



Page Proof Instructions and Queries

Please respond to and approve your proof through the “Edit” tab, using this PDF to review figure and table formatting and placement. This PDF can also be downloaded for your records. We strongly encourage you to provide any edits through the “Edit” tab, should you wish to provide corrections via PDF, please see the instructions below and email this PDF to your Production Editor.

Journal Title: Journal of Empirical Research on Human Research Ethics
Article Number: 1083384

Thank you for choosing to publish with us. This is your final opportunity to ensure your article will be accurate at publication. Please review your proof carefully and respond to the queries using the circled tools in the image below, which are available in Adobe Reader DC* by clicking **Tools** from the top menu, then clicking **Comment**.


Please use *only* the tools circled in the image, as edits via other tools/methods can be lost during file conversion. For comments, questions, or formatting requests, please use . Please do *not* use comment bubbles/sticky notes .



*If you do not see these tools, please ensure you have opened this file with **Adobe Reader DC**, available for free at get.adobe.com/reader or by going to Help > Check for Updates within other versions of Reader. For more detailed instructions, please see us.sagepub.com/ReaderXProofs.



No.	Query
GQ1	Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
GQ2	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
GQ3	Please confirm that the Funding and Conflict of Interest statements are accurate.
GQ4	Please ensure that you have obtained and enclosed all necessary permissions for the reproduction of artistic works, (e.g. illustrations, photographs, charts, maps, other visual material, etc.) not owned by yourself. Please refer to your publishing agreement for further information.
GQ5	Please note that this proof represents your final opportunity to review your article prior to publication, so please do send all of your changes now.
GQ6	Please note, only ORCID iDs validated prior to acceptance will be authorized for publication; we are unable to add or amend ORCID iDs at this stage.
AQ1	The spelling of Ioannidis & Trikolinos (2007) has been changed to Ioannidis & Trikalinos (2007) in the text to match the entry in the references list. Please provide revisions if this is incorrect.
AQ2	Please provide the issue number in Baker (2016), Begley and Ioannidis (2015), Bem (2011), Boutron and Ravaut (2018), Bruns and Ioannidis (2016), Bruton et al. (2020), Brydges (2019), Chatard et al. (2017), Cheung (2019), Cuddy et al. (2018), de Winter and Dodou (2015), Ebersole et al. (2016), Erdfelder and Heck (2019), Fanelli (2018), Faul et al. (2007), Galak et al. (2012), Gervais (in press), Head et al. (2015), Ioannidis (2012), Ioannidis and Trikalinos (2007), John et al. (2012), Kerr (1998), Motyl et al. (2017), Open Science Collaboration (2015), Oswald et al. (2015), Ritchie et al. (2012), Sacco and Brown (2019), Schimmack (2012), Simmons et al. (2011), Simmons and Simonsohn (2017), Simonsohn et al. (2014a), Simonsohn et al. (2014b), Simonsohn et al. (2015), Tjldink et al. (2014), Tsipursky (2017), Wicherts et al. (2016).
AQ3	Please provide DOI number in Brydges (2019), Gervais (in press), Open Science Collaboration (2015), Tsipursky (2017).
AQ4	Reference Cheung (2019) is listed in the reference list but not cited in the text. Please cite in the text, else delete from the list.
AQ5	Please provide volume number in Gervais (in press).
AQ6	Please provide page range in Gervais (in press), Open Science Collaboration (2015).
AQ7	Reference Oswald et al. (2015) is listed in the reference list but not cited in the text. Please cite in the text, else delete from the list.
AQ8	Reference Simonsohn et al. (2015) is listed in the reference list but not cited in the text. Please cite in the text, else delete from the list.

Preliminary Evidence for an Association between Journal Submission Requirements and Reproducibility of Published Findings: A Pilot Study

Journal of Empirical Research on
Human Research Ethics
1–8
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15562646221083384
journals.sagepub.com/home/jre


 **Mitch Brown**¹,  **Robert E. McGrath**², and **Donald F. Sacco**³ 

Abstract

 Systemic efforts have been employed to improve the reproducibility of published findings in psychology. To date, little  research has been conducted evaluating how well these efforts work. In an effort to bridge this gap, the current study looked at journal submission requirements intended to encourage authors to engage in best practices for facilitating reproducible science and offers preliminary evidence for their potential efficacy. We calculated reproducibility indices (p -curves) for three randomly selected empirical studies published in each of 23 psychology journals in 2019 and correlated quantitative results from those analyzes with the number of submission requirements for each journal that intended to ensure compliance with best reporting practices. Results indicated a greater number of submission requirements at a given outlet was associated with indices indicating greater likelihood of reproducibility of findings. We frame findings as impetus for future, more extensive, research to identify causal links between submission requirements and reproducibility.

