Social and Personality Psychology in the Wake of a Crisis

By

SARAH R. SCHIAVONE

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Mijke Rhemtulla, Chair

Simine Vazire

Andrew Fox

Committee in Charge

2023

Abstract

The past decade has brought to light many questions and concerns about the validity of psychological research. In chapter 1, I argue that the field of social and personality psychology must reckon with the wave of doubts about the credibility of our research that emerged during the replication crisis and credibility revolution in the 2010s. To do so, we must take stock of the state of the field and empirically evaluate whether self-correction has occurred before declaring the crisis to have passed. I propose an agenda for metascientific research and review approaches to empirically evaluate and track where we are as a field (e.g., analyzing the published literature, surveying researchers). I describe one such project, SPPSPSSPP, underway in our research group and emphasize the need for empirical evidence to evaluate the credibility of research in social and personality psychology.

Validity is a critical component of research quality, and one that is both paramount and complicated for a field to assess. In chapter 2, I introduce a tool (seaboat.io) to aid researchers and reviewers in identifying potential threats to the validity of empirical research. This tool was developed through an iterative consensus-based process of eliciting expert feedback to select potential validity threats that are most common and most serious in psychological science. Reviewers can visit seaboat.io to identify validity threats relevant to the research they are evaluating and generate a report that can be shared alongside peer traditional review reports or used in post-publication peer review.

In chapter 3, I investigate researchers' in social and personality psychology perceptions of the state of the field and of the published literature. To explore how researchers perceive the field to have changed over time, I compare their perceptions of articles published in 2010, just before the advent of the replication crisis, vs. articles published a decade later. I also examine researchers' perceptions of their own work, what qualities they consider important when evaluating research quality, and explore individual

differences among researchers' perceptions. Overall, these findings indicate that researchers perceive the quality of the published literature to have improved in many ways over the last decade, where significant strides are thought to have been made, and what weaknesses remain.

Acknowledgements

When I started graduate school, I had heard that it would be hard—and it was. Yet, it was not difficult in the ways that I expected. The last six years have been a whirlwind of ups and downs on many scales, from the general chaos of the last 6 years in the U.S., moving across the country, working through a global pandemic, to finishing my dissertation while participating in the largest strike in the history of higher education. I have learned many things during this time, about science and about myself. I consider myself incredibly lucky to have met and shared this time with some truly remarkable people. There are many thanks I wish to give and many people whose support has meant the world to me, who I have become stronger just by knowing, and who I cannot imagine having gotten here without.

Simine, thank you for being such a truly wonderful advisor. Very early in my academic career I remember overhearing grumbles and some not-so-hushed complaints about an alleged troublemaker named Simine who dared to question and criticize the status quo, and quickly knowing I was going to like her. After later meeting Simine, I indeed concluded that she was a badass and someone who I admired for the values that she held and most importantly, for her willingness to do something about them. The last few years of working with and getting to know Simine have only made me grow to respect her even more. There is so much I could say and so many stories I could share, but to say it briefly–I am so happy to have met you and to have had you as my advisor, as well as my friend.

Thank you for your unwavering support, through the good times and the tough times. Thank you for standing up for me and standing with me. It is much easier to be brave when you are around others who are. Thank you for endeavoring into this metascience adventure we have so absurdly named SPPSPSSPP. I am proud of the work we are doing, and I have had a lot of fun doing it.

In his book *Letters to a Young Contrarian*, Christopher Hitchens wrote "The noble title of "dissident" must be earned rather than claimed; it connotes sacrifice and risk rather than mere disagreement, and it has been consecrated by many exemplary and courageous men and women." The more I have gotten to know Simine, the more she has shown herself to be someone willing to take risks, to hold to her values, and to do the right thing even when it is costly. I am both grateful and proud to work with you, Simine, and so excited for the next two years.

To the friends I have made, thank you for the fun, the laughs, the distractions when we should have been working, the silliness, the somber moments, and the steadfastness.

Dr. Lenhausen/Mads, when I came to California, I had such big plans. These included focusing on the work I wanted to do and not spending too much time making friends since I did not plan to be in Davis very long. Well, I am uncharacteristically happy to say that my plan failed. Thank you for being the best friend and roommate I could have ever asked for. I love our friendship and how quickly we are to devolve into laughter. We both went through a lot in a short amount of time—navigating a pandemic, losing loved ones, finishing our dissertations, etc. Yet we somehow managed to find ways of having so much fun despite the terrible things happening. Thank you for always being there for me, all the beers we shared, the video games we beat, the burritos we made, the days we danced in the living room, and the countless adventures we had and glowsticks we went through.

Scott and Chelsea, thank you both for always being such a solid rock of support, love, and friendship. I am so happy and lucky to have you in my life. Thank you for all the encouragement, for letting me crash with you so many times, and all the fun we have had. You mean the world to me.

To my former lab mate, Maxine, thank you for what you bravely said at that dinner many years ago. It is striking how such small events can dramatically alter the trajectory of

one's life. This dissertation most certainly would not exist had you not and I am incredibly grateful that you did.

To Kay, Andrew, Sam, Steph, and the Hot Business Office (Sarah Beth, Travis, Nava, Maxine, and Brian), thank you for being the best part of those years. I am so proud of you all.

To Jesse, I am so glad our paths crossed, and we were "fake first years" together. I am thankful we were able to have some adventures before Covid derailed all our plans. Thank you for all the support and all the science talks.

To my twitch community and the lovely friends that I have made as a result, thank you all for the support, for watching my sometimes-mediocre gameplay, and for being there to laugh with me and sometimes at me.

To Larry Bates and Richard Hudiburg, thank you for helping me discover how much I love research and for the support and mentorship you provided. This dissertation would not exist if you had not taken the time to invest and believe in me.

To Daniël Lakens and the Red Square Lab, thank you for welcoming me into your lab, for all the kindness you have shown me, and for being a wonderful example of the power and importance of teamwork.

To Mom, Dad, Kristin, David, Kate, Becca, and Esther, thank you for supporting and loving me, and for always wanting me to come home.

To Andrew and Mijke, thank you for the encouragement, the feedback, and for being on my committees. Mijke, thank you for so kindly adopting me into your lab.

To Simine, my lab mate Beth, Fiona, and the entire MetaMelb lab in Melbourne, I can't wait for us all to finally be working on the same side of the planet, let alone in the same time zone and country!

