al. (2016) found that introducing badges at *Psychological Science* led to an increase in sharing, which indicates the superficial success of this policy—particularly compared with other initiatives (see Rowhani-Farid & Barnett, 2018, and Rowhani-Farid et al., 2020, who found lower and no increase in data sharing at *Biostatistics* and *BMJ Open*, respectively). Hardwicke et al. (2021) investigated the analytic reproducibility of articles that received Open Data badges at *Psychological Science* between 2014 and 2015; they were able to reproduce the results of 36% of articles without author involvement and a further 24% with author involvement. Obels et al. (2020) examined data sharing and computational reproducibility of registered reports in general psychological research; 36 of the 62 articles assessed (58%) provided both data and code, of which 21 (58%) were computationally reproducible.

Whereas Hardwicke et al. (2021) and Obels et al. (2020) were concerned with computational or analytic reproducibility per se, we focused on computational reproducibility as a measure of the effectiveness of the *Psychological Science* Open Data badge policy. If this policy was effective, the results in the April 2019 issue should be independently and precisely reproducible. If these results are wholly or partially irreproducible, then any issues we identify during reproduction attempts may inform the improvement of the policy at *Psychological Science* and other journals. Our focus on

Statement of Relevance

Open science badges are incentives for encouraging researchers to participate in open science practices such as preregistration and the sharing of data or experimental materials. These practices are thought to be desirable as a means for enhancing both transparency and reproducibility, which are important to scientific inquiry. In particular, the results of a study should be at least computationally reproducible using the same data and analyses. In the present study, we aimed specifically to investigate the effectiveness of the Open Data badge at *Psychological Science*, the stated purpose of which is to ensure the reproducibility of results. We found that the Open Data badge policy did not work as intended, and we suggest possible changes in how the badge could be awarded. We hope to contribute to improving the badge program at *Psychological Science* as well as reproducibility and transparency in psychology.

one practice in one issue of *Psychological Science* allows for in-depth examination of the effectiveness of this specific measure for incentivizing data sharing as implemented and advertised at this journal.

Open Practices Statement

The individual and summary reports, as well as the informal reproducibility ratings and code to create Tables 1 and 2, are publicly accessible at <https://osf.io/xzke7/>. This study was not preregistered.

Method

Sample

The scope of our investigation was all 14 Research Articles published in the April 2019 issue of *Psychological Science*, the journal's first issue in which all Research Articles were awarded the Open Data badge (Bae & Luck, 2019; Dorfman et al., 2019; Garcia & Rimé, 2019; Geniole et al., 2019; Hakim et al., 2019; Hilgard et al., 2019; Johnson & Wilson, 2019; Lindsay et al., 2019; Obaidi et al., 2019; Olsson-Collentine et al., 2019; Vardy & Atkinson, 2019; Wójcik et al., 2019; Woolley & Fishbach, 2019; Yousif & Keil, 2019). To emphasize our focus on *Psychological Science's* Open Data badge policy and not these individual articles, we will refer to them as Articles 101 to 114, the numbers having been randomly assigned. A superficial examination of the repositories linked to the articles shows that all articles are associated

Table 1. Initial Ratings: Reproducers' Ratings of Their Initial Reproduction Attempts for Each Article

Article	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Modal rating
101	Partially	Essentially	Partially			Partially
102	Partially	Essentially	Partially	Essentially	Partially	Partially
103	Partially	Partially	Partially			Partially
104	Partially	Partially	Essentially			Partially
105	Not at all	Not at all	Not at all			Not at all
106	Not at all	Mostly not	Not at all			Not at all
107	Not at all	Not at all	Not at all	Not at all		Not at all
108	Partially	Exactly	Exactly			Exactly
109	Partially	Partially	Essentially			Partially
110	Mostly not	Essentially	Mostly not			Mostly not
111	Essentially	Essentially	Essentially			Essentially
112	Mostly not	Not at all	Mostly not			Mostly not
113	Partially	Partially	Partially			Partially
114	Partially	Essentially	Essentially	Essentially		Essentially

with at least some data. No code is provided in the linked repository for six of the articles (Articles 101, 105, 107, 111, 112, and 113).

Design

This is an observational, descriptive, one-group study. We did not compare the April 2019 issue of *Psychological Science* with any other issue or journal but rather to the ideal of the policy of the Open Data badge as implemented at *Psychological Science*.

In the present study, we were mainly concerned with this Open Data badge policy's effectiveness, not with reproducibility per se. Our informal reproducibility ratings are a proxy measure of that effectiveness. Although

we did not establish any criteria for successful reproduction in advance, for a study to count as reproducible, its results should at least be reproducible by a competent external researcher (National Academies of Sciences, Engineering, and Medicine, 2019), such as a PhD student with some experience and training in a similar field. When we say that a study was or was not reproducible, this is specific to our team of reproducers. Our informal reproducibility rating items were "exactly reproducible," which represented the ideal of the Open Data badge in which there were no deviations from the reported results; "essentially reproducible," meaning that there were minor deviations in the decimals or obvious typographical errors (e.g., 2.39 vs. 2.93); "partially reproducible," indicated that there were more than minor deviations but the

Table 2. Summary Ratings: Reproducers' Ratings of the Group's Reproduction Attempts for Each Article

Article	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Modal Rating
101	Essentially	Essentially	Partially			Essentially
102	Partially	Essentially	Partially	Partially	Partially	Partially
103	Partially	Partially	Partially			Partially
104	Partially	Partially	Partially			Partially
105	Not at all	Not at all	Not at all			Not at all
106	Mostly not	Not at all	Not at all			Not at all
107	Not at all	Mostly not	Not at all	Not at all		Not at all
108	Exactly	Exactly	Exactly			Exactly
109	Partially	Essentially	Essentially			Essentially
110	Partially	Partially	Partially			Partially
111	Essentially	Essentially	Essentially			Essentially
112	Mostly not	Not at all	Mostly not			Mostly not
113	Partially	Partially	Partially			Partially
114	Partially	Essentially	Partially	Essentially		Partially

results were mostly numerically consistent; “mostly not reproducible,” meaning that there were major deviations and few numerically consistent results; and “not at all reproducible” if there was no numerical consistency between the reported results and the ones that we found, or a reproduction attempt was otherwise not possible.

Procedure

Reproducer assignment. The last author initially recruited 13 researchers of varying experience and career levels to attempt to reproduce studies from the April 2019 issue of *Psychological Science* on the basis of the data and, where available, code shared by the original authors. They were asked to indicate their ability to access and use four software packages: Excel, MATLAB, R, and SPSS. Each reproducer was asked to attempt to reproduce four of the 14 articles, the selection being determined by (a) the match between the reproducer’s access to software and the format of the code or data provided by the original authors, and (b) the aim to have distinct sets of researchers working on each article, where possible. Because of an error in the assignment process, two reproducers (J.M. and S.L.) were asked to reproduce the same four articles. No two articles were reproduced by the exact same set of researchers. Two reproducers dropped out and did not complete any reproduction reports. Furthermore, reproducers were unable to complete individual reproduction attempts because of technical limitations in three cases (B. J. B., Article 106; S. C., Article 110; S. J. G., Article 112). One further reproducer joined the project at a later stage. In total, 12 reproducers completed three to five reproductions each. For each of the 14 articles, at least three researchers were assigned to, and completed, individual reproduction reports (46 individual reports in total).

Reproduction process. The reproduction process was split into two stages. In the first stage, each researcher independently attempted to reproduce their assigned studies and wrote an individual reproduction report on their experience and findings. These initial reports were unstructured; some reproducers included further information such as code, whereas others focused on the narrative report of their reproduction attempts. Results were initially not shared, and reproducers were encouraged to stay as masked as possible (i.e., not discussing results with other reproducers until their own analyses were completed). In the second stage, on the basis of the individual reports, the groups of reproducers for each article agreed on a summary report of their overall findings. After the reproduction process, they rated the reproducibility of each article they had attempted to reproduce on the basis of (a) their individual, initial experience reproducing the article and

(b) the summary findings and discussions among the group for each article.

All of our reproduction attempts were carried out independently of the articles’ original authors. We then contacted the authors prior to preprinting and submission to explain the nature of the project; all our analyses and conclusions were finalized by that point. In the case of two articles, the last author of the present article had previously (i.e., before the other coauthors joined the project in May 2020) contacted the corresponding authors for reproduction advice before realizing that this was not compatible with the overarching aim of the project. Consequently, he did not write an individual report on these articles, and he did not contribute to the associated group discussions.

Results

Reproducibility

Only one of the 14 articles was rated to be exactly reproducible (Article 108), and three further articles were rated essentially reproducible with minor deviations by a majority of the researchers who reproduced them, on the basis of the summary reports (Articles 101, 109, and 111). Both the initial reproducibility ratings based on the individual reproduction attempts (Table 1) and the summary ratings based on the article group’s combined reproduction attempts (Table 2) varied, and there were four changes between the modal majority-agreed initial and summary ratings (Articles 101, 109, 110, and 114).

The individual reports (46 total) and summary reports (14 total) are available on the OSF alongside further information about each reproduced article (see <https://osf.io/xzke7/>). The reports provide in-depth qualitative and quantitative information in the form of narrative descriptions of each reproduction attempt, often including numerical results.

Issues encountered

The following section qualitatively and nonexhaustively summarizes the issues that we encountered (for a further summary of the shared data and code, see Table 3). General issues include (a) a lack of documentation of data and/or code; (b) minor discrepancies in several results, likely due to use of random numbers without fixed seeds in bootstrapped analyses; (c) minor discrepancies in individual results, likely due to typographical or copy-paste errors; (d) unclear reporting of procedures in the article text, including the criteria for inclusion in subgroups, lack of or incorrect reporting of the variables used for regression models, and

Table 3. Summary of the Results Reported in the Summary Reproduction Reports for Each Article

Article	Results (summary rating)	Analytic code	Data	Readme file	Variable key	Other
101	Essentially reproducible	Missing	Postprocessed provided	Missing	Missing	Missing data for one experiment
102	Partially reproducible	Provided	Postprocessed provided	Missing	Missing	Inconsistencies in data from what was reported in article
103	Partially reproducible	Provided	Raw provided	Missing	Missing	Broken GitHub links, key file not linked to in repository
104	Partially reproducible	Provided	Postprocessed provided	Provided	Provided	Different reproducers had different issues running code
105	Not reproducible	Missing	Postprocessed Provided	Missing	Provided	Data for Supplemental Material were missing
106	Not reproducible	Insufficient	Raw provided	Missing	Missing	Required extra MATLAB packages
107	Not reproducible	Missing	Raw provided	Provided	Provided	Insufficient information
108	Exactly reproducible	Provided	Raw provided	Missing	Provided	Package dependency issues
109	Essentially reproducible	Provided	Postprocessed provided	Missing	Missing	Unclear whether data were raw or postprocessed
110	Partially reproducible	Insufficient	Raw provided	Missing	Missing	Corrupt data/unable to download data
111	Essentially reproducible	Missing	Postprocessed provided	Missing	Missing	Preregistration discrepancies
112	Mostly not reproducible	Insufficient	Postprocessed Provided	Provided	Provided	Required extra MATLAB packages
113	Partially reproducible	Missing	Postprocessed provided	Missing	Missing	Unclear variable identification
114	Partially reproducible	Provided	Postprocessed provided	Missing	Provided	Corrupt data/unable to download data

unreported one-sided analyses; (e) data storage issues on the OSF, including files being either corrupt or not downloadable at all (Article 110); and (f) ambiguous labeling of studies in the article's Open Practices statement (Article 109). Data-specific issues include (a) provision of cleaned data without raw data, (b) provision of raw data without cleaned data, and (c) no description of, or code for, the data-cleaning process. Code-specific issues include (a) a lack of shared analysis code or modeling code and (b) issues with package or software versions (often resolvable but sometimes only with considerable effort).