Keywords

questionable research practices, p -curve, reproducibility, publication, submission requirements

The critical role of science in understanding and solving societal problems necessitates the reproducibility of scientific results. Reproducibility is particularly appealing to a lay public beyond research reported to confirm hypotheses (Ebersole et al., 2016). Nonetheless, scientists are often motivated by institutional and reputational pressures that may implicitly or explicitly influence them to make decisions compromising the accuracy of their findings. Though engaging in these detrimental behaviors could ensure findings meet the threshold considered suitable for publication (i.e., statistically significant findings), they come at the expense of misrepresenting null findings. The discussion of possible detrimental practices has led to concern about the validity of many scientific findings and whether they represent true phenomena or are artifacts of researcher behavior, with numerous published works failing to replicate (Open Science Collaboration, 2015). This concern has coupled with eroding public trust in science, leaving many skeptical of research, and undermining public support of science (Tsipursky, 2017).

Concern over both the accuracy of reporting and protecting public trust in the scientific enterprise has since led to various efforts to mitigate the proliferation of unreliable findings. One increasingly popular means to improve the reproducibility of science, particularly in psychology, is the adoption of submission requirements for a journal that necessitate authors report their engagement in best practices

(e.g., Elsevier, 2019). As the adoption of these policies has become standard in the editorial process, this pilot study sought to provide preliminary evidence on whether and how submission requirements impact the reproducibility of published findings.

Pervasiveness and Consequences of Questionable Research Practices

The scientific community has become increasingly concerned with the reproducibility of many published empirical research findings following the widely publicized “reproducibility crisis,” with psychology frequently being at the forefront of this conversation. First, in a survey of research scientists incentivized to disclose their ethicality in research, 70% admitted being unable to replicate another researcher’s work; more than 50% of respondents had even failed to replicate their own findings (Baker, 2016). Additional efforts to replicate Ben’s (2011) controversial findings for the

¹University of Arkansas, Fayetteville, Arkansas, USA

²Fairleigh Dickinson University, Teaneck, New Jersey, USA

³The University of Southern Mississippi, Hattiesburg, Mississippi, USA

Corresponding Author:

Mitch Brown, Department of Psychological Science, University of Arkansas, 216 Memorial Hall, Fayetteville, AR 72701, USA.
Email: mbl03@uark.edu

existence of extrasensory perception using the same methodology as the original paper have consistently failed (Galak et al., 2012; Ritchie et al., 2012). These were followed by a mass replication effort of 100 well-known psychology studies, finding only 36% of attempts resulted in statistical significance (Open Science Collaboration, 2015). Non-reproducibility, coupled with research documenting pervasiveness of detrimental practices in research, has sparked concern in the scientific community about the trustworthiness of many findings (Baker, 2016; but see Fanelli, 2018).

Although results sections of publications increasingly report null findings (Fanelli, 2018), as of 2014 (the last year on record for this research), a substantial proportion of published p -values indicating significance fell in the range of .041 to .049 (de Winter & Dodou, 2015). A truncated distribution of p -values clustering near .05 is considered indicative of “ p -hacking,” an ethically detrimental behavior designed to ensure results attain conventional significance considered publishable (Simmons et al., 2011). Despite the probability typically set at 5% that statistically significant results are due to chance (i.e., Type I error), methodological and statistical decisions that tend to increase the probability of attaining statistically significant results inflate Type I error rates up to 50% (Simmons et al., 2011). These techniques are frequently called questionable research practices (QRPs). Though falling outside of federal definitions of research misconduct (i.e., falsification, fabrication, plagiarism), these behaviors remain deleterious because of the increased likelihood of false positives they may introduce into the scientific literature. Examples of QRPs include the use of theoretically unjustified covariates chosen on a post hoc basis (Simmons et al., 2011), omitting null findings (Ioannidis & Trikalinos, 2007), and HARKing (hypothesizing after the results are known; Kerr, 1998). Over 90% of surveyed scientists reported previously using at least one QRP (John et al., 2012). A survey of Flemish scientists found 15% of respondents admitted to having engaged in QRPs in the past three years that could have contributed to non-reproducibility (Tijdkink et al., 2014).