Table of Contents

Abstract ii

Acknowledgements iv

Table of Contents vii

Chapter 1: Reckoning with Our Crisis: An Agenda for the Field of Social and Personality Psychology 1

An Agenda for a Reckoning 3

Observations of the Published Literature 4

Beyond the Published Literature 12

Integrating Approaches 15

Conclusion 19

Chapter 2: A Consensus-Based Tool for Evaluating Threats to Validity 22

Identifying Potential Threats to Validity 23

Caveats and Limitations 27

Recommended Uses 29

Method and Results 30

Expert Reviewers 31

Wave 1 31

Wave 2 32

Wave 3 33

Chapter 2 Acknowledgements 37

Chapter 3: Researchers' Perceptions of the State of Social and Personality Psychology 39

Understanding Perceptions of the Field Overall, the Published Literature, and Researchers' Own Work (Part I) 41

Perceptions of Field Overall 41

Perceptions of Published Literature 43

Perceptions of Researchers' Own Work 46

Understanding What Researchers Value When Evaluating Research (Part II) 48

Understanding Individual Differences Among Researchers' Perceptions (Part III) .. 51

Intellectual Humility 52

Support for the Open Science Movement 52

Career Stage 53

Method 57

Transparency and Openness 57

Participants and Recruitment Strategy 57

Procedure and Measures 60

General Analytic Approach 63

Results and Discussion 65

Part I. Overall Perceptions of the Field of Social and Personality Psychology, the Published Literature, and Researchers' Own Work 65

Part II. Understanding What Researchers Value When Evaluating Research 88

Part III. Understanding Individual Differences Among Researchers' Perceptions 94

General Discussion 119

References 130

Chapter 1: Reckoning with Our Crisis: An Agenda for the Field of Social and Personality Psychology

When historians of science look back on the 2010s in social and personality psychology, the decade will likely stand out as a period of exceptional doubt and self-scrutiny within the field. During the 2010s, a great deal of the field's attention was consumed by the 'replication crisis' and resulting reform movement, sometimes called the credibility revolution (a phrase borrowed from Angrist & Pischke, 2010; Vazire, 2018). This crisis hit many fields beyond social and personality psychology, but our field was arguably at or near the epicenter of the crisis. This led to questions from within and outside the field about our field's credibility as a science. Having just come out of this tumultuous decade, we have a responsibility as a field to take stock and ask ourselves what we have learned, and what, if anything, we ought to do differently moving forward.

How a field responds when its credibility comes under serious threat is an important marker of its commitment to scientific values. Our credibility as a science depends upon our ability to demonstrate, with concrete actions, that we are committed to self-correction. What we do now, in the 2020s, will determine how we look to future members of our field, to historians of science, and to the general public. Did our crisis in the 2010s drive us to thoroughly and systematically examine and subsequently improve our practices? Or did we make superficial changes and largely continue with business as usual?

The crises and concerns brought to the forefront of the field of social and personality psychology over the past decade have elicited a range of reactions and responses. At one extreme, researchers declared the literature untrustworthy and called for large-scale change. At the other extreme, researchers denied there being any cause for concern, arguing that the status quo in social and personality psychology was working as it should. Many researchers' reactions fell somewhere in between. Indeed, a common narrative is that while

our field may have had serious problems prior to the crisis, those problems have been addressed by the reforms made in response to the crisis (e.g., Kahneman, 2022). According to this view, if social and personality psychology's credibility was ever shaky, it has now been largely shored up through incremental changes that are part of the normal progress of science and sufficient to keep a field on track.

We argue that there are serious risks in prematurely declaring the crisis a thing of the past without first collecting evidence. Specifically, we risk further undermining our credibility if it were to become apparent that we had not taken the necessary steps to self-correct. Here, we argue that we have relatively little evidence about what has changed, and we cannot know how much progress has been made—and where we are as a field—without careful empirical examination of our practices and published research.

We must reckon with what the 2010s brought to light. This means carefully studying and reforming our practices where they fall short of our standards, and attempting to identify the structural features of our field that contributed to the crisis in the first place. The coming decade is a critical period for our field. We will be judged on what we do now, what we learn from the past decade, and what actions we take to change the course of the field towards becoming a more credible science. A full reckoning with our turbulent decade of crisis requires us to undertake systematic analyses of our past and our present. What problems were (or were not) addressed and how? Have new problems emerged as a result of these changes? Where do we still need to improve?

To begin to examine these questions, we propose an agenda for the field's reckoning and imagine what commitment to self-correction might look like for the field of social and personality psychology. We give examples of how metascience can help to track and evaluate where we are as a field by highlighting previous and ongoing metascientific projects, and argue that large-scale metascientific efforts to combine and synthesize across

multiple approaches, measures, timepoints, and sources of data are needed to better evaluate and understand the state of our science. Empirical evidence about the state of our field is necessary if we are to take self-correction seriously, and if we hope to avert future crises.

An Agenda for a Reckoning

What would it mean to reckon with our field's problems? We propose that a promising approach is to use metascience to track improvements in the quality of research produced by the field. The issue of defining quality in scientific research is a complex one (Shadish, 1989). Luckily, as scientists who study abstract and complex constructs, we are well-positioned to tackle this challenge with the tools and approaches provided by metascience. There is no single perfect way to measure the state of a field. The best approach is to combine a range of methods and measures that vary on a number of dimensions from narrow to broad, objective to subjective, among others, while being careful to interpret each in light of its strengths and limitations.

While metascience has been around for decades, there has recently been a rapid growth in tools and methods specifically aimed at examining the state of scientific disciplines. Below, we review some of the most relevant tools and approaches that can be used to examine the state of social and personality psychology (for a review of the broader need and value of metascience within the sciences, see Fidler & Wilcox, 2021; Hardwicke et al., 2020; Ioannidis et al., 2015). These approaches use familiar methods: observations over time, surveys of researchers, and experiments and quasi-experiments. By turning these methods on ourselves, we can empirically examine whether our field has improved, how effective our reform efforts have been, what unintended side effects there may be to these reforms, and what work remains to be done.

Observations of the Published Literature

Perhaps the most direct way to track whether the field of social and personality psychology is improving is to investigate changes in the published literature. Many researchers believe that the best way to determine quality is to personally inspect the content of articles. However, due to constraints on their time and resources, researchers often rely on less relevant but easier-to-obtain indicators of quality (e.g., journal or author reputation, impressions based on article abstracts) as stand-ins for in-depth personal inspection (Harney et al., 2021; Tenopir, 2014). Likewise, metascientists may not be able to collect experts' holistic evaluations of quality for every article in their sample. Moreover, metascientists may be interested in tracking not only quality, but other features of the published literature that can be coded more efficiently and objectively. One of the advantages of this approach is that historical data can be collected by coding existing articles, allowing us to assess change over time, and how actual practices compare to "best practices" and standards within the field. The published literature provides a window into the current norms and standards with regards to sampling, research design, statistical analysis, transparency, reproducibility, replicability, interpretations, and citations, among other practices.

Sampling

Examining who—and how many people are participating in psychological research is one way that metascience can help us track progress. For example, after years of repeated calls for greater attention to statistical power and increases in sample size, metascience studies documenting trends in reported sample size (e.g., Fraley & Vazire, 2014; Motyl et al., 2017; Sassenberg & Ditrich, 2019; Singleton Thorn, 2020) can reveal how the field has (or has not) responded to these calls. Similarly, examining how often, and how well, authors justify their sample sizes, can provide a glimpse into norms and standards in the field.