Open Data badge eligibility

Overall, we found that eight articles (Articles 101, 105, 106, 107, 110, 111, 112, and 113) did not provide, even in principle, sufficient information for independent

exact reproduction of their results by our team. In these cases, reproduction would require analysis code or syntax, as the descriptions of the methodology and the shared data files did not provide enough information on their own.² This means that (a) these articles did not meet the standard for receiving the Open Data badge at *Psychological Science* according to the explicit requirements stated in the submission guidelines, and (b) the authors of these articles may have interpreted the less explicit requirements of the Open Practices Disclosure statement in a rather minimalist way.

Provision of both analysis code and data was a requirement for the award of an Open Data badge at *Psychological Science* at the time of submission, according to the explicit requirements stated in the submission guidelines. These requirements appear to not have been met in these cases. Articles missed these explicit requirements of the journal submission guidelines to different extents. Six

articles (Articles 101, 105, 107, 111, 112, and 113) did not provide any code in the linked repository (some modeling code was provided for Article 112 on a separate GitHub page not linked to from the article), and Article 101 additionally provided only summarized and incomplete data. Therefore, these articles do not appear to have met the requirements for receiving the Open Data badge, according to the explicit requirements in the submission guidelines that were in force at *Psychological Science* when the articles were first submitted. Arguably, given this stipulation, Articles 106 and 110 were also not eligible for the Open Data badge because they provided some code files but not the statistical analysis code. This field-leading policy was certainly introduced and implemented with the best of intentions, but there appear to have been some oversights by the journal in its execution, as the OSF guidelines recommend at least a cursory check by the journal before the badge is awarded.

On top of these clearer eligibility issues regarding the provision of sufficient information and/or analysis code for independent exact reproduction, on a strict interpretation of the badge eligibility criteria at *Psychological Science*, our reproduction results arguably imply that only one of the 14 articles met the requirements for an Open Data badge. Eight articles did not share both data and analysis code or otherwise sufficient information, and of the remaining six articles that did attempt to share sufficient information for independent reproduction in the form of analysis code, only one was exactly reproducible by our team. However, the reproducibility of the articles that shared data and analysis code likely decreased since publication (because of issues such as “software rot”; Hinsén, 2019). Therefore, it is unclear how we can make an inference from current reproducibility to past Open Data badge eligibility in the case of the articles that share both data and analysis code but were not exactly reproducible.

Discussion

The disclosure method did not ensure the required higher standard for the Open Data badge at *Psychological Science*, at least in its April 2019 issue. Of 14 articles, eight did not share both data and analysis code and so failed to meet the eligibility requirements. Of the remaining six, only one was exactly reproducible, but we do not know whether the other five were exactly reproducible at the time of submission. We make several recommendations for improving the specific badge policy at *Psychological Science* and comparable initiatives at other journals (for further general recommendations on improving data sharing and computational reproducibility, see Stodden et al., 2016; Trisovic et al.,

2022; Wilson et al., 2017). Excellent and more in-depth recommendations and tutorials for authors to ensure that their shared data and code are eligible for an Open Data badge are provided by, for example, Arslan (2019), Eberle (2022), Klein et al. (2018), Levenstein and Lyle (2018), Peikert and Brandmaier (2021), and Van Lissa et al. (2021). Moreover, the provision of further incentives, in particular by funding agencies and institutions, may help make data sharing more common and effective (Houtkoop et al., 2018).

First, authors wanting to share their data and code could take further steps to ensure eligibility for an Open Data badge. It might be argued that the average psychology researcher lacks the necessary technical skills. Any journal offering open science badges could support its authors in making their data and code reproducible and usable by providing guidance on (a) documentation of data, code, and the online repository; (b) sharing the rawest possible data (within ethical and logistical limits) alongside the cleaned data; and (c) guidance on recommendations for avoiding dependency and version issues (e.g., by using a platform such as Docker or Code Ocean; Clyburne-Sherin et al., 2019; Nüst et al., 2020; or if working in R by using, e.g., *groundhog* or *renv*; Simonsohn & Gruson, 2022; Ushey, 2022). There are many resources for making a reproducible workflow accessible, particularly concerning data and code sharing (see above). Authors can also ensure machine-actionable reusability of their data by following the findable, accessible, interoperable, and reusable (FAIR) guidelines (Wilkinson et al., 2016). It is commendable when authors attempt to share their data—data and code imperfectly shared are typically better than data and code perfectly kept to oneself. Indeed, our study would have been impossible without the introduction of the Open Data badge. The badge is a step in the right direction, but the corresponding policy needs to be improved to better support and incentivize transparent and reproducible research.

Second, there are improvements that could be made by badge-awarding journals that require both data and code for Open Data badge eligibility. If such journals rely on the disclosure method over the peer-review method, they could better describe the specific badge criteria and clarify that code, syntax, or a detailed analysis description needs to be shared alongside the data—for example, as required by the submission guidelines at *Psychological Science*. Many journals, and the baseline open science badge guidelines (Blohowiak et al., 2022), do not explicitly include the sharing of analysis code as an eligibility criterion; whether they should do so depends on the purpose of the Open Data badge. If the purpose is data reusability, not sharing code may be acceptable. If the purpose includes reproducibility,

however, code should always be included. This particularly applies to complex analyses, as verbal descriptions are unlikely to cover the information necessary for exact or essential reproduction (as demonstrated by our difficulties reproducing Article 112; see also Seibold et al., 2021). In simpler cases, not sharing code might seem acceptable (e.g., we essentially reproduced Article 111), but verbal reports can still fail, and sharing of analysis code ensures that all relevant information is available. By requiring the sharing of analysis code, *Psychological Science* is going beyond the basic requirements of the Open Data badge in order to achieve both reusability and reproducibility. Nevertheless, we still found that insufficient code was in fact shared for more than half of the examined articles. Badge-awarding journals requiring not only data but also code could more explicitly require authors to provide working code—where necessary—that enables straightforward reproducibility and produces clearly annotated output (see Bauer, 2022, for a reaffirmation of this requirement).

Third, it may be sensible to focus on other methods of awarding the open science badges. Given our results, as well as those of Hardwicke et al. (2021), a badge check may be needed as part of peer review at badge-awarding journals, including *Psychological Science*. This provides earlier verification and allows authors to upload all materials before publication and award of the badges. One way of doing this is to move to the peer-review method of awarding the Open Data badge (as opposed to the disclosure method; Blohowiak et al., 2022). The standard required by the peer-review method is open to interpretation by the specific journal: For the Open Data badge, this could range from a formal but brief review of the materials to independent reproduction of the reported results.³ The expected standard should match up with the standard stated in the submission guidelines; in the case of *Psychological Science*, data and code are already nominally required to enable precise or exact reproducibility, at least at the time of submission (Psychological Science, 2022). This work could be done by peer reviewers, dedicated badge reviewers, editors, or dedicated editorial staff (Blohowiak et al., 2022) and should be as straightforward as running the code or scripts on the data and requiring corrections if this does not lead to an exact reproduction. A checkbox could be provided for reviewers or dedicated badge reviewers to confirm that they executed the code successfully. If the analysis methods are complex or time consuming, then it should be incumbent on the authors to provide appropriate tools and assistance to the reviewers. If this responsibility is made clear to researchers before submission, this can incentivize more straightforwardly reproducible research. Alternatively, authors could provide proof of

a successful reproduction attempt, either independently or from within the research team (which would be an improvement, as analyses are commonly carried out by single team members; Veldkamp et al., 2014).⁴ This could be a condition for the award of the badge, or for an alternative Open Data+ badge, similar to the existing Preregistered+ badge (Blohowiak et al., 2022). Another approach would be to break the badge down into checkboxes of what was shared (e.g., raw and/or processed data, full or partial analysis code), thereby both lowering the threshold for participation and increasing transparency and usefulness of the badge.⁵ Regardless, whether authors fill in their disclosure items appropriately should continue to be monitored—a recent study found low adherence even to mandatory data availability statements in biomedical research manuscripts (Gabelica et al., 2022).

Limitations

The focus of our study was limited to the April 2019 issue of *Psychological Science*, a nonrandom sample of all articles in *Psychological Science* that received an Open Data badge. An advantage of this approach was that we could investigate each article in more depth than would be feasible for a larger sample, resulting in 46 individual reports in total, at least three per article. In comparison, Hardwicke et al. (2021) focused only on the numerical results of a subset of substantive findings for each article, meaning that reproducibility was not as fully evaluated as in our study. Our rich qualitative and quantitative results can be a starting point for further investigation. Building on our reproduction experiences may allow us to better anticipate the roadblocks that reproducers will face.

A possible limitation of our focus is that data-sharing practices may have improved overall since the publication of the issue under investigation. However, our results show only a slight improvement over those found by Hardwicke et al. (2021), who looked at articles published between 2014 and 2015 (using their less strict definition of reproducibility, equivalent to our “essential” reproduction). The Open Data badge eligibility criteria have not substantially changed since, so there is no reason to believe that a more current issue would show substantial improvement in a shorter time frame. Specifically, the eligibility criteria for the award of an Open Data badge at *Psychological Science* have included sharing of the relevant analysis code since at least November 2017 (Psychological Science, 2017).

Where reproducers had to recreate all or part of the analyses, our reproduction attempts may not be correct. This can result from unclear reporting or a lack of code (or other issues, identified above) but also from

a reproducer's expertise and evolving abilities as a researcher. However, we believe that competent graduate students should be able to reproduce the results of an article with an Open Data badge in their field of training. For an article that was awarded the Open Data badge at *Psychological Science*, reproduction should simply be a matter of running the code on the data.

An advantage of publicly shared data—over data unshared or available “on request”—is that they are available, and ideally useful, without the original authors' involvement. Contacting authors is not always easy: Researchers change institutions or email addresses and are mortal. Sometimes authors refuse to share data, even if required by the journal. Stodden et al. (2018) assessed the effectiveness of a policy of mandatory sharing on request at the journal *Science* and found that, despite this policy, they received data for only 44% of articles. Hence, the independence of the reproduction attempts in our study is one of its strengths. Doubtless we could have exactly or essentially reproduced more articles by contacting the original authors. We did not do this, as we wanted to investigate the effectiveness of the specific Open Data badge policy at *Psychological Science*, not the analytic or computational reproducibility of individual studies. The out-of-the-box reproducibility of each article indicates that effectiveness—if a successful reproduction requires contacting the authors, the badge was unsuccessful.

Conclusion

Recent advances in open and reproducible science have been rapid, and associated journal policies are constantly improving (see *Psychological Science's* move to Transparency and Openness Promotion [TOP] guidelines Level 2; Bauer, 2022). The stopgap, however, cannot be to award Open Data badges to articles that do not meet the minimum criteria. This study provides insight into the importance of sharing data for reproducibility and reuse as well as into the experience of reproducing studies that received the Open Data badge. We hope it can motivate improvements of the Open Data badge policy, or its implementation by the authors, at *Psychological Science* and other journals committed to promoting open science.