Mitigation and Identification of Nonreproducible Findings

Various efforts have been employed to reduce the proliferation of unreliable scientific findings. One approach has involved educational interventions. Although several such efforts have demonstrated efficacy in reducing scientists' endorsement of QRPs as ethically defensible (e.g., Bruton et al., 2020; Sacco & Brown, 2019), they rely on self-policing that is difficult to implement on a systemic level (Ioannidis, 2012). Alternatively, journals have enhanced their gatekeeping by instituting submission checklists in which authors must state their explicit engagement in best practices prior to submission (Wicherts et al., 2016). For example, authors

submitting to *Journal of Experimental Social Psychology* must state in the manuscript that they have reported all measures they administered (Elsevier, 2019). These checklists make adherence explicit before review, in contrast to peer reviewers having to extrapolate whether the authors employed best practices through potentially ambiguous reporting. However, broad initiation of checklists has only occurred over the last few years. Their efficacy as a safeguard against unreliable findings has not yet been evaluated.

Several metrics have been developed to determine the likelihood of reported findings being the product of p -hacking or QRPs. Metrics have demonstrated efficacy in identifying the degree reported findings are reproducible. For example, the increasingly popular p -curve analysis assesses the degree a range of conventionally significant p -values are negatively skewed within the critical range of significance (Simonsohn et al., 2014a). Significant effects influenced by QRPs tend towards negative skew because researchers often manipulate findings until they just barely achieve significance, so significant effects tend to cluster around the threshold for significance. These analyzes afford researchers the opportunity to identify the presence of true effects with limited information (i.e., conventionally significant inferential statistics) from published studies. These additionally provide estimates for the possibility certain published findings may have been selectively reported (Simonsohn et al., 2014b). As a result, p -curve analyzes became popular in meta-scientific investigations of reproducibility to determine whether effects reported in published findings represent likely true effects in psychological research (e.g., Chatard et al., 2017; Cuddy et al., 2018; Simmons & Simonsohn, 2017).

Current Study

Although p -curve has demonstrated utility for diagnosing how robust scientific findings are, it has typically been used to determine the reproducibility of a specific finding or the work of a specific author or journal (e.g., Motyl et al., 2017). The purpose of the current study is to provide preliminary evidence for the utility of p -curve analyzes for assessing the efficacy of systemic efforts to improve the reproducibility of findings as a pilot investigation. Although a pilot study, we expected for higher numbers of journal submission requirements at a psychology journal to be associated with more reproducible research as indicated by p -curve analyzes as a tentative prediction. We provide all data and information for this study at: https://osf.io/yfstv/?view_only=5b0f99f405594e33b8a226dd4e31ed8b

Method

Journal Selection

For this project, we identified 23 psychology journals across a variety of psychology subfields (e.g., social, cognitive,

clinical) that publish empirical research using quantitative methods to provide an ostensibly representative sample of the field across disciplines. Journals for this analysis were selected by the first author (see Table 1 for selected journals). Selection of specific journals was not random, as we sought to include outlets with differing levels of potential impact and topics within the different subfields of psychology. Journals ranged in impact from relatively high (e.g.,

Journal of Personality and Social Psychology) to relatively low (e.g., *Experimental Psychology*). Focal topics of the journal also varied from broad (e.g., *Psychological Science*) to more specialized (e.g., *Memory*).

Our inclusion criterion for journals was an impact factor of at least 1.00 in 2019. An impact factor of that minimal level could suggest greater venerability of an outlet that would have existed before the implementation of submission guidelines, thereby providing greater range of potential significance values. Because of the pilot nature of these findings, we sought to identify a minimal number of journals to analyze efficiently in preparation for a potential larger-scale study if preliminary evidence suggested an association between study variables.

Journals selected had an average impact factor that would be considered relatively high for psychology ($M = 2.93$, $SD = 1.47$; Range: 1.00–6.13). Another consideration was ensuring some variability across journals in their submission requirements. Table 2 provides the list of requirements identified by reviewing journal websites.

Table 1. Journals Included in the Study.