Documenting the diversity of samples (Thalmayer et al., 2021), such as the country from which participants were reportedly recruited, allows us to assess how well research in psychology represents the human population. For example, have concerns about the WEIRDness of samples in psychological science (Henrich et al., 2010) led researchers to attend more closely to the type of participants they recruit? Metascientific approaches can also be used to better understand what researchers tend to disclose and report about the people they study. For example, when and how often are the countries from which samples were recruited reported in the title or abstract of research articles (Kahalon et al., 2021)? Such work can also reveal trends and changes in how participants are being recruited, including shifts towards the recruitment of online samples (Sassenberg & Ditrich, 2019; Zhou & Fishbach, 2016). By better understanding which groups are—and importantly are not—being included in our research, we can determine how well we are doing as a field at generating knowledge that reflects the populations we claim to study.

Design

Beyond decisions about how to sample participants, researchers must make decisions about what research designs to use and how to operationalize the constructs of interest to them. Design decisions often involve trade-offs, and those trade-offs evolve with developments in technology. For example, the opportunity to collect data online may push researchers to decide between collecting larger samples with methods more easily administered online (e.g., self-reports) or collecting smaller in-person samples with more intensive methods that may be better suited to their research question (e.g., behavioral observation). Decisions about study design, which (and how many) measures and manipulations to use, and what—if any—quality checks to include can impact the results and the validity of inferences drawn. Tracking the particular designs and methods that researchers choose, and how those do or do not align with their research aims and

constructs of interest, can tell us how researchers are navigating these trade-offs. In addition to examining the suitability of the designs and methods authors select to test their research questions, we can also examine the quality of authors' specific methodological choices, such as the validity of the measures used (Flake et al., 2017).

Statistical Results

The statistical results in published papers can also provide valuable information about trends in a field. In social and personality psychology, Null Hypothesis Significance Testing is very prevalent, and the results of significance tests provide clues about the state of the field. For example, comparing the distribution of *p*-values reported in published research to what is expected under different data-generating models (e.g., high-powered research with no questionable research practices) can point to potential problems in the field. Many studies have now used these types of techniques, raising questions and sparking debate about whether or not there are more *p*-values just below .05 than we should expect, and what this might mean for the prevalence of questionable research practices (see Hartgerink et al., 2016; Krawczyk, 2015; Lakens, 2015; Masicampo & Lalande, 2012). One such tool is *p*-curve (Simonsohn et al., 2014), which can also be used to evaluate the evidential value of a set of statistical results after correcting for selective reporting. The strength of these techniques comes in part from the ability to assess and detect suspicious patterns within a group of findings that are otherwise not detectable when looking at individual studies or articles. However, because these techniques get their value from comparing observed distributions of *p*-values to expected distributions, the process of how the observed *p*-values are collected, and what assumptions underlie the expected distributions, are crucial and should be carefully specified and documented.

In-depth analysis of statistical reporting can also reveal strengths and weaknesses in the inferential practices that are commonly used in the field, including evaluating the

quality and accuracy of statistical inferences. For example, researchers have studied the published literature to estimate the prevalence of unsupported inferences drawn from mediation models (Fiedler et al., 2018) and inappropriate inferences from nonsignificant results (Aczel et al., 2018). Finally, evaluating the proportion of null-to-positive results can help reveal a potentially troublingly rate of statistically significant results in the published literature in psychology—or, at the least, spur conversation about what the optimal proportion should be (Fanelli, 2010; Scheel et al., 2021; Sterling, 1959; Sterling et al., 1995).

Statistical Errors

Another way to assess the state of a field is to look at the prevalence of statistical errors. Here, again, the homogeneity of statistical practices in social and personality psychology (i.e., reliance on NHST) presents an opportunity. Tools have been developed to detect the prevalence of basic statistical reporting errors such as statcheck (Epskamp & Nuijten, 2016) to test the coherence of NHST results (i.e., whether the degrees of freedom, test statistics, and reported *p*-values are statistically consistent with each other). Statcheck has been used to estimate and track the prevalence of statistical reporting errors in psychology articles published from 1985 to 2013 (Nuijten et al., 2016). Other tools have even broader applicability, such as the GRIM test (Brown & Heathers, 2017), which checks for inconsistencies between reported means (of integer data) and their corresponding number of items and sample sizes, and SPRITE (Heathers et al., 2018), which uses summary statistics to create plausible distributions of data that could produce the summary statistics that were reported. At a minimum, we should expect these types of errors to become less prevalent once tools are developed and available to identify them more easily, as authors and journals can use these tools to verify the accuracy of manuscripts before publication. As new tools become available, we can retrospectively assess a broader and broader range of errors. As with all tools, these can be misused or misinterpreted, and researchers should be

careful to appropriately use and apply them, particularly when strong claims or assumptions are made about the cause of these errors.

Transparency

There is likely little controversy about the direction of change in social and personality psychology with respect to transparency-related practices, in particular the open sharing of data, materials, and study preregistrations. Nevertheless, even when progress is apparent, it is important to document and quantify such change (e.g., Hardwicke, Thibault, et al., 2021; Vanpaemel et al., 2015) as empirical data can provide benchmarks for evaluating future progress or comparing the pace of progress in various subsets of the field (e.g., different research areas, journals, methods). In addition, progress likely varies across transparency-related practices. Understanding which practices are slow to change can inform decisions about where to focus our efforts, and can shed light on potential mechanisms that enable and impede change in the field. Finally, transparency is more than just sharing data, materials, and preregistration plans. We should prioritize collecting information on how our field is doing on aspects of transparency that are fundamental to scientific integrity, such as transparency about potential conflicts of interest, declaring author/contributor roles, making the peer review process more transparent, and open access to the research articles themselves, among others.

Reproducibility

The analytic and computational reproducibility of published results (i.e., whether reanalyzing the same data produces the same results) provides another way to assess the research practices and norms in our field. In order to track reproducibility, original data must be available to reanalyze, which makes this difficult, or even impossible, to systematically assess for many articles. Nevertheless, tests of analytic reproducibility have been attempted, for example, to estimate the reproducibility rates of findings published in

journals that had introduced policies and incentives aimed at increasing the sharing of original data and code (Hardwicke, et al., 2018; 2021). Computational reproducibility has also been examined in articles published as Registered Reports to test how many findings can be reproduced using publicly shared data and analysis scripts (Obels et al., 2020). Reproducibility checks require making sense of original data and analyses, and as such require attempting to clarify steps that are often left unreported in published articles. As a result, reproducibility checks often inadvertently bring to light deeper problems, such as misreported results or undisclosed flexibility (e.g., Chalkia et al., 2020).