Transparency

Action Editor: Patricia J. Bauer

Editor: Patricia J. Bauer

Author Contributions

Sophia Crüwell: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Deborah Apthorp: Data curation; Formal analysis; Investigation; Writing – review & editing.

Bradley J. Baker: Data curation; Formal analysis; Investigation; Writing – review & editing.

Lincoln Colling: Formal analysis; Investigation; Writing – review & editing.

Malte Elson: Formal analysis; Investigation; Writing – review & editing.

Sandra J. Geiger: Data curation; Formal analysis; Investigation; Writing – review & editing.

Sebastian Lobentanzer: Formal analysis; Investigation; Writing – review & editing.

Jean Monéger: Data curation; Formal analysis; Investigation; Writing – review & editing.

Alex Patterson: Data curation; Formal analysis; Investigation; Validation; Writing – review & editing.

D. Samuel Schwarzkopf: Formal analysis; Investigation; Writing – review & editing.

Mirela Zaneva: Data curation; Formal analysis; Investigation; Writing – review & editing.

Nicholas J. L. Brown: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported from Economic and Social Research Council (Data Analytics & Society Centre for Doctoral Train) and Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (no. 1706dgn006).



ORCID iDs

Sophia Crüwell <https://orcid.org/0000-0003-4178-5820>
 Deborah Apthorp <https://orcid.org/0000-0001-5785-024X>
 Bradley J. Baker <https://orcid.org/0000-0002-1697-4198>
 Lincoln Colling <https://orcid.org/0000-0002-3572-7758>
 Malte Elson <https://orcid.org/0000-0001-7806-9583>
 Sandra J. Geiger <https://orcid.org/0000-0002-3262-5609>
 Jean Monéger <https://orcid.org/0000-0003-1178-1896>
 Alex Patterson <https://orcid.org/0000-0002-7780-4192>
 Mirela Zaneva <https://orcid.org/0000-0003-3569-931X>
 Nicholas J. L. Brown <https://orcid.org/0000-0003-1579-0730>

Notes

1. We carried out our study on the basis of the clear requirements stated in the submission guidelines.
2. According to the arguably ambiguous requirements as phrased in the Open Practices Disclosure (but not according to the explicit requirements in the submission guidelines), Article 111 may be considered eligible for an Open Data badge even without code. However, although the analyses were straightforward and *mostly* described, they were not sufficiently described to enable *exact* independent reproduction by our team.

3. Note that although these standards are compatible with the Transparency and Openness Promotion (TOP) guideline levels, they are not necessarily equivalent. Thus, a journal operating on TOP Levels 0 to 2 may introduce a version of the peer-review method of awarding Open Data badges without thereby moving to TOP Level 3. Adopting TOP Level 3 is more restrictive than moving only to the peer-review badge-awarding method.
4. We thank Daniel Simons and an anonymous reviewer for this suggestion.
5. We thank a further anonymous reviewer for this suggestion.

References

- Arslan, R. C. (2019). How to automatically document data with the *codebook* package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science*, 2(2), 169–187. <https://doi.org/10.1177/2515245919838783>
- Bae, G., & Luck, S. J. (2019). Reactivation of previous experiences in a working memory task. *Psychological Science*, 30(4), 587–595. <https://doi.org/10.1177/0956797619830398>
- Bauer, P. J. (2022). Psychological science stepping up a level. *Psychological Science*, 33(2), 179–183. <https://doi.org/10.1177/09567976221078527>
- Blohwiak, B. B., Cohoon, J., de-Wit, L., Eich, E., Farach, F. J., Hasselman, F., Holcombe, A. O., Humphreys, M., Lewis, M., Nosek, B. A., Peirce, J., Spies, J. R., Seto, C., Bowman, S., Green, D., Nilsson, G., Grahe, J., Wykstra, S., Hofelich Mohr, A., . . . Lowrey, O. (2022, February 4). *Badges to acknowledge open practices*. OSF. <https://osf.io/tvyxz>
- Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational reproducibility via containers in psychology. *Meta-Psychology*, 3, Article MP.2018.892. <https://doi.org/10.15626/MP.2018.892>
- Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, 15(4), Article e0230416. <https://doi.org/10.1371/journal.pone.0230416>
- Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal inference about good and bad outcomes. *Psychological Science*, 30(4), 516–525. <https://doi.org/10.1177/0956797619828724>
- Eberle, J. W. (2022). *Improving the computational reproducibility of clinical science: Tools for open data and code*. PsyArXiv. <https://doi.org/10.31234/osf.io/bf28t>
- Eich, E. (2014). Business not as usual. *Psychological Science*, 25(1), 3–6. <https://doi.org/10.1177/0956797613512465>
- Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, 150, 33–41. <https://doi.org/10.1016/j.jclinepi.2022.05.019>
- Garcia, D., & Rimé, B. (2019). Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, 30(4), 617–628. <https://doi.org/10.1177/0956797619831964>
- Geniole, S. N., Procyshyn, T. L., Marley, N., Ortiz, T. L., Bird, B. M., Marcellus, A. L., Welker, K. M., Bonin, P. L., Goldfarb, B., Watson, N. V., & Carré, J. M. (2019). Using a psychopharmacogenetic approach to identify the pathways through which—and the people for whom—testosterone promotes aggression. *Psychological Science*, 30(4), 481–494. <https://doi.org/10.1177/0956797619826970>
- Hakim, N., Adam, K. C. S., Gunseli, E., Awh, E., & Vogel, E. K. (2019). Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. *Psychological Science*, 30(4), 526–540. <https://doi.org/10.1177/0956797619830384>
- Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., & Frank, M. C. (2021). Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: An observational study. *Royal Society Open Science*, 8(1), Article 201494. <https://doi.org/10.1098/rsos.201494>
- Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L., & Bartholow, B. D. (2019). Null effects of game violence, game difficulty, and 2D:4D digit ratio on aggressive behavior. *Psychological Science*, 30(4), 606–616. <https://doi.org/10.1177/0956797619829688>
- Hinsen, K. (2019). Dealing with software collapse. *Computing in Science & Engineering*, 21(3), 104–108. <https://doi.org/10.1109/MCSE.2019.2900945>
- Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, 1(1), 70–85. <https://doi.org/10.1177/2515245917751886>
- Johnson, D. J., & Wilson, J. P. (2019). Racial bias in perceptions of size and strength: The impact of stereotypes and group differences. *Psychological Science*, 30(4), 553–562. <https://doi.org/10.1177/0956797619827529>
- 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., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14(5), Article e1002456. <https://doi.org/10.1371/journal.pbio.1002456>
- Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., IJzerman, H., Nilsson, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, 4(1), Article 20. <https://doi.org/10.1525/collabra.158>
- Levenstein, M. C., & Lyle, J. A. (2018). Data: Sharing is caring. *Advances in Methods and Practices in Psychological Science*, 1(1), 95–103. <https://doi.org/10.1177/2515245918758319>
- Lindsay, L., Gambi, C., & Rabagliati, H. (2019). Preschoolers optimize the timing of their conversational turns through flexible coordination of language comprehension and production. *Psychological Science*, 30(4), 504–515. <https://doi.org/10.1177/0956797618822802>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. <https://doi.org/10.17226/25303>
- Nüst, D., Sochat, V., Marwick, B., Eglén, S. J., Head, T., Hirst, T., & Evans, B. D. (2020). Ten simple rules for writing Dockerfiles for reproducible data science. *PLOS*

- Computational Biology*, 16(11), Article e1008316. <https://doi.org/10.1371/journal.pcbi.1008316>
- Obaidi, M., Bergh, R., Akrami, N., & Anjum, G. (2019). Group-based relative deprivation explains endorsement of extremism among Western-born Muslims. *Psychological Science*, 30(4), 596–605. <https://doi.org/10.1177/0956797619834879>
- 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>
- Olsson-Collentine, A., van Assen, M. A. L. M., & Hartgerink, C. H. J. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, 30(4), 576–586. <https://doi.org/10.1177/0956797619830326>
- Peikert, A., & Brandmaier, A. M. (2021). A reproducible data analysis workflow with R Markdown, Git, Make, and Docker. *Quantitative and Computational Methods in Behavioral Sciences*, 1(1), Article e3763. <https://doi.org/10.5964/qcmb.3763>
- Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, Article e175. <https://doi.org/10.7717/peerj.175>
- Psychological Science. (2017, November 15). *Submission guidelines*. https://web.archive.org/web/20171115110444/https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#OPEN
- Psychological Science. (2022, April 15). *Psychological Science submission guidelines*. https://www.psychologicalscience.org/publications/psychological_science/ps-submissions
- Rowhani-Farid, A., Aldcroft, A., & Barnett, A. G. (2020). Did awarding badges increase data sharing in *BMJ Open*? A randomized controlled trial. *Royal Society Open Science*, 7(3), Article 191818. <https://doi.org/10.1098/rsos.191818>
- Rowhani-Farid, A., & Barnett, A. G. (2018). Badges for sharing data and code at *Biostatistics*: An observational study. *F1000Research*, 7, Article 90. <https://doi.org/10.12688/f1000research.13477.2>
- Seibold, H., Czerny, S., Decke, S., Dieterle, R., Eder, T., Fohr, S., Hahn, N., Hartmann, R., Heindl, C., Kopper, P., Lepke, D., Loidl, V., Mandl, M., Musiol, S., Peter, J., Piehler, A., Rojas, E., Schmid, S., Schmidt, H., . . . Nalenz, M. (2021). A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. *PLOS ONE*, 16(6), Article e0251194. <https://doi.org/10.1371/journal.pone.0251194>
- Simonsohn, U., & Gruson, H. (2022). *groundhog: The simplest solution to version-control for CRAN packages*. <https://cran.r-project.org/package=groundhog>
- Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317), 1240–1241. <https://doi.org/10.1126/science.aah6168>
- Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences, USA*, 115(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>
- Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, 9(1), Article 60. <https://doi.org/10.1038/s41597-022-01143-6>
- Ushey, K. (2022). *renv* (Version 0.15.5) [Computer software]. GitHub. <https://rstudio.github.io/renv/>
- Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A. L., Peikert, A., Struiksma, M. E., & Vreede, B. M. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, 4(1), 29–49. <https://doi.org/10.3233/DS-210031>
- Vardy, T., & Atkinson, Q. D. (2019). Property damage and exposure to other people in distress differentially predict prosocial behavior after a natural disaster. *Psychological Science*, 30(4), 563–575. <https://doi.org/10.1177/0956797619826972>
- Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in *Psychological Science*. *PLOS ONE*, 9(12), Article e114876. <https://doi.org/10.1371/journal.pone.0114876>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. 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>
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6), Article e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>
- Wójcik, M. J., Nowicka, M. M., Bola, M., & Nowicka, A. (2019). Unconscious detection of one's own image. *Psychological Science*, 30(4), 471–480. <https://doi.org/10.1177/0956797618822971>
- Woolley, K., & Fishbach, A. (2019). Shared plates, shared minds: Consuming from a shared plate promotes cooperation. *Psychological Science*, 30(4), 541–552. <https://doi.org/10.1177/0956797619830633>
- Yousif, S. R., & Keil, F. C. (2019). The additive-area heuristic: An efficient but illusory means of visual area approximation. *Psychological Science*, 30(4), 495–503. <https://doi.org/10.1177/0956797619831617>


Publication 2: Citation Patterns Project



Hardwicke, T. E., Szűcs, D., Thibault, R. T., **Crüwell, S.**, van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021a). Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology. *Advances in Methods and Practices in Psychological Science*.