Journal	Field	# Rules	2019 IF
<i>Attention, Perception, & Psychophysics</i>	Cognitive/Developmental	6	1.79
<i>Body Image</i>	Clinical	0	3.12
<i>Developmental Psychology</i>	Cognitive/Developmental	2	3.34
<i>Emotion</i>	Experimental/General	4	3.12
<i>Evolution and Human Behavior</i>	Social/Personality	1	2.96
<i>Experimental Psychology</i>	Experimental/General	7	1.00
<i>Journal of Abnormal Psychology</i>	Clinical	0	5.52
<i>Journal of Consulting and Clinical Psychology</i>	Clinical	4	4.36
<i>JEP: General</i>	Experimental/General	0	3.50
<i>JEP: Human Perception & Performance</i>	Cognitive/Developmental	0	2.94
<i>JEP: Learning, Memory, & Cognition</i>	Cognitive/Developmental	0	2.67
<i>Journal of Experimental Social Psychology</i>	Social/Personality	8	3.29
<i>Journal of Personality and Social Psychology</i>	Social/Personality	3	5.92
<i>Journal of Research in Personality</i>	Social/Personality	4	2.57
<i>Journal of Social and Personal Relationships</i>	Social/Personality	0	1.68
<i>Journal of Social Psychology</i>	Social/Personality	2	1.10
<i>Memory</i>	Cognitive/Developmental	2	1.71
<i>Personal Relationships</i>	Social/Personality	2	1.09
<i>Personality and Individual Differences</i>	Social/Personality	4	2.00
<i>Personality and Social Psychology Bulletin</i>	Social/Personality	4	2.65
<i>Psychological Science</i>	Experimental/General	5	6.13
<i>Social Psychological and Personality Science</i>	Social/Personality	8	3.60
<i>Social Psychology</i>	Social/Personality	7	1.36

Note. JEP = Journal of Experimental Psychology; # Rules = number of submission requirements reflecting best practices; IF = impact factor.

Procedure

After selecting the journals, the first author identified all unique submission requirements for a journal related to transparency and best practices, as opposed to submission requirements related to formatting or style. Upon identification of unique requirements, the second and third author performed a reliability check to confirm their presence. On average, journals had 3.17 best practice requirements (SD

Table 2. Submission Requirements Identified.

Journal rules
Report power or sensitivity analyzes
Report effect size and confidence intervals
Report all manipulations
Disclose multiple tests
Report outliers and exclusions
Report all studies
Report all dependent variables
Report the availability of data/data repository link
Share data (required or recommended)
Follow Journal Article Reporting Standards
Register trials (required or recommended)
Report psychometric properties
Report scoring protocols
Report exact p -values
Report descriptive statistics
Justify choice of mediators
Make all materials available
Make all code available
Provide a file of study materials as presented to participants for reviewers' edification

= 2.70). Each rule was verified to be currently in effect as of 2019, as confirmed by the journal website.

We then randomly selected three empirical articles from each selected journal to construct a p -curve for that journal, a number of papers that would provide a sufficient number of p -values to calculate this metric and minimize the statistical impact of any single selected article. Articles were selected based on being accepted for publication or paginated in 2019 to ensure their review following the implementation of the enumerated submission requirements. Articles authored by the current research team, their departmental colleagues, or any prior collaborators (as of September 2019) of the authors were excluded to prevent potential conflicts of interest. The random selection further prevented potential stimulus effects that could persist in deliberate selections of articles by topic. The first author then identified the relevant statistical information for p -curve analyzes as outlined by Simonsohn et al. (2014b). Data were independently checked by two graduate students. Discrepancies were resolved by discussion.

In studies with multiple p values, different strategies have been suggested for dealing with the non-independence of these findings. For example, Head et al. (2015) recommended only using the first p value reported for a study. For the primary analysis, we chose to include all significant findings from the study, or the first study with significant findings in a multi-study paper. However, we also generated p -curves using the first- and last-reported significant values in each study as suggested by Simonsohn et al. (2014b) to evaluate the impact of non-independent observations on our conclusions. For studies in which the first-reported analysis involved two p values (e.g., a cross-over interaction, or multiple pairwise comparisons), both were included in the computations of p -curve. A power analysis using G*Power (Faul et al., 2007) indicated the primary analysis was associated with a power of .44. In interpreting results, we will therefore note significance, but will also consider the size of correlations we found.

Results

We entered the relevant inferential statistics into the online calculator at <https://p-curve.com>, separately for each journal, three times: all p values, first value per study, and last value. Each p -curve analysis generated three statistics that were used to evaluate replicability across the three articles for that journal. First was the p value for a binomial test evaluating the degree to which the significance levels for that journal were consistent with the studies providing evidential (replicable) results. Lower values suggested greater evidence of replicability. Second was the p value for a binomial test whether the results were inadequate (unreplicable). Higher values suggested greater evidence of replicability. Finally, the power of the studies was estimated, with

higher values again suggesting greater replicability of significant findings.