Replicability

Of course, tracking the replicability of published findings (i.e., whether repeating a study by collecting new data produces similar results) is another way to systematically assess the state of the field. Indeed, psychology's crisis of the 2010s came to be known as the "replication crisis" in large part because many of the triggering events had to do with failed replications, most notably the "Reproducibility [sic] Project: Psychology" published in 2015 (Open Science Collaboration, 2015). Replication efforts have been used to test findings sampled from articles published in prominent journals (see Camerer et al., 2018; Open Science Collaboration, 2015) and to focus more narrowly on testing the replicability of specific effects or areas of research (e.g., Flore et al., 2018; O'Donnell et al., 2021). Large scale collaborations have been used to coordinate multi-site replications, such as through the Many Labs projects (Ebersole et al., 2016, 2020; Klein et al., 2014, 2014, 2019), the Psychological Science Accelerator (Jones et al., 2021), and Registered Replication Reports (Simons et al., 2014). These projects provide information about the replicability of individual studies, as well as informing debates about the overall rate of replicability in the field.

As with the process of testing the reproducibility of published findings, testing the replicability of past studies can reveal previously unexamined or unappreciated problems with the research process that may be affecting the credibility of psychological research. By bringing to light details of the original study's procedures and design, for example, replication efforts can help identify the often hidden day-to-day workings of research labs (e.g., how decisions about data exclusions are made, how protocols are documented) that shape our literature.

Interpretation

Another window into a field's norms is how researchers make sense of their findings and what claims and conclusions are permitted. Analyzing the interpretations that researchers draw can provide insight into what kind of evidence a field tends to require before making strong claims. A field's standards about, for example, how much hype is tolerated (or encouraged), when it is considered reasonable to make claims about policy implications, and how much evidence and what quality evidence is required before findings are popularized, speak to the field's commitment to accuracy and calibration. The words and phrases that authors use in their articles, such as the use of hedging and boosting words, may also provide some information on the degree of caution or hype researchers use when presenting their findings (Riddle, 2017).

Other indicators of calibration might include: the extent to which important caveats and limitations are stated in article abstracts, the standard of evidence (e.g., sample size and representativeness, validity of measures, robustness of results, etc.) that characterizes articles that present policy recommendations, and the degree to which authors own vs. excuse the limitations of their work (Hoekstra & Vazire, 2021; Whitcomb et al., 2017). Authors' interpretations and claims, and how well they match the quality of the evidence presented, are among the most challenging aspects of research practice to code. Thus,

perhaps unsurprisingly, we have seen few examples of metascience attempting to tackle these domains, suggesting that developing and validating tools to study these practices should be a priority.

Citation Practices

Citation practices can also shed light on the state of a field. First, we can examine whether a field engages in problematic citation practices, such as continuing to cite articles after they have been retracted (Teixeira da Silva & Bornemann-Cimenti, 2017), or citing articles whose findings have been conclusively overturned without acknowledging the new evidence (Hardwicke, Szűcs, et al., 2021; Schafmeister, 2021). Trends in these practices can alert us to potential impediments to self-correction—if a community fails to update its citation practices when the evidence changes, this suggests that publishing new and better evidence is not enough for the field to course-correct. Another potential sign of dysfunction in a field is excessive self-citations, especially when citations are highly rewarded (e.g., if self-citation is strongly associated with greater status or recognition in the field; Fowler & Aksnes, 2007).

Citation trends can also provide a measure of a field's status in the broader scientific community, including its popularity and influence. Some scholars have raised concerns that the reforms in response to social and personality psychology's crisis will make us unpopular and irrelevant (Baumeister, 2016). While citation impact should not be the measure of a field's value, it offers one way of measuring its scientific impact. Similarly, measures of impact outside of science (e.g., Altmetrics) can provide data on how much attention and engagement a field receives from society more broadly. Thus, tracking trends in citations and engagement with social and personality psychology articles can help address concerns about the potential side effects of reforms. It is also worth considering how these measures of scientific impact may be affected by improvements in the research practices of the field.

For example, if authors' claims become more calibrated with the quality of their evidence, some research may see decreases on these metrics. This may not be a bad sign, and rather could be a sign of a healthier and more intellectually humble science.

Finally, because citations and impact play such an outsized role in how researchers are evaluated, tracking trends in what is being cited can help us identify whether our incentive systems are changing, and whether we are rewarding the kinds of research we wish to reward. By examining trends in the topics, methods, and author characteristics that predict citation impact, we can gain insight into whether incentives are becoming better aligned with scientific values (e.g., if making extravagant, unwarranted claims is becoming a weaker predictor of impact over time), detect signs of bias (e.g., if authors' demographic characteristics continue to predict citation impact; e.g., Ghiasi et al., 2018), and track which practices are gaining in popularity (e.g., by examining which methods papers are being cited, and how; Simmons et al., 2018).

Beyond the Published Literature

Surveying Researchers

Another way to better understand the state of social and personality psychology is to survey the researchers working in the field to measure their attitudes, values, and opinions. For example, surveys have examined researchers' attitudes and beliefs about the field more broadly, and their perceptions of, and reactions to, the specific issues raised and reforms proposed over the last decade (for examples, see Agnoli et al., 2021; Christensen et al., 2019; Motyl et al., 2017; Toribio-Flórez et al., 2021; Washburn et al., 2018). Others have conducted in-depth interviews and ethnographies of researchers in psychology (Peterson, 2016; Peterson & Panofsky, 2020). These methods can provide valuable insight into researchers' overall impressions of their field and the developments (or crises) occurring within them.

In addition to providing a glimpse into researchers' perceptions of the field and how it is changing, survey-based approaches have been used to better understand and estimate the prevalence of specific research practices and behaviors among researchers. Perhaps most notable are surveys collecting estimates of the prevalence of questionable research practices (Agnoli et al., 2017; Fiedler & Schwarz, 2016; John et al., 2012) or transparency-related practices (Christensen et al., 2019). While most of these surveys rely on self-reports, others use indirect techniques to minimize self-report biases (e.g., Bayesian truth serum; John et al., 2012). Studies such as these can be used to estimate what researchers are doing, or at least what they say they are doing (which can be interesting in its own right), and how that varies across (sub)disciplines or over time. Surveys can also be well-suited to practices and experiences that are often hidden from the public record. For example, a well-designed survey could address the prevalence of experiences of racism or other forms of bias and discrimination, or ask researchers about their knowledge of, beliefs about, or even involvement in research fraud. Both of these problems are likely much larger than many researchers would like to think, and can be difficult to detect without asking people about their private experiences and beliefs.

Surveys have also been used to investigate—and to test—researchers' knowledge, abilities, and decision-making processes. Such surveys have assessed researchers' (mis)understanding of statistics, including intuitions about statistical power (Bakker et al., 2016) and the prevalence of the misinterpretation of confidence intervals (Hoekstra et al., 2014). Other surveys study the research process, assessing the types of decisions that researchers make. For example, after giving researchers the same dataset and research questions to investigate, the Many Analysts project was able to document the diversity of different operational and analytic choices that researchers made in response to the same information and goal (Silberzahn et al., 2018). These techniques, repeated over time, can

provide valuable insights into whether researchers' skills and decision-making processes are improving.