DOI: <https://doi.org/10.1177/25152459211040837>

Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology



Advances in Methods and Practices in Psychological Science
 July–September 2021, Vol. 4, No. 3,
 pp. 1–14
 © The Author(s) 2021
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/25152459211040837
www.psychologicalscience.org/AMPPS


Tom E. Hardwicke^{1,2} , Dénes Szűcs³, Robert T. Thibault^{4,5} ,
 Sophia Crüwell^{2,6} , Olmo R. van den Akker⁷, Michèle B. Nuijten⁷,
 and John P. A. Ioannidis^{2,8,9}

¹Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands; ²Meta-Research Innovation Center Berlin (METRIC-B), QUEST Center for Transforming Biomedical Research, Charité – Universitätsmedizin Berlin, Berlin, Germany; ³Department of Psychology, University of Cambridge, Cambridge, England; ⁴School of Psychological Science, University of Bristol, Bristol, England; ⁵MRC Integrative Epidemiology Unit at the University of Bristol, Bristol, England; ⁶Department of History and Philosophy of Science, University of Cambridge, Cambridge, England; ⁷Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands; ⁸Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA; and ⁹Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, California, USA

Abstract

Replication studies that contradict prior findings may facilitate scientific self-correction by triggering a reappraisal of the original studies; however, the research community's response to replication results has not been studied systematically. One approach for gauging responses to replication results is to examine how they affect citations to original studies. In this study, we explored postreplication citation patterns in the context of four prominent multilaboratory replication attempts published in the field of psychology that strongly contradicted and outweighed prior findings. Generally, we observed a small postreplication decline in the number of favorable citations and a small increase in unfavorable citations. This indicates only modest corrective effects and implies considerable perpetuation of belief in the original findings. Replication results that strongly contradict an original finding do not necessarily nullify its credibility; however, one might at least expect the replication results to be acknowledged and explicitly debated in subsequent literature. By contrast, we found substantial citation bias: The majority of articles citing the original studies neglected to cite relevant replication results. Of those articles that did cite the replication but continued to cite the original study favorably, approximately half offered an explicit defense of the original study. Our findings suggest that even replication results that strongly contradict original findings do not necessarily prompt a corrective response from the research community.

Keywords

replication, citations, meta-research, self-correction, citation bias, open data, open materials, preregistered

Received 2/9/21; Revision accepted 7/23/21

It is often assumed that science is a self-correcting enterprise: The veracity of scientific knowledge should progressively improve as inaccurate claims are abandoned and accurate claims are reinforced (Vazire & Holcombe, 2020). Replication studies are considered to be a key driver of this process because they may indicate that prior results are exaggerated or erroneous (Ioannidis, 2012; Zwaan et al., 2018). Although interpreting the outcome

of replication studies is not necessarily straightforward (Collins, 1985; Earp & Trafimow, 2015; Maxwell et al., 2015), one might expect a replication result that strongly

Corresponding Author:

Tom E. Hardwicke, Department of Psychology, University of Amsterdam
 Email: tom.hardwicke@uva.nl



Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Table 1. Progressive or Regressive Responses to Strongly Contradictory Replication Results and Their Expected Impact on Citation Patterns for Original Studies

Progressive responses	Regressive responses
<p>Belief correction A decrease in the number of favorable citations may reflect a decline in belief in the credibility of the original finding (a <i>belief correction pattern</i>). This may be accompanied by a relative increase in the number of unfavorable citations (an <i>active belief correction pattern</i>) or relatively fewer/no unfavorable citations (a <i>passive belief correction pattern</i>).</p>	<p>Belief perpetuation A maintenance of or increase in the number of favorable citations may reflect a maintenance of or increase in belief in the credibility of the original finding (a <i>belief perpetuation pattern</i>). This may be accompanied by a relative increase in unfavorable citations (a <i>challenged belief perpetuation pattern</i>) or relatively fewer/no unfavorable citations (an <i>unchallenged belief perpetuation pattern</i>).</p>
<p>Citation balance Articles that cite the original study may also cite the contradictory replication, reflecting that relevant evidence has been considered (a <i>balanced citation pattern</i>).</p>	<p>Citation bias Articles that cite the original study may neglect to also cite the contradictory replication, reflecting that relevant evidence has been neglected either through lack of awareness or deliberate omission (a <i>citation bias pattern</i>).</p>
<p>Explicit defense Articles that favorably cite the original study and unfavorably cite the contradictory replication may offer concrete counterarguments that state why the credibility of the original finding has not been undermined (an <i>explicit defense pattern</i>).</p>	<p>Absent defense Articles that favorably cite the original study and unfavorably cite the contradictory replication may offer no concrete counterarguments that state why the credibility of the original finding has not been undermined (an <i>absent defense pattern</i>).</p>

contradicts¹ and outweighs the results of a prior (“original”) study to affect how that study is cited in subsequent academic literature. For example, if a replication undermines belief in the credibility of an original finding, one might expect to see a change in the frequency and valence (favorability) of citations to the original study, that is, a decrease in favorable citations accompanied by an increase in unfavorable citations. However, as discussed below, a variety of interesting patterns could emerge depending on how the research community responds to a replication result. The goal of the present study was to empirically explore and describe postreplication citation patterns in the context of four prominent multilaboratory replication attempts published in the field of psychology that strongly contradicted and outweighed the findings of prior studies.

Table 1 outlines several citation patterns that might follow a contradictory replication result, each reflecting different types of response by the research community. We have tentatively categorized these response patterns as being “progressive” or “regressive,” depending on their expected impact on the accumulation of scientific knowledge.² The first set of patterns, *belief correction/perpetuation*, refers to what is often considered a primary functional role of (contradictory) replication studies—to change belief in the credibility of exaggerated or erroneous original findings (Ioannidis, 2012; Vazire & Holcombe, 2020; Zwaan et al., 2018). In the absence of an explicit defense of an original study that convincingly explains a strongly contradictory replication result (see explicit

defense explanation below), a progressive response might involve a decrease in favorable citations and an increase in unfavorable citations, reflecting updated beliefs about the credibility of the original finding. Conversely, a regressive response might involve maintenance of (or even increase in) favorable citations and relatively few unfavorable citations, which suggests a perpetuation of belief in the credibility of the original finding despite the contradictory replication result. Prior research has documented how favorable citations to observational epidemiology studies can persist despite the claims of those studies being strongly contradicted in subsequent randomized trials (Tatsioni et al., 2007). Likewise, it has been reported that even when articles are retracted, they can continue to receive favorable citations (Budd et al., 1998; Fernández & Vadillo, 2020). Thus, there is evidence that belief in the credibility of original findings can perpetuate even when subsequent events cast doubt on their credibility; however, we are unaware of similar evidence in the context of studies that were explicitly designed to test the replicability of prior findings.

The second set of patterns in Table 1, *citation balance/bias*, generally refers to whether positive (supportive) evidence is preferentially cited relative to negative (nonsupportive) evidence (Bastiaansen et al., 2015; Greenberg, 2009). This pattern has previously been observed in the context of research on inclusion body myositis; citation content analysis showed that the accumulating literature heavily cited the theory that beta amyloid is involved, ignoring multiple studies that

Table 2. Sample Sizes and Effect Sizes for Replication Studies and Original Studies

Original study	Replication study	Effect	Total citations (original) ^a	Original sample size	Replication sample size	Original effect size [95% CI]	Replication effect size [95% CI]
Baumeister et al. (May, 1998)	Hagger et al. (July, 2016) ^b	Ego depletion	1,974	$k = 1$ $N = 67$	$k = 23$ $N = 2,141$	$d = 2.05$ [1.31, 2.79]	$d = 0.04$ [-0.07, 0.15]
Sripada et al. (April, 2014)	Hagger et al. (July, 2016) ^b	Ego depletion	36	$k = 1$ $N = 26$	$k = 23$ $N = 2,141$	$d = 0.68$ [0.09, 1.27]	$d = 0.04$ [-0.07, 0.15]
Strack et al. (May, 1988)	Wagenmakers et al. (October, 2016)	Facial feedback	708	$k = 1$ $N = 92$	$k = 17$ $N = 2,124$	MD = 0.82 [-0.05, 1.69]	MD = 0.03 [-0.11, 0.16]
Caruso et al. (July, 2012)	Klein et al. (January, 2014)	Money priming	57	$k = 1$ $N = 30$	$k = 36$ $N = 6,333$	$d = 0.8$ [0.05, 1.54]	$d = 0.01$ [-0.06, 0.09]
T. J. Carter et al. (July, 2011)	Klein et al. (January, 2014)	Flag priming	54	$k = 1$ $N = 70$	$k = 36$ $N = 4,896$	$d = 0.50$ [0.01, 0.99]	$d = 0.01$ [-0.07, 0.08]

Note: Publication dates are earliest available (i.e., “online first” if relevant). d = Cohen’s d ; MD = mean difference; k = number of data-collection sites; N = total number of participants; CI = confidence interval.

^aTotal citations to the original study between the publication date and December 31, 2019.

^bFor methodological reasons (see Hagger et al., 2016), the ego-depletion replication was aimed at a classic study in the field (Baumeister et al., 1998) but actually employed a modified computer-based version of the original paradigm (Sripada et al., 2014). We examined postreplication citation patterns for both studies.

contradicted this theory (Greenberg, 2009). In the present study, these patterns specifically refer to whether articles citing an original study also cite the subsequent contradictory replication study. A progressive response would be to cite both studies (citation balance) because this involves considering and reporting highly relevant evidence (even if the implications of the replication are disputed; see explicit defense explanation below). By contrast, citation bias could occur if articles citing an original study neglect to cite a relevant replication study. Regardless of whether this occurs through lack of awareness or deliberate omission, it can be considered a regressive response pattern because highly relevant evidence is not being reported or considered.

The third set of patterns in Table 1, *explicit/absent defense*, refers to whether researchers who continue to favorably cite the original finding despite the strongly contradictory replication result offer a concrete defense of the original study. As implied above, even when the results of a replication study strongly contradict the results of an original study, this does not necessarily nullify the credibility of the original findings; the same criticisms that one might apply to an original study to infer that its findings are erroneous or exaggerated may also be applied to replication studies (Collins, 1985; Earp & Trafimow, 2015; Maxwell et al., 2015). Thus, if proponents of the original claim mount an explicit defense that counters the implications of the replication results, this might still be considered a progressive response (although obviously one could disagree with the arguments that are presented). By contrast, if favorable citations to the

original study are not accompanied by explicit argumentation about the replication result (an absent defense), this might be considered a regressive response because the replication result is apparently discounted without providing any rationale.

In the present study, we explored the postreplication citation patterns described above in the context of four case studies in the field of psychology in which the findings of a replication study strongly contradicted and outweighed the findings of an original study (Table 2). In two of the cases, the replication studies were part of a single Many Labs project (Klein et al., 2014) and addressed the “flag priming effect” (T. J. Carter et al., 2011) and “money priming effect” (Caruso et al., 2013), respectively. The other two cases involved Registered Replication Reports (Simons et al., 2014) that examined influential demonstrations of the “facial feedback effect” (cf. Strack et al., 1988; Wagenmakers et al., 2016) and the “ego-depletion effect” (cf. Baumeister et al., 1998; Hagger et al., 2016; Sripada et al., 2014), respectively. For methodological reasons (see Hagger et al., 2016), the ego-depletion replication was aimed at a classic study in the field (Baumeister et al., 1998) but actually employed a modified computer-based version of the original paradigm (Sripada et al., 2014). For this particular case study, we examined citation patterns to both of these original studies.