We included two covariates based on a priori considerations: the journal's impact factor and number of p values included in the full analysis. The former was justified because impact factor is potentially an indicator of quality of publications in that journal. It is also possible a stronger reputation would allow journal editors to implement more stringent requirements. Similarly, p values for binomial tests could be influenced by the number of data points. We covaried both variables from the full analysis. For the analyzes considering the first- and last-reported values, we omitted number of p values, as they were essentially fixed across journals.

Table 3 provides results from the tests of replicability. The first column provides descriptive statistics across journals for p values associated with the binomial test of evidential results and the binomial test of inadequate results, and for the estimated power. We also computed partial correlations between these results and the number of submission requirements for the journal. The number of rules was significantly negatively associated with the binomial test p value in the primary analysis (all p values). Significance was not achieved for the secondary analyzes (based on the first and last p value) but all were negative as expected. The same pattern emerged in the positive direction for the

Table 3. Relationships between Predictors and p -curve Outcomes.

Analysis	M (SD)	Correlations		
		# Rules	Binomial	Binomial (In.)
All p values				
Binomial	.12 (.19)	-.48*		
Binomial (In.)	.88 (.17)	.46*	-.89**	
Power	.89 (.18)	.37	-.70**	.75**
First p value				
Binomial	.22 (.19)	-.25		
Binomial (In.)	.88 (.17)	.19	-.97**	
Power	.89 (.18)	.28	-.54*	.57*
Last p value				
Binomial	.30 (.23)	-.07		
Binomial (In.)	.81 (.22)	.08	-.98**	
Power	.81 (.29)	.31	-.52*	.48*

* $p < .05$. ** $p < .01$.

Note. "All p values" analysis included all significant p values from three studies per journal; "First p value" analysis included the first significant p value from each study; "Last p value" analysis included the last significant p value. Values are partial bivariate correlations controlling for impact factor and number of analyzes for all p values analysis, impact factor for first and last p value analyzes. Binomial = p value for the binomial test of evidentiary value; Binomial (In.) = p value for the binomial test of inadequate value; Power = estimated power of tests. Lower values for Binomial are considered evidence of greater replicability; higher values for Binomial (In.) and Power are indicative of greater replicability.

binomial test of inadequate results. Though none of the correlations between number of rules and power were significant, all were again in the positive direction. In every case, then, the direction of the relationship indicated journals with more rules were associated with results that p -curve analyzes indicated were more replicable. Six of these nine correlations were in a range considered medium-sized for correlations (Brydges, 2019; Hemphill, 2003). As could be expected, the last two columns indicate that all correlations between different tests of replicability were significant and in the expected direction.

Discussion

The prevalence of QRPs within a scientific literature creates an obligation for researchers to identify systemic efforts that may effectively improve the reproducibility of published findings (Begley & Ioannidis, 2015; John et al., 2012). Results from this pilot study suggest some value of submission requirements as a method of increasing reproducibility. We found that a greater number of submission requirements instituted by the journal was associated with p -curve binomial results and power estimates that were more indicative of reproducibility, namely p values that clustered less around the threshold of significance as would be indicative of p -hacking (Simonsohn et al., 2014a). Only the correlations for binomial tests in the most powerful set of analyzes (those where all p values were included) were significant. Effects based on the first and last p value were smaller, particularly those for the last p value, and consistently non-significant. It is unclear why that would be. The restriction of analyzes to a single p value per study versus the inclusion of all p values is recommended as an analytic approach by Simonsohn et al. (2014b), but the effects of these different approaches on the outcome has not been extensively evaluated. It may well be that first p values often reflect preliminary analyzes preceding the tests of key hypotheses, whereas final p values are more likely to be associated with secondary analyzes; in other words, there may be a tendency for the most important analyzes to appear near but not quite at the beginning of the statistical results. If so, the analyzes most likely to be the product of p -hacking could well appear in the middle of the mix. That is purely post hoc speculation on our part, however.

That said, all relationships were in the expected direction, and most were at least medium-sized. The findings suggest publication requirements could be deemed as a mitigating tool that reduces the likelihood authors rely on researcher degrees of freedom to ensure their data conform to their hypotheses and remain “publishable” under conventional standards (Simmons et al., 2011).