Assessing Interventions

Other approaches have sought to evaluate how changes in journal policies and publishing models impact researchers' behaviors and the characteristics of the research being published. Ideally, this would be done with experiments, but these are often impossible or impractical to conduct. However, careful reasoning from observational studies can also shed light on how effective such interventions may be at bringing about the desired change. For example, researchers have considered the potential effects of transparency-related journal policies by tracking the number of articles published in *Psychological Science* that had (or claimed to have) open data and open materials before and after the journal introduced "badges" that could be earned for open practices (Kidwell et al., 2016; c.f. Bastian, 2017; Rowhani-Farid & Barnett, 2020). Another study compared the availability and computational reproducibility of data published in *Cognition* before and after the adoption of an open data policy (Hardwicke et al., 2018).

The introduction of Registered Reports is an area ripe for considering how publishing models shape the quality and characteristics of the literature published in a field. In Registered Reports, authors typically submit a Stage 1 manuscript before data collection, outlining their research question and study design and analysis plan. The journal reviews this manuscript and can give the authors receive an "in principle acceptance" that commits the journal to publishing the article regardless of the results and commits authors to following the plan in the Stage 1 manuscript. This model should, in principle, reduce the opportunity for study results to be influenced by bias on the part of authors or reviewers/editors.

Several studies have sought to empirically compare research published under the Registered Reports model versus the traditional model. For example, Scheel et al. (2021) compared the prevalence of positive results in articles published as Registered Reports and articles published under the more traditional model of peer review. Another study investigated researchers' evaluations of papers when blinded to whether they were or were not published as Registered Reports (Soderberg et al., 2021), assessing not just perceptions of rigor, but also creativity and novelty. As the number of articles published under this model increases, even more work will be possible to assess how alternative publishing models are being used and how they are shaping the state of the published literature.

Integrating Approaches

Together, these metascientific approaches offer a window into what was happening in the past, what is happening now, and what might help to address our problems in the field of social and personality psychology. Each of the approaches can provide valuable evidence that, when considered carefully, can help constrain the range of possible models of our field and shape the next steps for the field: decisions about research training and best practices, policies for journals and funders, allocation of resources and critics' attention, etc. Of course, all of these decisions will depend on value judgments and priorities—they cannot be made on the basis of metascientific evidence alone. But evidence will help, and without rigorous metascientific evidence, we will be in a much worse position to make these decisions.

Most of the existing metascientific studies applying one of the approaches described above examine a relatively small sample of published articles and only one or two variables (e.g., an inspection of sample sizes in 824 articles). This relatively narrow focus is understandable given the effort involved with coding each article for each variable of interest. Moreover, many of these early metascientific papers have served the dual purpose

of introducing a method for metascientific analysis, and applying it to a narrowly-defined problem. However, by looking at many different aspects of the published literature in isolation, we may be missing important relationships among these variables, or trends in how these variables are related to each other. Now that these methods and tools have begun to be developed, the time is ripe to adopt them on a larger scale. As such, we hope to see more research that combines these approaches to compare and triangulate information collected across different measures and lead to new insights about the state of the field.

A thorough reckoning with our situation requires not only looking in individual nooks and crannies, as metascientific studies using a single approach do, but mapping out the full landscape and taking stock of the larger picture. Combining a broad range of measures will allow us to obtain a more complete picture of the state of social and personality psychology, and ask questions that cannot be addressed with any single approach. For example, if researchers have shifted their practices to meet rising expectations for sample size, they may have compensated by saving resources elsewhere, such as avoiding behavioral or physiological methods even when they would maximize construct validity, or conducting a study online even when that provides less experimental control than in-person settings. To detect such dynamics, trends in all of these variables need to be measured together, and ideally these should be complemented with surveys or interviews asking researchers about their decision-making process when designing their studies.

Understanding how trends in research practices covary can also provide fodder for debates about potential side effects and unintended consequences of reform, though of course conclusive causal evidence would be rare. Empirical evidence could help constrain the range of plausible explanations, however, and would provide concrete information for cost-benefit analyses. Formal models and simulation-based approaches can also provide

valuable insight to guide future investigations in these areas. Ultimately, we may not be able to conclusively answer whether, for example, the push for greater rigor is leading social and personality psychologists to produce boring, unimportant research that receives little attention (Baumeister, 2016), but such debates can be made more tractable with a combination of the approaches described above.

In our research group, we are currently undertaking such a "kitchen-sink" approach to empirically assess the state of the field—a project we call Surveying the Past and Present State of Published Studies in Social and Personality Psychology (SPPSPSSPP; in case it is not obvious, the acronym is poking fun at our field's obsession with acronyms consisting of S's and P's.) Our project assesses the state of the published literature in social and personality using a broad range of methods, including many of the methods described here. Our corpus includes over 8,000 social and personality psychology articles published between 2010 and 2020 across seven journals: *Collabra: Psychology*, *Journal of Experimental Social Psychology*, *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *PLOS One*, *Psychological Science*, and *Social Psychological and Personality Science*. This project has been the main focus of our research team, including a dozen or so collaborators and over 50 research assistants, for the last two years. This project is still underway, but our efforts have already provided many valuable lessons, for example, about what is easy and hard to measure, and where there are gaps in the existing tools and methods. Our experiences with this project have informed the agenda we have presented here.

The design of the SPPSPSSPP project aims to assess the literature on a broad range of qualities and features beyond replicability (Vazire et al., in press) including the "four validities" (Shadish et al., 2002) To do this, we make use of many of the approaches described above. The observational coding of the literature includes attempts to code

aspects of the articles' samples (e.g., sample size, sample size justification, population sampled), design and methods (e.g., experimental vs. observational, between- vs. within-persons design, type of method used), statistical analyses and results (e.g., *p*-curve, null results as primary finding), statistical errors, transparency-related practices (e.g., claims of open data, materials, and preregistration, declarations of conflicts of interest), interpretation (e.g., limitations reported, hedging and boosting language), and more. Of course, we could not code every variable for all articles in our sample. We were able to code every article for variables that are relatively easy to code (e.g., sample size, type of participants recruited), but are coding only a subset of articles for variables that required more intensive scrutiny to code (e.g., p-curve, limitations reported). In some cases, we gave up on variables that proved too difficult to code well (e.g., causal claims in abstracts). Where possible, we are attempting to develop and validate semi-automated tools to code some variables (e.g., link to open data, open materials, or preregistration), and to link our dataset to existing metadata about the same articles (e.g., citation impact and Altmetric data). We also selected a subsample of articles and recruited experts to read the full article and rate the quality of the research on various dimensions (e.g., the four validities, novelty, interestingness). Finally, we surveyed authors of published articles to capture their views and attitudes about norms and practices in the field, and how these have changed from 2010 to 2020.