We adopted a case-study approach to develop a “narrow and deep” understanding of the topic, as opposed to a “broad and shallow” approach, which would have required a deliberate representative sampling strategy. We chose these particular case studies because they involved

prominent preregistered multilaboratory replication attempts with sample sizes 23 times to 211 times larger than the original studies, thus providing highly visible and highly credible evidence that strongly contradicted and outweighed earlier findings.³ This facilitates additional interpretative clarity about the citation patterns one might expect to observe.

The present study was exploratory in nature and intended to provide descriptive observations rather than test hypotheses. The three sets of expected citation patterns outlined in Table 1 were used to guide our study design and interpretation, but we do not claim that this is a comprehensive typology of the postreplication patterns that may occur. Such patterns may be more complex and idiosyncratic in other topic domains. To examine belief correction/perpetuation patterns, we downloaded citation histories (a list of citing articles) for each original study and classified the valence (favorable, equivocal, or unfavorable) of a set of prereplication and postreplication citations. To examine citation balance/bias patterns, we manually checked whether postreplication citations of the original study were accompanied by citations to the replication study. Finally, to examine explicit/absent defense patterns, we extracted and categorized any counterarguments offered in articles that cited the replication.

Method

The study protocol (rationale, methods, and analysis plan) was preregistered on April 7, 2018 (<https://osf.io/eh5qd/>). An amended protocol was registered partway through data collection on May 1, 2019, primarily because we extended the sampling frame to cover additional months (<https://osf.io/pdvb5/>). All deviations from these protocols are explicitly acknowledged in Supplementary Information A in the Supplemental Material available online. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Design

This was a retrospective observational study consisting of four case studies. Primary outcome variables were annual citation counts for original studies, citation valence (favorable, equivocal, unfavorable), co-citation of original and replication studies, and frequency/type of counterarguments.

Sample

We examined four case studies in which a prominent preregistered and multilaboratory replication study strongly contradicted and outweighed the findings of an

original study (Table 2). As shown, all the original studies found modestly large to very large effects, and all of them were relatively small studies (thus they would have been underpowered to detect small effects). Conversely, all the replication efforts comprised very large sample sizes, and they would be very well powered to detect even small effects; however, they all obtained null results.

Procedure

Annual citation counts. Citation histories (i.e., bibliographic records for all articles that cite the original study) from the publication date of each original study through December 31, 2019, were downloaded from Clarivate Analytics Web of Science Core Collection, accessed via the Charité – Universitätsmedizin Berlin on August 12, 2020. We also obtained citation histories for a reference class—all articles published in the same journal and the same year as each original study—from the same source. For example, for Baumeister et al. (1998), the reference class was all articles published in 1998 in the *Journal of Personality and Social Psychology*. Citation counts were standardized in each case study by setting the citation count in the replication year to the standardized value of 100 and then adjusting the counts in other years according to the same transformation ratio. For example, if the raw citation count in the replication year was 1,000, citation counts in each year would be standardized by dividing by 10. This computation was performed separately for citations to the reference class and citations to the original article.

Qualitative assessment. Qualitative assessment of citation patterns was limited to a time period starting 1 year before the year of publication of the replication study until December 31, 2019, excluding the year in which the replication was published. We excluded the replication year because it may be unreasonable to expect citing articles already in the publication pipeline to cite the replication study. For the Baumeister case, the qualitative analysis was based on a random sample of 40% of citing articles from the prereplication period and postreplication period because of the large number of citations to the original study ($n = 1,974$; for details, see Supplementary Information B in the Supplemental Material).

For each citing article undergoing qualitative assessment, we attempted to retrieve the full text via several methods in the following order: (a) search of at least two of the institutional libraries we are affiliated with; (b) general Internet search for the article title, including the Google and Google Scholar search engines and Research Gate; (c) email requests to the corresponding author; and (d) interlibrary loan request. Articles that remained inaccessible after all of these methods were

exhausted were excluded. Articles written in a non-English language were translated by one of the authors or by using Google Translate (see Supplementary Information D in the Supplemental Material). For articles for which we could obtain the full text, we classified the research design according to the categories in Table 3 and recorded whether the replication study was cited after manual inspection of the reference section (see Table 1: citation balance/bias).

To examine the belief correction/perpetuation pattern (Table 1), one of six primary coders (T. E. Hardwicke, D. Szűcs, R. T. Thibault, S. Crüwell, O. R. van den Akker, and M. B. Nuijten) manually extracted the “citation context” of the original study and the replication study (i.e., all relevant verbatim text surrounding each in-text citation). The primary coder then classified the citation valence as “favorable,” “equivocal,” “unfavorable,” or “unclassifiable.” Favorable citations were those used to support a positive claim about the phenomenon of interest, whereas unfavorable citations were used to support a negative claim about the phenomenon of interest. Citations were considered equivocal if the authors did not take a predominantly favorable or unfavorable position. Citations that did not endorse or oppose the phenomenon of interest (e.g., simply referring to the procedures of the original study) were designated as unclassifiable. Because this process was inherently subjective, the citation contexts and classifications were also examined by one of six secondary coders (T. E. Hardwicke, D. Szűcs, R. T. Thibault, S. Crüwell, O. R. van den Akker, and M. B. Nuijten). Disagreements were resolved through discussion, and a third coder arbitrated when necessary. Valence classifications by the primary coder were modified after discussion with the secondary coder in 31 (5%) cases.

To examine the explicit/absent defense pattern (Table 1), the primary coder flagged articles that co-cited the original and replication studies and also contained any explicit defense of the original study. Subsequently, two team members (O. R. van den Akker and S. Crüwell) reexamined all of the flagged cases, extracted verbatim counterarguments, and developed a post hoc categorization scheme that summarized them as concisely and informatively as possible. Coding disagreements were resolved through discussion, and a third coder (T. E. Hardwicke) arbitrated when necessary.

In additional exploratory (not preregistered) analyses, we examined overlap of authorship for articles that provided counterarguments with (a) any of the authors of the original studies and (b) any prior collaborators of the first authors of the original studies. These analyses are complicated by the fact that author names in bibliographic records do not always adhere to the same grammatical standards—for example, whether forenames are initialized or middle names are included—so

Table 3. Counts and Percentages for Article Type Classifications of Articles Included in Qualitative Analyses

Article type	Count (%)
No empirical data (e.g., editorials, commentaries [without reanalysis], simulations, news, and reviews)	197 (33)
Data synthesis - meta-analysis	11 (2)
Empirical data - commentary including analysis	4 (1)
Empirical data - case study	1 (< 1)
Empirical data - survey	79 (13)
Empirical data - field study	40 (7)
Empirical data - laboratory study	248 (41)
Empirical data - multiple study types are reported	24 (4)

it is not straightforward to isolate individual authors within bibliographic databases. To identify prior collaborators of the first authors of the original studies, we downloaded bibliographic records (on February 2, 2021) for all articles published by each of the original study first authors according to their author record in the Web of Science Core Collection. These author records are automatically generated by an algorithm that attempts to identify all documents likely published by an individual author using several variations of their name (e.g., “Hardwicke, Tom E.,” “Hardwicke, Tom,” “Hardwicke, T. E.”), but errors can still occur, and incomplete database coverage means that this method likely misses some of the authors’ prior publications and consequently some of their collaborators. Nevertheless, the method supports a reasonable lower bound estimate of authorship overlap with articles providing counterarguments. To identify authorship overlap, we used string manipulation tools in R to extract only author surnames from bibliographic records and then used string matching to automatically detect the presence of original author or collaborator surnames among the surnames of authors of articles that provided counterarguments. When a match was detected, it was verified by manual examination of the authors’ full names.

Results

In total, 2,829 articles cited one of the original studies, of which 632 articles (after taking a 40% random sample in the Baumeister case) fell within the time period designated for qualitative assessment. Of these 632 articles, we excluded 28 from the qualitative analysis because (a) we could not access the full text ($n = 22$), (b) they included a citation to the original study in the reference section but not in the main text ($n = 5$), or (c) manual inspection indicated that they did not actually appear to cite the original study at all ($n = 1$). Article type classifications for the remaining 604 articles included in the qualitative analysis are shown in Table 3.

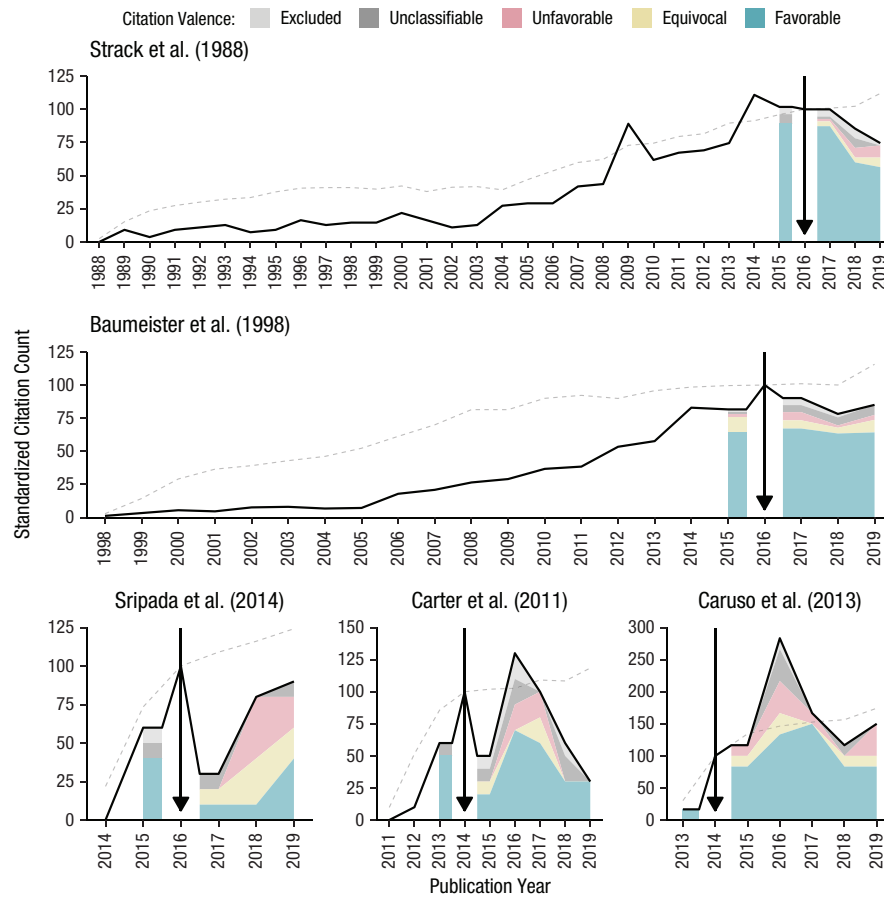


Fig. 1. Standardized annual citation counts (solid line) for the five original studies with citation valence (favorable, equivocal, unfavorable, unclassifiable) illustrated by colored areas in prereplication and postreplication assessment periods. The dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Citation valence classifications for the Baumeister case are extrapolated to all articles in the assessment period according to a 40% random sample.