These results additionally provide initial evidence for the potential utility of reproducibility indices as an outcome to measure research norms’ influence on the quality of published findings in a field in general. Previous research

using these indices have focused largely on the degree to which a given journal or researcher is associated with reproducible findings (e.g., Cuddy et al., 2018; Motyl et al., 2017). However, it remains unclear whether these previous analyzes’ results were based in the encouragement of best practices without considering systemic means designed specifically for that purpose (Bruns & Ioannidis, 2016; Erdfelder & Heck, 2019). The preliminary evidence in this analysis affords researchers an opportunity to determine whether submission requirements designed to encourage best practices are associated with reproducible research beyond the more descriptive nature of previous analyzes.

Though the current project provides initial evidence that these journal policies are related to more reproducible results, these results are cross-sectional and so cannot be used to suggest these requirements *cause* an increase in reproducibility. To determine the causal link between these variables, future research would benefit from conducting p -curve analyzes using articles from before versus after the implementation of more stringent submission requirements. Such a longitudinal analysis would afford more conclusive evidence on the value of these policies. Additionally, more extensive journal review would increase the statistical power of the analyzes and allow for more granular analyzes such as a comparison of the effectiveness of different requirements. It is noteworthy the mean impact factor for the journals used in this preliminary review was quite high. These journals may demonstrate less variability in quality of publications than would be true for less-selective journals. For example, mean p values for the binomial test of inadequate value and power estimates were in all three analyzes $>.80$, suggesting studies of unusually high quality across the psychological literature. In keeping with the earlier comment this study suggests the potential for using reproducibility indices to evaluate a field in general, a larger investigation might also allow for comparisons across subfields within psychology.

Best Practices

As research moves forward in understanding the basis of reproducibility through these systemic efforts, it becomes necessary to identify and employ best practices. For example, future research would benefit from replicating these findings using larger samples across a wider gamut of journals and topics. Despite relatively robust effects with several medium effect sizes in a small sample, the pilot nature of this study precludes us from drawing larger inferences from our data beyond a preliminary analysis, particularly because of our relatively limited sample size. Nonetheless, it is hoped that the current findings spur larger-scale replications to determine the robustness of our findings and critical moderating variables. These larger-scale replications would further afford more opportunities to identify which computations of p -curves would be

most optimal in identifying effects. Within these larger replications could be further consideration of journals across a wider gamut of impact factors and stages of prolificity. Our methodological decision to focus on ostensibly established journals was in the service of identifying outlets that could be more influential in shaping perceptions of the fields reproducibility, especially given previous controversies of publishing findings in flagship journals (e.g., Bem, 2011). It should be noted that many newer outlets in psychology are specifically promoting best practices as part of their mission. For example, the recently established journal *Comprehensive Results in Social Psychology* requires pre-registration for submissions (Taylor & Francis, 2015). The European Union has further been publicly funding new outlets that employ similar submission requirements of more established outlets (e.g., PsychOpen, 2015). A future investigation could specifically address whether more established outlets experience different levels of reproducibility compared to newer outlets.

The current analysis indicated effects considering the entire range of reported values in a study for the p -curve was most robustly associated with submission requirements. However, such an analytic decision remains one of several possible computations as recommended by previous research (Head et al., 2015; Simonsohn et al., 2014b). Research would benefit from identifying which practice for computation is most consistently robust.

This future research employing larger samples would also afford added benefits to inform which specific practices are most associated with more reproducible science. That is, research could identify which best practices are specifically associated with the most reproducibility so journals could determine which submission requirements may be more imperative than others. Conversely, these analyzes could identify which rules are *unassociated* with reproducibility, thus leading outlets to abandon ineffective rules.

In addition to specific methodological considerations that would lead to best practices in research, it would prove advantageous to facilitate best practices in scientific discourse. The growing concern of “toxicity” in open science initiatives could dissuade participation (Gervais, in press). Discussions among psychological researchers in public forums (e.g., Twitter) have highlighted escalating hostility between researchers, particularly when the discussion involves reproducibility indices, including p -curves (e.g., Letzter, 2016). Use of various indices has also recently come under scrutiny for their potential for spreading libelous claims and issues with their overall reliability (Heine, 2021). Researchers should use caution when reporting analyzes and not immediately infer or convey pernicious intentions. Such exchanges or inappropriate use of metrics could undermine their utility in tracking shifts in observed reproducibility while creating hostility between scientists that could make open science practices unattractive.