Despite all of our resources and efforts, the SPPSPSSPP project will be flawed. Our measures will be imperfect, our sample of articles and researchers will be unrepresentative, our ability to draw conclusions about causal mechanisms will be limited, and we will have missed important variables that should be considered. Moreover, we are just one team of researchers, with significant conflicts of interest and very particular positions with respect to our subject matter. We are not dispassionate observers. Unfortunately, we are pessimistic

that dispassionate metascientists would invest so much effort in evaluating the state of one subdiscipline.

Fortunately, several other research groups in social and personality psychology are also tackling this challenge. Several recent projects looked specifically at the state of the published literature in social and personality psychology, and how it is changing over time, including Motyl and colleagues (2017) and Sassenberg and Ditrich (2019). These projects follow in the footsteps of similar projects in the history of social and personality psychology (e.g., Fisch & Daniel, 1982; Fried et al., 1973; Higbee & Wells, 1972; Quiñones-Vidal et al., 2004; Reis & Stiller, 1992; Sherman et al., 1999; West et al., 1992). These projects vary in quality and in scope, and, when assessing the value of these projects as well as ours, it is worth noting that bad metascience is worse than no metascience. In light of the serious concerns about our field's credibility, and the relative dearth of empirical evidence regarding the recent state and trajectory of our field, we hope that other researchers, with diverse opinions, values, and expectations, will contribute to these efforts to rigorously evaluate our field.

Conclusion

We have argued that if we want to know if the field of social and personality psychology is producing credible claims, and identify where we still need to improve, a good place to start is to study the state of the published literature. The published literature constitutes the majority of the outputs of our field, and also reflects the downstream outcomes of researchers' practices and of structural processes and incentives. Thus, evidence regarding the state of the published literature would provide valuable information regarding whether we are reckoning with our crisis. Surveys of researchers and tests of interventions would also help to provide a more complete picture of the field, and potentially identify hidden practices and norms.

Another important part of reckoning with our crisis, however, should be improving structural systems in our field, such as journals, peer review, university and department governance, hiring and promotion, training, awards, and professional societies. Many of these systems are inefficient, prone to error and bias, leave themselves open to corruption and exploitation, have few mechanisms for accountability or even transparency, and are designed to reinforce the status quo (e.g., by giving an outsized voice to a small group of successful researchers; for examples, see Bakker et al, 2021; Bol et al., 2018; Edwards & Roy, 2016; Gross & Bergstrom, 2019; Larivière, & Sugimoto, 2019; Smaldino & McElreath, 2016). Until we change these systems, changes to individual research practices and outputs are unlikely to be enough. Moreover, these structures are meant to provide safeguards against crises of credibility (e.g., by ensuring equity, fairness, and quality control). Until we fix the structural problems, we leave ourselves vulnerable to future crises.

One example of the field's dysfunction at a structural level is the lack of incentives, and even disincentives, for doing metascientific research. Although systematic data on this issue is lacking (which itself could be considered evidence of this problem), we contend that the field of social and personality psychology does not sufficiently value metascientific research on the field. This includes research on replication, error detection, and empirical audits of the literature. This is likely not specific to social and personality psychology—it is probably rare for an academic field to reward its own critics with jobs, grants, and awards. However, we would argue that recognizing and incentivizing those who rigorously assess the state of the field and point out problems would be one the best ways to demonstrate a commitment to self-correction. In many contexts, a group that is committed to quality control and accuracy might hire experts to "red team" their ideas and outputs to help find problems and weaknesses so that they can be addressed (Red Team Market, 2020).

Academic departments, funding agencies, and professional societies should similarly invest a proportion of their resources in metascientific research.

Metascientific studies can help identify the ways in which the field of social and personality psychology is doing well, ways in which it is improving, and ways in which it is falling short. Identifying these issues, and tracking progress, provides valuable information about the state of the field. By investing in empirical investigations into the state of our field, we can demonstrate our commitment to self-correction, and our willingness to hold ourselves accountable for fixing the problems that the last decade's crisis exposed. Moreover, the empirical findings from these metascientific investigations will help guide our approach to policies and interventions that address the structural problems in our scientific ecosystem that created the crisis in the first place. What we do now will determine how we look in the history books, and in the public's eye. The next decade is an opportunity for the field of social and personality psychology to demonstrate its commitment to self-correction.

Chapter 2: A Consensus-Based Tool for Evaluating Threats to Validity

When a researcher agrees to review a manuscript for a scientific journal, what should they focus on? Many journals give reviewers little guidance regarding what dimensions to consider when reviewing a manuscript (Hirst & Altman, 2012). Most researchers likely have an intuitive sense of what makes for a valuable contribution, but when we break down the many different aspects of research evaluation, it becomes clear that evaluating scientific papers is a very complex task. Moreover, reviewers are often overburdened and underpaid. Thus, providing support during the review process could potentially improve the quality of peer reviews. Tools like checklists are one such source of support, directing reviewers' attention to important aspects of research evaluation.

Although the use of checklists has been criticized, even experts with extensive training and experience routinely benefit from the support provided by checklists—including, for example, pilots, astronauts, and surgeons. A number of checklists have been developed for scientific peer review, particularly in the health and biomedical sciences. Although there is relatively little empirical research on the peer review process in science (Tennant & Ross-Hellauer, 2020), the use of checklists has shown some promise. For example, an exploratory study conducted with the *British Medical Journal* found adding a checklist to evaluate methodological and statistical features increased the quality of the studies published (Gardner & Bond, 1990). Editorial use of CONSORT and STROBE checklists has also been found to lead to small improvements in the quality of manuscripts (Cobo et al., 2011). However, other studies testing the impact of guidelines and checklists have found no significant effects on quality assessments (Cobo et al., 2007; Jefferson et al., 1998). More recently, methodological reporting in *Nature* papers was found to have improved following the introduction of a checklist into the article submission process (The NPQIP Collaborative Group, 2019; Han et al., 2017).

Checklists and guidelines have yet (Hirst & Altman, 2012) to be widely implemented in peer review. For example, in a survey of 116 health research journals, only 19 journals pointed reviewers to reporting guidelines such as CONSORT (Hirst & Altman, 2012). However, many resources have been developed, such as the transparency checklist (Aczel et al., 2020), APA's Journal Article Reporting Standards (Appelbaum et al., 20180118), and the EQUATOR Network's online toolkit (The EQUATOR Network, 2018) to help reviewers locate the reporting guidelines appropriate for the research they are reviewing. Most checklists and reporting guidelines focus on transparency or completeness of reporting (Davis et al., 2018; Valentine & Cooper, 2008). This is understandable as transparency is a prerequisite for evaluating quality (indeed, the EQUATOR website tells reviewers that their reporting guidelines "will help you decide whether a research manuscript contains enough detail to judge its quality.") However, fewer tools narrow in on helping reviewers evaluate the quality of the inferences and claims made in empirical articles.