Annual citation counts and citation valence

Figure 1 shows standardized annual citation counts for each original study and the respective reference class (citations to all articles published in the same year and same journal as the original study) and classifications of citation valence (favorable, equivocal, unfavorable, unclassifiable, or excluded). The data can also be viewed in tabular format in Supplementary Table C1 in the

Supplemental Material. All counts (n) reported in the text and table are raw counts (i.e., not standardized).

After the replication was published, citations to the reference classes were continuing their trend to plateau (Baumeister case) or increase (other cases). By contrast, citations to the original study appeared to undergo a modest decline in the Strack case (decreasing from 56 to 41 between 2015 and 2019) and a small decline followed by a small increase in the Baumeister case (increasing from 191 to 199 between 2015 and 2019). In the other

cases (Sripada, Carter, Caruso), the total citation counts were much lower, and there was considerable variability in the postreplication citation patterns; nevertheless, there was no substantial change in annual citations from prereplication to postreplication in these three cases (the maximum difference was +8 citations).

Before the replication, the vast majority of citations were favorable for all five articles (range = 67%–100%). In most cases (Strack, Sripada, Carter, and Caruso), there was a small postreplication increase in unfavorable citations and a small decrease in favorable citations, indicating a modest active correction pattern. However, the overall number of unfavorable citations was very low, and there was still a substantial majority of favorable citations. For example, in the Strack case, unfavorable citations increased from 0% in the prereplication period (2015) to 7% in the postreplication period, whereas favorable citations decreased from 88% to 78%. In the Baumeister case, the proportion of favorable citations remained stable from prereplication (79%) to postreplication (77%), a pattern consistent with belief perpetuation. The very small number of unfavorable citations (2017: $n = 7$, 7%; 2018: $n = 2$, 2%; 2019: $n = 2$, 4%) suggests that this is largely an unchallenged belief perpetuation pattern (see Table 1).

Citation balance and citation bias

Figure 2 shows the proportion of citing articles that also cited or did not cite the replication study after it was published (excluding the publication year itself). The data can also be viewed in tabular format in Supplementary Table C1 in the Supplemental Material. In the Strack and Baumeister cases, a considerable majority of articles citing the original study did not cite the replication study, which indicates substantial citation bias. In the Baumeister case, the proportion of articles citing the replication study remained stable (20% in 2017, 18% in 2019). In the Strack case, the proportion increased from 13% to 41%. In the Carter and Caruso cases, the proportion never exceeded 50%, also consistent with substantial citation bias. In the Sripada case, it was much more common for the replication study to be cited (> 88%), which reflects a balanced citation pattern.

Explicit defense and absent defense

Table 4 shows whether articles that cited the original study and replication study (“co-citing articles”) and the subset of co-citing articles that cited the original study favorably provided any explicit counterarguments to defend the credibility of the original finding (an explicit defense) or not (an absent defense). Overall, fewer than half of the 127 co-citing articles provided any

counterarguments. Of the 60 co-citing articles that cited the original study favorably, around half provided counterarguments. We identified 58 discrete counterarguments in 51 citing articles (45 of which were unique articles because six of them were cited in two of the case studies) and allocated them to one of three categories (Table 5).

In additional exploratory analyses (not preregistered), we examined other characteristics of the 45 unique articles that contained counterarguments. The articles were published in 34 individual journals; *Frontiers in Psychology* published seven of the articles, *Social Psychology* published four of the articles, and all other journals published only one or two of the articles. Seventeen of the articles did not involve empirical data, three involved reanalysis or meta-analysis of existing data, and 25 involved collection of novel data. The articles had 112 individual authors, of whom all contributed to a single article except for nine individuals who had authored or coauthored two articles. Three articles were authored or coauthored by one of the original authors, and nine articles were authored or coauthored by at least one prior collaborator of one of the first authors of the original articles. Seven of these articles did not involve empirical data, and five of them involved novel data collection.

Citations to replication studies and co-citation of original studies

A reviewer requested that we examine citation counts for replication studies and check whether citing articles also co-cited the relevant original study. To obtain the data, we downloaded bibliographic records from the Clarivate Analytics Web of Science Core Collection, accessed via the University of Amsterdam on April 16, 2021, for articles that cited each replication study (up to the study endpoint—December 2019) and cross-checked them with our sample of articles that cited the original study. As shown in Table 6, the replication studies also have a life of their own, and they are often cited independently of the specific original study. Often this is easy to explain. For example, Klein et al. replicated 13 original studies, not just the two that were of interest in our analysis. These studies are also likely to have accrued citations by virtue of being among the first highly prominent examples of preregistered multilaboratory replications in psychology.

Discussion

It has been proposed that replication studies can facilitate scientific self-correction by modifying scientists’ belief in the credibility of published findings (Ioannidis, 2012; Vazire & Holcombe, 2020; Zwaan et al., 2018); however,

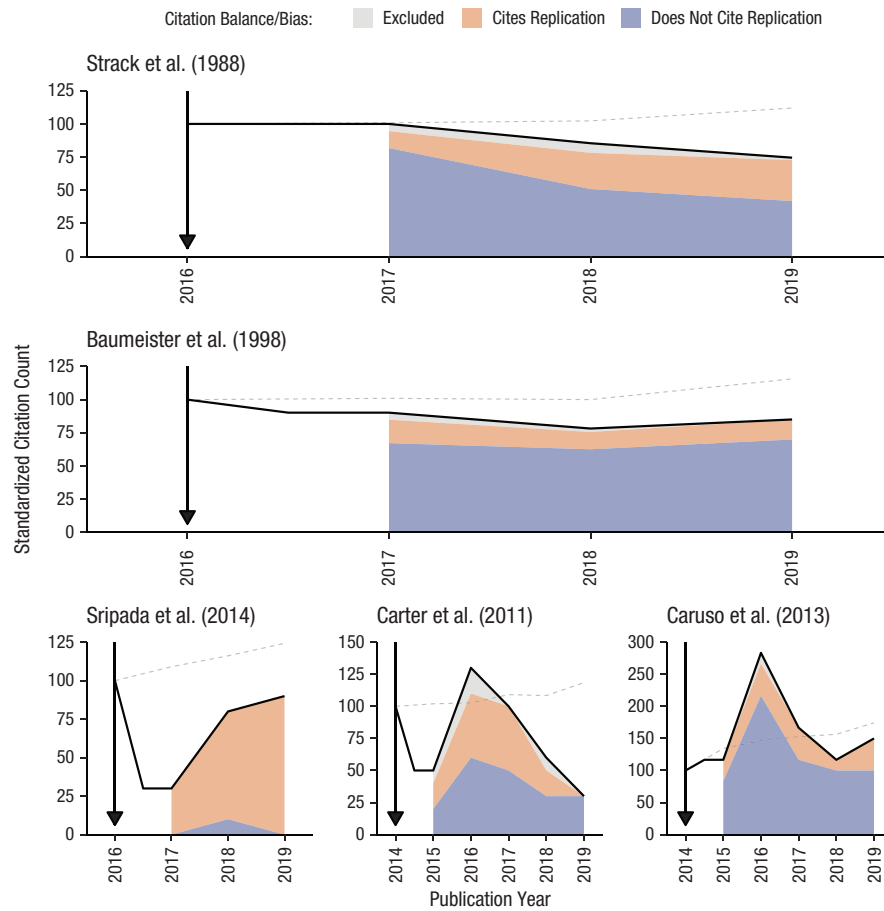


Fig. 2. Standardized annual citation counts (solid line) for the five original studies with citation balance/bias (i.e., whether the replication is cited) illustrated by colored areas in the postreplication assessment period. The dashed line depicts citations to the reference class (all articles published in the same journal and same year as the target article). Annual citation counts are standardized against the year in which the replication was published (citation counts in the replication year, indicated by a black arrow, are set at the standardized value of 100). Replication citation proportions for the Baumeister case are extrapolated to all articles in the assessment period according to a 40% random sample.

the extent to which this occurs in practice is unclear. In this study, we investigated how the research community responded to four strongly contradictory replications in the field of psychology by examining postreplication citation patterns for original studies. We observed some progressive response patterns in the form of modest active correction (a small decline in favorable citations and a small increase in unfavorable citations) and even more prominent regressive response patterns in the form of unchallenged belief perpetuation (sustained levels of favorable citations and few unfavorable citations,

particularly in the Baumeister case) and considerable citation bias (neglecting to cite the replication study; in all cases aside from Sripada). When authors cited the original study favorably despite the replication result, only half of the articles provided any explicit counterarguments in defense of the original study (an explicit defense). Overall, these findings are consistent with prior observations that favorable citation patterns appear relatively unperturbed by subsequent publication of contradictory results in studies with lower risk of bias (Tatsioni et al., 2007) or even by full retraction (Budd et al., 1998;

Fernández & Vadillo, 2020) and that positive (supportive) evidence is preferentially cited relative to negative (non-supportive) evidence once a theory gets entrenched despite overwhelming evidence against it (Bastiaansen et al., 2015; Greenberg, 2009).

It is reasonable to question whether the replication results in these case studies should (rationally) have instigated belief change in the research community and thus triggered a more sizable decline in favorable citations than we observed. Note that the case studies we selected were deliberately chosen because the replication results were superior to the original results in terms of both credibility and evidential value. The replications were preregistered, multilaboratory studies with large sample sizes. By contrast, the original studies had much smaller sample sizes and arguably had a much higher risk of bias given that they were not preregistered, were performed by single teams, and arose in domains affected by publication bias and other questionable research practices (E. C. Carter et al., 2015; Coles et al., 2019; Vadillo, 2019; Vadillo et al., 2016). Thus, it seems reasonable to suggest that the compelling replication results should have reduced belief in the credibility of the original results. However, one could also contend that if there were some flaw in the replication study that undermined its validity (Brandt et al., 2014; Fabrigar et al., 2020; Vazire et al., 2021), then this may provide justification to continue favorably citing the original study. Indeed, in these particular case studies, the validity of the replication results has been challenged by proponents of the original findings (Table 5; Baumeister & Vohs, 2016; Strack, 2016). Because this debate remains far from settled (Coles et al.,

Table 4. Counts and Percentages for Whether Articles That Cited Both the Original Study and Replication Study Provided Any Explicit Argumentation to Defend the Original Study

Case	Did co-citing articles provide argumentation to defend the original study?			
	All citation valences		Favorable citation valence	
	No	Yes	No	Yes
Baumeister	24 (56%)	19 (44%)	11 (46%)	13 (54%)
Carter	11 (79%)	3 (21%)	3 (75%)	1 (25%)
Caruso	8 (67%)	4 (33%)	1 (50%)	1 (50%)
Sripada	10 (53%)	9 (47%)	2 (40%)	3 (60%)
Strack	23 (59%)	16 (41%)	12 (48%)	13 (52%)
All cases	76 (60%)	51 (40%)	29 (48%)	31 (52%)

Note: Data are displayed for co-citing articles with any citation valence classification and the subset of co-citing articles with favorable citation valence classifications.

2019; Vadillo, 2019), ideally any favorable citation of the original studies should at a minimum be accompanied by co-citation of the replication results and some discussion of the discrepant findings.