Research Agenda

Moving beyond larger samples and longitudinal designs, future research on this topic would also benefit from using a wider variety of potential indicators of reproducibility. A major limitation of p -curve analyzes is its reliance on statistically significant findings. This index therefore omits effects failing to attain conventional significance. For example, the R-Index or Magic Index estimates the likelihood of certain findings achieving significance via QRPs by comparing the number of significant effects with the expected number given estimated power (Schimmack, 2012, 2014). Given the small size of this pilot, we did not want to proliferate the number of analyzes we conducted. Future studies could potentially generate multiple replicability indices simultaneously, given their reliance on complementary information.

Another avenue for future research would expand investigation beyond psychological literature. Although the reproducibility crisis in psychology has been among the most publicized because of mass replication efforts (e.g., Open Science Collaboration, 2015), it is not the only science under scrutiny for concerns over unreliable findings. Biomedical research has also been criticized for unreliable findings, including some high-profile retractions (e.g., Begley & Ioannidis, 2015; Boutron & Ravaud, 2018; Retraction Watch, 2020). As scientists continue to identify the systemic efforts that promote best practices in psychology, it would be similarly critical to identify the utility of submission requirements for other scientific fields.

Educational Implications

The educational implications for these indices could focus around informing editorial staffs on the utility of submission requirements that proliferate reproducible findings. Upon extensive analyzes of various requirements and how they associate with reproducibility, researchers could provide discrete recommendations to editors on what would optimally facilitate reproducibility. This could lead to evidence-based publication practices.

Conclusion

Although various requirements have been implemented by journals to ameliorate proliferation of non-reproducible findings, researchers have yet to conduct large-scale investigations on the effects of these policies. As these policies have become normative in psychology, the current study provides initial evidence on the degree to which these policies are related to the publication of more reproducible findings. This empirical starting point should encourage future investigations in identifying how to optimize systemic efforts to improve published research.



Declaration of Conflicting Interests

GQ3 The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

GQ6 Mitch Brown  <https://orcid.org/0000-0001-6615-6081>
Donald F. Sacco  <https://orcid.org/0000-0001-6017-5070>

References

- Baker, M. (2016). Reproducibility crisis. *Nature*, *533*, 353-366. **AQ2** <https://doi.org/10.1038/nature17990>
- Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*, 116-126. <https://doi.org/10.1161/CIRCRESAHA.114.303819>
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425. <https://doi.org/10.1037/a0021524>
- Boutron, I., & Ravaud, P. (2018). Misrepresentation and distortion of research in biomedical literature. *Proceedings of the National Academy of Sciences*, *115*, 2613-2619. <https://doi.org/10.1073/pnas.1710755115>
- Bruns, S. B., & Ioannidis, J. P. (2016). *p*-curve and *p*-hacking in observational research. *PLoS One*, *11*, e0149144. <https://doi.org/10.1371/journal.pone.0149144>
- Bruton, S. V., Brown, M., & Sacco, D. F. (2020). Ethical consistency and experience: An attempt to influence researcher attitudes toward questionable research practices through reading prompts. *Journal of Empirical Research on Human Research Ethics*, *15*, 216-226. <https://doi.org/10.1177/1556264619894435>
- Brydges, C. R. (2019). Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging*, *3*, igz036. **AQ3**
- Chatard, A., Bocage-Barthélémy, Y., Selimbegović, L., & Guimond, S. (2017). The woman who wasn't there: Converging evidence that subliminal social comparison affects self-evaluation. *Journal of Experimental Social Psychology*, *73*, 1-13. <https://doi.org/10.1016/j.jesp.2017.05.005>
- Cheung, M. W. L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*, 387-396. <https://doi.org/10.1007/s11065-019-09415-6> **AQ4**
- Cuddy, A. J., Schultz, S. J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, *29*, 656-666. <https://doi.org/10.1177/0956797617746749>
- de Winter, J. C., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, e733. <https://doi.org/10.7717/peerj.733>
- Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLoS Biology*, *14*, e1002460. <https://doi.org/10.1371/journal.pbio.1002460>
- Elsevier (2019). *Journal of Experimental Social Psychology: Guide for authors*. <https://www.elsevier.com/journals/journal-of-experimental-social-psychology/0022-1031/guide-for-authors>
- Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and p-hacking with the *p*-curve tool. *Zeitschrift für Psychologie*, *227*, 249-260. <https://doi.org/10.1027/2151-2604/a000383>
- Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, *115*, 2628-2631. <https://doi.org/10.1073/pnas.1708272114>
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, *39*, 175-191. <https://doi.org/10.3758/BF03193146>
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, *103*, 933-948. <https://doi.org/10.1037/a0029709>
- Gervais, W. M. (in press). Practical methodological reform needs good theory. *Perspectives on Psychological Science*. **AQ5**
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*, e1002106. <https://doi.org/10.1371/journal.pbio.1002106> **AQ6**
- Heine, S. J. [@StevenHeine4] (2021). "Many people feel ashamed..." [Tweet]. Twitter. <https://twitter.com/StevenHeine4/status/1365414572031574018>
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78-79. <https://doi.org/10.1037/0003-066X.58.1.78>
- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645-654. <https://doi.org/10.1177/1745691612464056>
- Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245-253. <https://doi.org/10.1177/1740774507079441>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524-532. <https://doi.org/10.1177/0956797611430953>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196-217. https://doi.org/10.1207/s15327957pspr0203_4
- Letzter, R. (2016). Scientists are furious after a famous psychologist accused her peers of "methodological terrorism". *Business Insider*.
- Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, *113*, 34-58. <https://doi.org/10.1037/pspa0000084>
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial

- discrimination: Small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, 108, 562-571. <https://doi.org/10.1037/pspa0000023>
- AQ7 PsycOpen (2019). Author guidelines for Social Psychological Bulletin. Retrieved on January 20, 2022. <https://spb.psychopen.eu/index.php/spb/author-guidelines>
- Retraction Watch (2020). Retracted coronavirus (COVID-19) papers. <https://retractionwatch.com/retracted-coronavirus-covid-19-papers/>
- Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's "retroactive facilitation of recall" effect. *PLoS One*, 7, e33423. <https://doi.org/10.1371/journal.pone.0033423>
- Sacco, D. F., & Brown, M. (2019). Assessing the efficacy of a training intervention to reduce acceptance of questionable research practices in psychology graduate students. *Journal of Empirical Research on Human Research Ethics*, 14, 209-218. <https://doi.org/10.1177/1556264619840525>
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551-566. <https://doi.org/10.1037/a0029487>
- Schimmack, U. (2014). Quantifying statistical research integrity: The replicability index. *Unpublished Manuscript*.
- 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, 1359-1366. <https://doi.org/10.1177/0956797611417632>
- Simmons, J. P., & Simonsohn, U. (2017). Power posing: *p*-curving the evidence. *Psychological Science*, 28, 687-693. <https://doi.org/10.1177/0956797616658563>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666-681. <https://doi.org/10.1177/1745691614553988>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534-547. <https://doi.org/10.1037/a0033242>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *p*-curves: Making *p*-curve analysis more robust to errors, fraud, and ambitious *p*-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144, 1146-1152. <https://doi.org/10.1037/xge0000104>
- AQ8 Taylor & Francis (2015). Aims and scope. *Statement from Editorial Board of Comprehensive Results in Social Psychology*. Retrieved January 20, 2022. <https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=rrsp20>
- Tijdsink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics*, 9, 64-71. <https://doi.org/10.1177/1556264614552421>
- Tsipursky, G. (2017). Towards a post-lies future: Fighting "alternative facts" and "post-truth" politics. *The Humanist*, 77, 12.
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, 7, 1832. <https://doi.org/10.3389/fpsyg.2016.01832>

Author Biographies

Mitch Brown is an instructor and researcher in psychology at University of Arkansas, USA. His research primarily concerns the social and evolutionary underpinnings of motivated social perception and cognition, which has additionally considered how individuals perceive various forms of ethical research behavior. He was involved in the conceptualization of this research project and implemented its conductance. He wrote the initial draft of this manuscript and was involved in data analysis.

Robert McGrath is a professor of psychology at Fairleigh Dickinson University in New Jersey. His primary research interests fall in the area of character and virtue. He maintains secondary research programs in psychological measurement and professional issues in psychology. He was involved in the conceptualization of the research project and design of the data analysis, and assisted in the writing of the manuscript.

Donald Sacco is an associate professor of psychology at The University of Southern Mississippi, USA. His research interests include developing interventions to reduce engagement in questionable research practices and foster best practices in scientific research. He was involved in the conceptualization of the research project and assisted in the writing of the manuscript.