Identifying Potential Threats to Validity

One fundamental dimension of research quality is the validity of the research methods, design, and analyses for the inferences drawn. Does the study provide an adequate test of the research question? Are the data analyzed appropriately? Do the conclusions match the evidence? The issue of validity is arguably at the heart of research quality and should be one of the central foci of research evaluation during peer review. Ensuring that researchers draw valid inferences from their studies is one of—if not the—most important functions of peer review. Journals may also expect reviewers to discern other qualities, such as transparency, novelty, fit to the journal's scope, potential impact, or applied value, but a journal that directs reviewers not to weigh any given dimension on this list could still be considered to be conducting peer review. A journal that directs reviewers not to weigh the

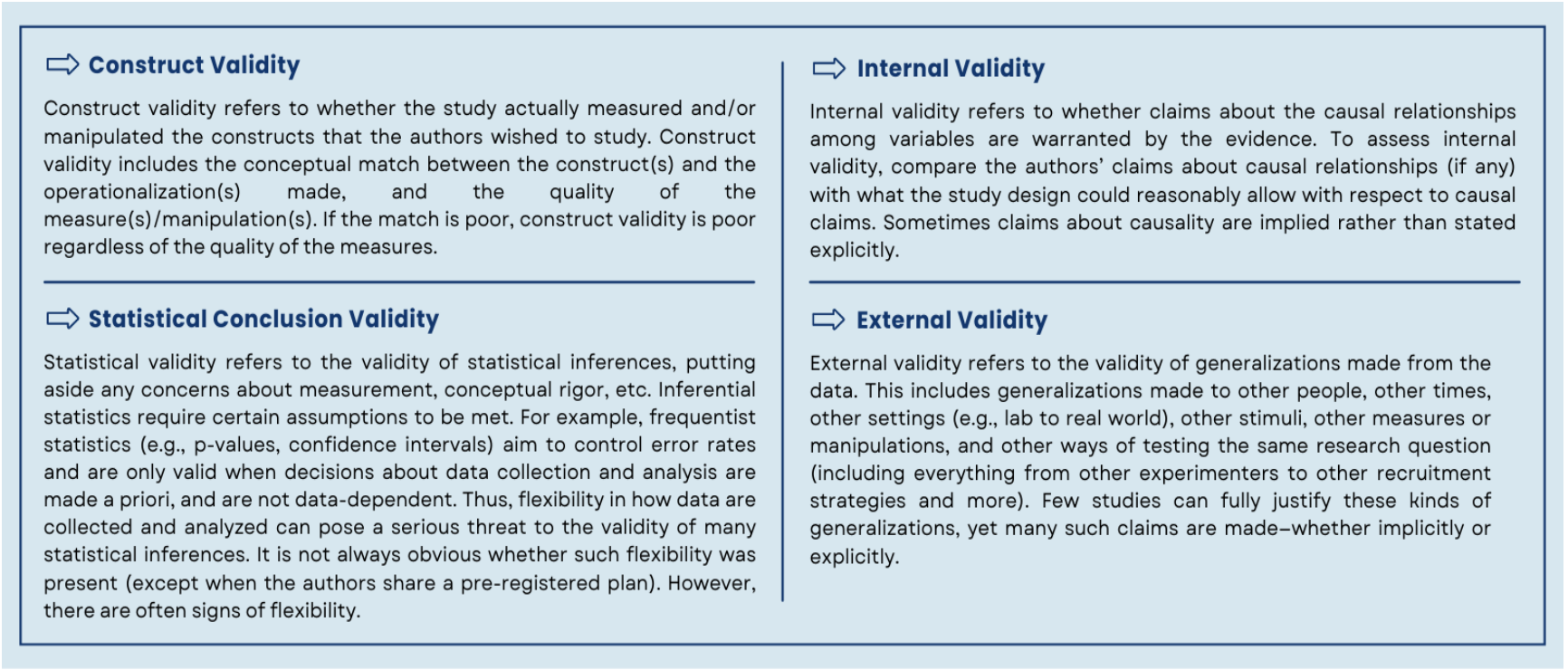
validity of the inferences drawn by the authors is arguably no longer engaged in scientific peer review.

Perhaps one of the reasons that there are few tools available to help researchers evaluate the validity of scientific papers is that judgments of validity (or threats to validity) are remarkably complex. Validity is often far less straightforward to evaluate than dimensions such as transparency or reporting completeness. Moreover, judgments of validity depend a great deal on subtle contextual factors (e.g., a measure or operationalization that is valid in one context may not be valid in another). These factors make developing a checklist for validity threats extremely challenging. However, it is because evaluating validity threats is so arduous and important that we believe reviewers would benefit from tools to help them navigate this complicated terrain.

To aid reviewers in evaluating threats to the validity of empirical research, we developed Seaboat, an online tool (available at seaboat.io) using the "four validities" framework (Shadish et al., 2002) popular in the social sciences. The four validities are: 1) construct validity, the validity of inferences about constructs that are measured or manipulated (i.e., validity of operationalizations), 2) internal validity, the validity of causal inferences, 3) external validity, the validity of generalizations made (e.g., to other people, settings, times, measures, stimuli), and 4) statistical conclusion validity, the validity of statistical inferences (for more detailed definitions, see Figure 1). Although we acknowledge this is one among many plausible models of validity, and the four categories sometimes overlap, we believe it provides a useful framework and one that is likely to resonate with reviewers in the social sciences (even if they are unfamiliar with the explicit framework or labels). We designed this tool to be used in evaluations of quantitative empirical papers in

Figure 1

The Four Validities



psychology, however researchers and reviewers in nearby disciplines may find it useful or easy to adapt[1].