The clear evidence of citation bias that our study documents may have two main contributory factors: (a) a lack of awareness about the replication results and/or (b) a decision to ignore the replication results. The practical issue of awareness is not necessarily straightforward to address. Individual scientists can find it difficult to keep up to date with the voluminous literature that is

Table 5. Categorization of Counterarguments Provided to Defend the Original Study in Light of the Contradictory Replication Result

Category	Definition and examples	Count
Methodological differences and moderators	Methodological features of the replication study or other moderating factors may explain the absence of an effect. Example: "Ego depletion exists but its occurrence seems to depend on moderating conditions. Therefore, we think that the search of moderators concerning ego depletion (or ego depletion as a moderator, respectively) is justified" (Kühl & Bertrams, 2019, p. 9).	45
Additional evidence	Evidence from other studies supports the existence of the effect. Example: "Although there has been a challenge to this original finding, there have been many replications of the principle and a meta-analysis shows a robust facial-feedback effect" (Lewis, 2018, p. 2).	11
Expertise	Inadequate expertise of the replicating authors explains why they could not replicate the effect. Example: "In other words, it is easier to be successful at non-replications while it takes expertise and diligence to generate a new result in a reliable fashion" (Strack, 2017, p. 3).	2

Note: Fifty-eight discrete counterarguments were identified in 51 articles (45 unique articles across cases).

Table 6. Citations Counts for Replication Studies and Co-Citation Counts to Relevant Original Studies

Replication study	Citation count	Co-citations of original study
Hagger et al. (2016)	258	136 (Baumeister et al., 1998) 22 (Sripada et al., 2014)
Klein et al. (2014)	316	15 (T. J. Carter et al., 2011) 12 (Caruso et al., 2013)
Wagenmakers et al. (2016)	80	36 (Strack et al., 1988)

relevant to their research. Recently, the reference manager Zotero introduced a feature that alerts users when an article in its database has been retracted (Zotero, 2019). One could imagine a similar feature being introduced for replication studies, perhaps based on databases that explicitly identify replication studies. However, it is much less straightforward to define and identify a relevant replication study (Neuliep & Crandall, 1993), and users would need to be alerted that this requires some scientific judgment rather than simple article metadata. Another solution could be to encourage researchers to focus less on individual studies and more on up-to-date evidence summaries (i.e., reviews and meta-analyses) in which relevant evidence is systematically identified and collated. This would require that high-quality and contemporary evidence summaries are available; however, in psychology, systematic reviews and meta-analyses can be of low quality, and their results may still be inflated and nonreproducible (Kvarven et al., 2020; Maassen et al., 2020; Polanin et al., 2020). Moreover, empirical studies are often not included in any form of evidence synthesis (Hardwicke et al., 2021).

The second issue of authors ignoring highly relevant replications seems undesirable and implies a biased appraisal or presentation of the evidence. Contradictory replication results do not necessarily nullify the credibility of an original study (Collins, 1985; Earp & Trafimow, 2015; Maxwell et al., 2015), but we would still expect highly relevant replication results to be cited and explicitly debated. In fact, we found that when authors continued to favorably cite original studies despite the replication results, around half did not provide explicit counterarguments in defense of the original study. Although our study did not examine researchers' individual beliefs, one recent study reported that when research psychologists were confronted with replication evidence, they often did update their (self-reported) beliefs, albeit modestly (McDiarmid et al., 2021). However, there are several reasons to be uncertain about whether the results from this artificial setting might generalize to real-world settings, including potential participant reactivity (i.e., participants behaving

differently because they are under observation and/or responding to the perceived expectations of the research team) and the possibility that individuals may behave differently in settings in which they have substantial personal investment and may be publicly scrutinized. In addition, cognitive psychology studies have obtained some evidence of a "continued influence effect" wherein an individual's beliefs and behavior can continue to be influenced by false or misleading information despite subsequent efforts to reject it (Lewandowsky et al., 2012). It is plausible that various cognitive biases, such as confirmation bias (preferentially seeking out and processing evidence that supports preexisting beliefs) or motivated reasoning (constructing and evaluating arguments according to what is desirable rather than what is rationally justifiable), may partly explain researchers' tendency to ignore or dismiss the replication evidence (Bishop, 2019; Kunda, 1990; Nickerson, 1998).

We observed that when counterarguments were raised, most of them tried to dismiss the contradictory replication by claiming that the original and the replication studies differed in important ways that moderated the absence/presence of the effect under investigation. In some cases, authors pointed to evidence from other studies as a rationale for their continued belief in the effect. In a minority of cases, the authors challenged the competence of replicators. We found that articles presenting counterarguments were published in a variety of journals (rather than clustered in a few journals) and involved collection of new data in around half of the cases. They were also published by a sizable group of investigators, only a minority of whom were one of the original authors or had previously collaborated with one of the original first authors. This suggests that the explicit defense of the original study came from a variety of sources rather than being confined to a small number of investigators. However, note that this analysis may underestimate authorship overlap because of the difficulties isolating individual researcher identities (see Method section) and articles published in the same year as the replication not being included.

The findings presented here are inherently limited by the observational nature of the study design, which complicates straightforward conclusions about the causal impact of the replications. Although the use of a reference class enables us to detect the influence of exogenous factors to some extent, we cannot rule out their contribution. For example, the modest decline in favorable citations observed in most cases could be attributable, at least in part, to a more general awareness in the research community about methodological issues (e.g., that the sample sizes of the original studies may not have provided adequate statistical power). We have also focused only on the replication study and the original study in each case study without considering the impact

of other potentially relevant events. Note that metaresearch studies have detected signatures of publication bias and other questionable research practices in the fields to which these case studies belong (E. C. Carter et al., 2015; Coles et al., 2019; Vadillo, 2019; Vadillo et al., 2016), and other relevant replication studies contesting prior findings have been published (e.g., Rohrer et al., 2015).

We have been able to gauge reactions to replications only to the extent that they are reflected in relatively short-term citation patterns. A potential explanation of the apparently cursory treatment of the replication studies could be that researchers became aware of them only after their own research projects had begun and/or even had been completed. If one of the original studies had been a key motivator for one's own study, then it may be difficult to accommodate the strongly contradictory replication results. It may even be tempting to ignore them or give them only superficial treatment. Moreover, researchers who are convinced that the replication study has squarely refuted the original may no longer be interested in doing research on a topic for which they see no future potential. In addition, examination of citation patterns would not detect whether there had been a correction effect among individuals who would not typically cite the original study, such as students, members of the public, or researchers working in other fields. A contested study may continue to be cited favorably by its proponents who remain working in the field. This will suffice to create belief perpetuation in the published literature even though other scientists may simply no longer be interested in getting involved with such a strongly contested research topic. Relatedly, we did not examine citation patterns beyond 3 to 5 years after replication. Some perspectives envision scientific self-correction unfolding over a much longer time scale (Lauden, 1981; Peterson & Panofsky, 2020) and suggest that the process is characterized less by the impact of individual study results and more by the gradual accumulation of converging evidence, gradual revision of theoretical understanding, and/or informal sociological processes (e.g., researchers choosing alternative topics to study). Thus, although the current findings may contradict the expectations of a more direct and expedited view of the corrective impact of replication studies (Ioannidis, 2012; Vazire & Holcombe, 2020), they are not necessarily inconsistent with a slower and more indirect process of self-correction. Future research could employ longitudinal designs or older historical case studies to evaluate citation patterns unfolding over a longer time scale. Finally, because we examined only citation patterns, this study could not capture other potential responses to replication results, such as changes in research practices.

Generalization beyond these four case studies requires caution. Note that the replication studies examined here

were some of the first large-scale, multilaboratory replication attempts conducted in the field of psychology. This gave them particular prominence and initiated considerable debate, which resulted in broader ramifications beyond the research community that typically studies the topics under scrutiny (Nelson et al., 2018). Also note that we deliberately selected case studies in which the replication studies were high-profile and had yielded high credibility evidence that strongly contradicted and outweighed the original findings. A correction effect may be less expected in cases in which replication results are more ambiguous, less consequential, or less well known. For example, in a situation in which two high-credibility studies with similar evidential value yield contradictory results, it would be premature to lose confidence in one of the studies before further investigation has probed the cause of the discrepancy. Pursuit of potential moderating factors may be entirely rational in such circumstances (Gershman, 2019).

We also note that particular aspects of our study were inherently subjective, specifically, the identification of citation context, the classification of citation valence, and the identification, extraction, and categorization of counterarguments. To minimize subjectivity, a team of six investigators performed coding in duplicate, with a third investigator arbitrating if necessary. Because disagreements between primary and secondary coders were infrequent, we are confident that the classifications are meaningful, but there may be some edge cases when an independent observer might reasonably disagree with our classifications.

In conclusion, postreplication citation patterns in four case studies indicated that the anticipated corrective impact of strongly contradictory replication results did not materialize to any substantive degree. A lack of awareness of replications and/or a decision to discount or omit them appears to have played a significant role. This highlights potential practical problems with the discoverability of replication studies and psychological or sociological issues related to belief change. The findings also indicate that scientific self-correction may not be as expedient or straightforward as one might hope (Ioannidis, 2012), which adds further impetus toward efforts to improve the quality of the academic literature (Hardwicke et al., 2020; Nelson et al., 2018).

Transparency

Action Editor: Julia Strand

Editor: Daniel J. Simons

Author Contributions

T. E. Hardwicke, D. Szűcs, and J. P. A. Ioannidis designed the study. T. E. Hardwicke, D. Szűcs, R. T. Thibault, S. Crüwell, O. R. van den Akker, and M. B. Nuijten performed the data extraction and coding. T. E. Hardwicke and S. Crüwell performed the data analysis. T. E. Hardwicke

wrote the manuscript. All of the authors provided feedback and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The authors received no specific funding for this work. The Meta-Research Innovation Center at Stanford (METRICS) is supported by a grant from the Laura and John Arnold Foundation. The Meta-Research Innovation Center Berlin (METRIC-B) is supported by a grant from the Einstein Foundation and Stiftung Charité. The work of J. P. A. Ioannidis is supported by an unrestricted grant from Sue and Bob O'Donnell. R. T. Thibault is supported by a postdoctoral fellowship from the Fonds de la recherche en santé du Québec. T. E. Hardwicke receives funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement 841188. D. Szűcs receives funding from the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition (Grant 220020370). O. R. van den Akker is supported by a Consolidator Grant (IMPROVE) from the European Research Council (Grant 726361).

Open Practices

Open Data: <https://osf.io/w8h2q/>


Open Materials: <https://osf.io/w8h2q/>


Preregistration: <https://osf.io/eh5qd/>


All data and materials have been made publicly available via OSF and can be accessed at <https://osf.io/gyzbm/files/>. The protocol and analysis plans were preregistered via OSF and can be accessed at <https://osf.io/eh5qd/>. To facilitate reproducibility, this manuscript was written by interleaving regular prose and analysis code using knitr (Xie, 2017) and papaja (Aust & Barth, 2020) and is available in a Code Ocean container (<https://doi.org/10.24433/CO.4225975.v3>) that recreates the software environment in which the original analyses were performed. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

Tom E. Hardwicke  <https://orcid.org/0000-0001-9485-4952>

Robert T. Thibault  <https://orcid.org/0000-0002-6561-3962>

Sophia Crüwell  <https://orcid.org/0000-0003-4178-5820>

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/25152459211040837>

Notes

1. Determining whether a replication finding contradicts an original finding is a complex issue that we largely sidestep here by

examining only cases that are *strongly contradictory*—that is, according to several reasonable quantitative criteria (e.g., effect size magnitude, *p* values, Bayes's factors), the results of the original study and replication study lead to opposing inferences (e.g., absence vs. presence of an effect). Also note that one can accept that results are strongly contradictory and remain agnostic about the explanation for the contradiction.

2. This terminology was inspired by but does not directly mirror terminology proposed by Imre Lakatos in his work on the rationality of the research community's response when a scientific theory is contracted by empirical evidence (Lakatos, 1970).

3. We deliberately focused on a select group of case studies rather than other potentially larger samples to aid interpretative clarity; for example, the extent to which the results of the large-scale Reproducibility Project in Psychology (RPP; Open Science Collaboration, 2015) actually contradicted the original studies has been contested (Etz & Vandekerckhove, 2016). In addition, original studies were not actually cited in the RPP research report, which may have diluted awareness about relevant replication studies.

References

- Aust, F., & Barth, M. (2020). *Papaja: Create APA manuscripts with Rmarkdown*. <https://github.com/crsh/papaja>
- Bastiaansen, J. A., de Vries, Y. A., & Munafò, M. R. (2015). Citation distortions in the literature on the serotonin-transporter-linked polymorphic region and amygdala activation. *Biological Psychiatry*, *78*(8), E35–E36. <https://doi.org/10.1016/j.biopsych.2014.12.007>
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. <https://doi.org/10.1037/0022-3514.74.5.1252>
- Baumeister, R. F., & Vohs, K. D. (2016). Misguided effort with elusive implications. *Perspectives on Psychological Science*, *11*(4), 574–575. <https://doi.org/10.1177/2F1745691616652878>
- Bishop, D. V. (2019). 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>
- Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. <https://doi.org/10.1016/j.jesp.2013.10.005>
- Budd, J. M., Sievert, M., & Schultz, T. R. (1998). Phenomena of retraction: Reasons for retraction and citations to the publications. *JAMA*, *280*(3), 296–297. <https://doi.org/10.1001/jama.280.3.296>
- Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. <https://doi.org/10.1037/xge0000083>
- Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward

- republicanism up to 8 months later. *Psychological Science*, 22(8), 1011–1018. <https://doi.org/10.1177/0956797611414726>
- Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, 142(2), 301–306. <https://doi.org/10.1037/a0029288>
- Coles, N. A., Larsen, J. T., & Lench, H. C. (2019). A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychological Bulletin*, 145(6), 610–651. <https://doi.org/10.1037/bul0000194>
- Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. SAGE.
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6, Article 621. <https://doi.org/10.3389/fpsyg.2015.00621>
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE*, 11(2), Article e0149794. <https://doi.org/10.1371/journal.pone.0149794>
- 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>
- Fernández, L. M., & Vadillo, M. A. (2020). *Retracted papers die hard: Diederik Stapel and the enduring influence of flawed science*. PsyArXiv. <https://doi.org/10.31234/osf.io/cszpy>
- Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, 26(1), 13–28. <https://doi.org/10.3758/s13423-018-1488-8>
- Greenberg, S. A. (2009). How citation distortions create unfounded authority: Analysis of a citation network. *BMJ*, 339, Article b2680. <https://doi.org/10.1136/bmj.b2680>
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., . . . Zwienerberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546–573. <https://doi.org/10.1177/1745691616652873>
- Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. A. (2020). Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application*, 7(1), 11–37. <https://doi.org/10.1146/annurev-statistics-031219-041104>
- Hardwicke, T. E., Thibault, R. T., Kosie, J., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. (2021). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*. Advance online publication. <https://doi.org/10.1177%2F1745691620979806>
- Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645–654. <https://doi.org/10.1177/1745691612464056>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- Kühl, T., & Bertrams, A. (2019). Is learning with elaborative interrogation less desirable when learners are depleted? *Frontiers in Psychology*, 10, Article 707. <https://doi.org/10.3389/fpsyg.2019.00707>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Kvarven, A., Strömland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4), 423–434. <https://doi.org/10.1038/s41562-019-0787-z>
- Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In A. Musgrave & I. Lakatos (Eds.), *Criticism and the growth of knowledge* (Vol. 4, pp. 91–196). Cambridge University Press. <https://doi.org/10.1017/CBO9781139171434.009>
- Lauden, L. (1981). Peirce and the trivialization of the self-corrective thesis. In *Science and hypothesis* (pp. 226–251). Springer.
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131. <https://doi.org/10.1177/1529100612451018>
- Lewis, M. B. (2018). The interactions between botulinum-toxin-based facial treatments and embodied emotions. *Scientific Reports*, 8, 14720. <https://doi.org/10.1038/s41598-018-33119-1>
- Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, 15(5), Article e0233107. <https://doi.org/10.1371/journal.pone.0233107>
- 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>
- McDiarmid, A., Tullett, A., Whitt, C. M., Vazire, S., Smailino, P. E., & Stephens, E. E. (2021). *Self-correction in psychological science: How do psychologists update their beliefs in response to replications?* PsyArXiv. <https://doi.org/10.31234/osf.io/hjcm4>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s Renaissance. *Annual Review of Psychology*, 69(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Neuliep, J. W., & Crandall, R. (1993). Everyone was wrong: There are lots of replications out there. *Journal of Social Behavior and Personality*, 8(6), 1–8.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>

- Peterson, D., & Panofsky, A. (2020). *Self-correction in science: The diagnostic and integrative motives for replication*. SocArXiv. <https://doi.org/10.31235/osf.io/96qvx>
- Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. <https://doi.org/10.1177/1745691620906416>
- Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, *144*(4), e73–e85. <https://doi.org/10.1037/xge0000058>
- Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science*, *9*(5), 552–555. <https://doi.org/10.1177/1745691614543974>
- Sripada, C., Kessler, D., & Jonides, J. (2014). Methylphenidate blocks effort-induced depletion of regulatory control in healthy volunteers. *Psychological Science*, *25*(6), 1227–1234. <https://doi.org/10.1177/0956797614526415>
- Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*(6), 929–930. <https://doi.org/10.1177/1745691616674460>
- Strack, F. (2017). From data to truth in psychological science. A personal perspective. *Frontiers in Psychology*, *8*, Article 702. <https://doi.org/10.3389/fpsyg.2017.00702>
- Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the Facial Feedback Hypothesis. *Journal of Personality and Social Psychology*, *54*(5), 768–777. <https://doi.org/10.1037/0022-3514.54.5.768>
- Tatsioni, A., Bonitsis, N. G., & Ioannidis, J. P. A. (2007). Persistence of contradicted claims in the literature. *JAMA*, *298*(21), 2517–2526. <https://doi.org/10.1001/jama.298.21.2517>
- Vadillo, M. A. (2019). Ego depletion may disappear by 2020. *Social Psychology*, *50*(5–6), 282–291. <https://doi.org/10.1027/1864-9335/a000375>
- Vadillo, M. A., Hardwicke, T. E., & Shanks, D. R. (2016). Selection bias, vote counting, and money-priming effects: A comment on Rohrer, Pashler, and Harris (2015) and Vohs (2015). *Journal of Experimental Psychology: General*, *145*(5), 655–663. <https://doi.org/10.1037/xge0000157>
- Vazire, S., & Holcombe, A. O. (2020). *Where are the self-correcting mechanisms in science?* PsyArXiv. <https://doi.org/10.31234/osf.io/kgqzt>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2020). Credibility beyond replicability: Improving the four validities in psychological science. *PsyArXiv*. <https://doi.org/10.31234/osf.io/bu4d3>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., . . . Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. <https://doi.org/10.1177/1745691616674458>
- Xie, Y. (2017). *Dynamic documents with R and knitr*. CRC Press.
- Zotero. (2019, June 14). *Retracted item notifications with retraction watch integration*. <https://www.zotero.org/blog/retracted-item-notifications>
- 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>

Curriculum Vitae

My curriculum vitae does not appear in the electronic version of my work for reasons of data protection.

Publication list

First authorships: 4

Co-authorships: 4

- 1) **Crüwell, S.**, van Doorn, J., Etz, A., Makel, M.C., Moshontz, H., Niebaum, J., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2019). Seven Easy Steps to Open Science. *Zeitschrift für Psychologie*. <https://doi.org/10.1027/2151-2604/a000387>
Impact Factor (2017): 1.364
- 2) **Crüwell, S.**, Stefan, A.M., Evans, N.J. (2019). Robust Standards in Cognitive Science. *Computational Brain & Behaviour*. <https://doi.org/10.1007/s42113-019-00049-8>
Impact Factor (2021): N/A, journal started in 2018
- 3) Hardwicke, T.E., Serghiou, S., Janiaud, P., Danchev, V., **Crüwell, S.**, Goodman, S., Ioannidis, J.P.A. (2020). Calibrating the Scientific Ecosystem Through Meta-Research. *Annual Review of Statistics and Its Application*, 7, 11-37. <https://doi.org/10.1146/annurev-statistics-031219-041104>
Impact Factor (2021): 7.917
- 4) Hardwicke, T. E., Wallach, J. D., Kidwell, M., 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>
Impact Factor (2017): 2.504
- 5) Hardwicke, T. E., Szűcs, D., Thibault, R. T., **Crüwell, S.**, van den Akker, O. R., Nuijten, M. B., Ioannidis, J.P.A. (2021). Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/25152459211040837>
Impact Factor (2021): 15.817
- 6) **Crüwell, S.**, Evans, N.J. (2021) Preregistration in Diverse Contexts: A Preregistration Template for the Application of Cognitive Models. *Royal Society Open Science*. <https://doi.org/10.1098/rsos.210155>
Impact Factor (2019): 2.646
- 7) Kent, B. A., Holman, C., Amoako, E., Antonietti, A., Azam, J. M., Ballhausen, H., Bediako, Y., Belasen, A.M., Carneiro, C. F. D., Chen, Y., Compeer, E. B., Connor, C. A. C., **Crüwell, S.**, Debat, H., Dorris, E., Ebrahimi, H., Erlich, J. C., Fernández-Chiappe, F., Fischer, F., Gazda, M. A., Glatz, T., Grabitz, P., Heise, V., Kent, D. G., Lo, H., McDowell, G., Mehta, D., Neumann, W., Neves, K.,

Patterson, M., Penfold, N. C., Piper, S. K., Puebla, I., Quashie, P. K., Quezada, C. P., Riley, J. L., Rohmann, J. L., Saladi, S., Schwessinger, B., Siegerink, B., Stehlik, P., Tzilivaki, A., Umbers, K. D. L., Varma, A., Walavalkar, K., de Winde, C. M., Zaza, C., & Weissgerber, T. L. (2022). Recommendations for empowering early career researchers to improve research culture and practice. *PLoS biology*, 20(7), e3001680. <https://doi.org/10.1371/journal.pbio.3001680>

Impact Factor (2020): 8.029

- 8) **Crüwell, S.**, Apthorp, D., Baker, B. J., Colling, L. J., Elson, M., Geiger, S. J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D. S., Zaneva, M., Brown, N. J. L. (2023). What's in a Badge? A Computational Reproducibility Investigation of the Open Data Badge Policy in one Issue of Psychological Science. *Psychological Science*. 34(4), 512-522. <https://doi.org/10.1177/09567976221140828>

Impact Factor (2020): 7.029

Acknowledgments

My acknowledgments do not appear in the electronic version of my work for reasons of data protection.