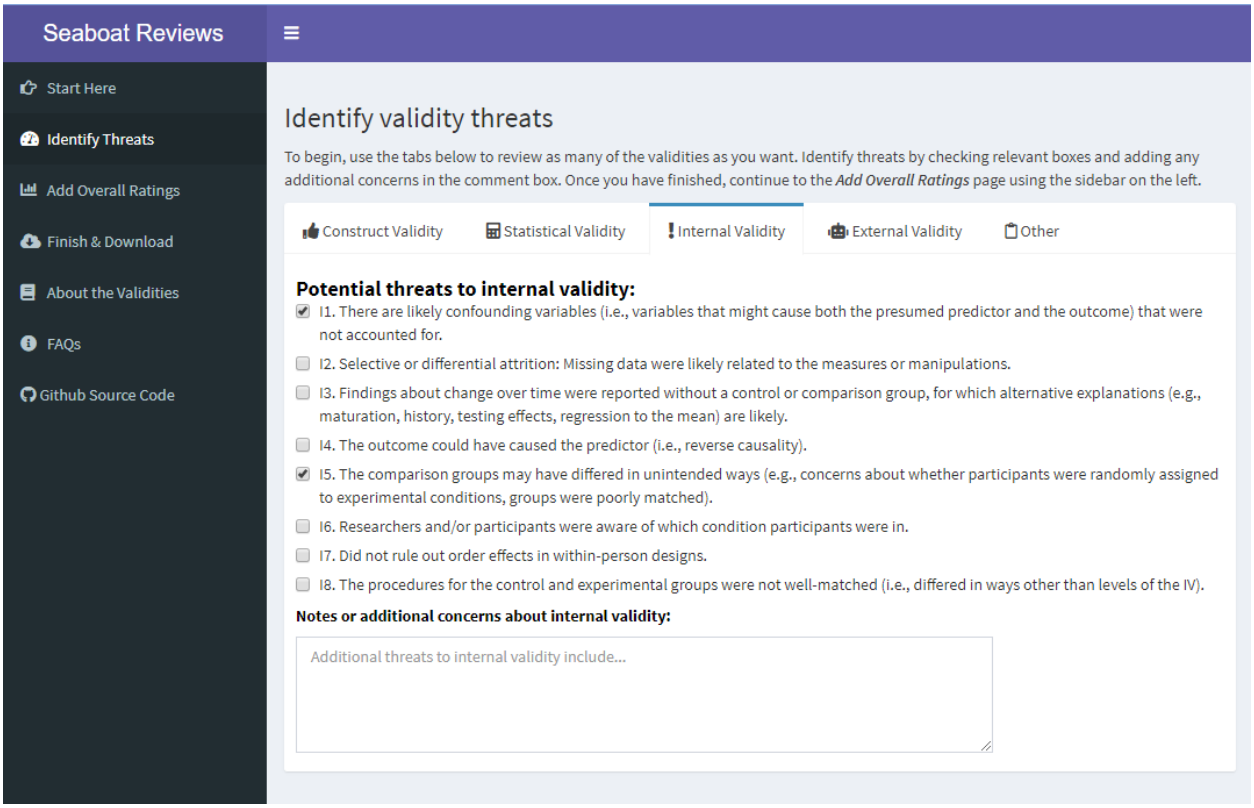
To select the validity threats to include in this tool, we used a 'reactive-Delphi' expert consensus process in which over 50 experts provided feedback across multiple rounds of item generation and refinement. Additional details about the method and results are available at osf.io/6rfu9 and in the supplemental materials. Our preregistered protocol guided our decisions about which threats to retain, revise, or remove. Our aim was to select a manageable number of threats in each of the four validity categories, prioritizing threats that are most common and most serious in psychological research. The final list includes 32 potential threats to validity (5 to construct validity, 8 to internal validity, 6 to external validity, and 13 to statistical conclusion validity).

To use this tool, users may visit seaboat.io where they can begin a report to evaluate threats to validity and access resources to learn more about validity threats and the four validities framework. Users can navigate back and forth across the four validities to identify potential threats, specify additional threats to validity that are not listed, and elaborate in comments as needed (see Figure 2). After identifying specific threats, users are invited to rate the paper on each of the four validities on a Likert-type response scale. Here, users are reminded that their global rating on each validity need not correspond to the number of specific threats identified—any level on the global rating can in principle be consistent with any number or combination of specific threats. Finally, users are given the option to download a report of their response in a variety of formats (e.g., PDF, html, Word doc). This report can then be included in a narrative review for a journal, shared privately, or posted as a comment on a publicly available paper.

[1] Code is available at github.com/schiavone1/seaboat.

Figure 2

Seaboat.io User Interface



Note. Example of selecting potential threats to internal validity on Seaboat.io.

Caveats and Limitations

First, this tool does not provide a comprehensive set of criteria for evaluating research quality. Validity is but one aspect of quality that reviewers and consumers of research should consider—papers that make valid claims are not necessarily high quality in other ways. Conversely, even papers with serious validity threats can make valuable contributions when claims are calibrated.

Second, the threats described in this tool are not exhaustive. We aimed to find a balance between covering common threats and keeping the list manageable and approachable to the typical reviewer. Many threats unique to particular methods and analyses were not included, as were threats believed could be too easily misunderstood or

misapplied. Thus, users should not rely upon this app to cover all potential threats—or even all major threats—to the validity of the research being evaluated. Users are encouraged to add additional threats they identify when using this app.

Third, this tool could be misinterpreted or misused if users treat validity as the sum of its parts, for example by simply counting the number of threats identified as a measure of a paper's validity. Not all validity threats are equally damning or cause for concern. Papers with fewer identified threats are not necessarily more valid, and the same threat identified in two papers may have very different implications for the validity of the claims. Thus, threats should be considered in the context of the whole paper.

Fourth, this tool is intended for users with a background knowledge of research methods as the validity threats described assume some existing expertise. Users do not need to be familiar with the 'four validities' framework, nor with the labels used in this framework. We believe many researchers in the social sciences will recognize the potential threats included, even if they typically use slightly different language to describe them. Nevertheless, users should use their judgment when deciding what rises to the level of a threat to validity, which will require some training in research methods and critical thinking. Users with little training or no expertise in research methods will not find enough support within this tool alone to find it useful.

Fifth, this tool is not meant to replace narrative reviews or other existing peer review structures. While we see some benefits to structured reviews (e.g., can be shared publicly with fewer concerns about confidentiality, are easy for a wide range of readers to understand and benefit from, and can be aggregated/analyzed quantitatively), narrative reviews provide unique information and richness that cannot be captured by checklists and rating scales. Both can play an important role in research evaluation.

Finally, we have yet to collect empirical data to evaluate the usefulness of this tool. As such, we expect to continue to make improvements as we learn more about what features and changes would improve user experience and enhance the tool's effectiveness. That said, drawing attention to concerns about validity is unlikely to do harm or be worse than many current peer review processes which often offer little-to-no guidance to reviewers to inform their evaluations of research quality.

Recommended Uses

Reviewers

Our primary audience for this tool is reviewers evaluating papers for journal-based peer review. We suspect many reviewers would welcome tools to help make their peer review process more systematic and help them make sure they have thoroughly considered any concerns that may threaten the conclusions drawn. We anticipate that reviewers will find this tool helpful regardless of whether a journal explicitly encourages attention to validity threats or not. The tool helps reviewers more easily assess validity threats and communicate concerns about validity by including their customized report generated by the app alongside their narrative reviews.

Editors / Journals / Publishers / Societies / Funders

We encourage editors and those who set journal policies, to consider how tools (such as this and others) may improve their review process. Journals could offer these as optional resources for reviewers, conduct experiments to evaluate the effect of adding these tools, or directly incorporate them as required steps in their peer review process. This could not only elicit more consistent (and potentially more thorough) feedback but provide editors (and readers in the case of transparent peer review) a clearer picture of what reviewers considered in their evaluations. Funders and societies could similarly use such tools in their own evaluation and review processes.

Informal Research Evaluation

Evaluating potential threats to the validity of scientific research is also vital in many contexts outside of journal peer review where we hope this tool could be useful. For example, researchers can use it in post publication peer review by sharing their reports openly online (on blogs or platforms such as PubPeer or hypothes.is). We imagine this tool would also be useful in journal club settings to prompt discussion and to compare judgements about validity threats among readers.

Training

The tool could be used in graduate and undergraduate courses, workshops, and other training contexts. For example, journal editors can use the tool to provide reviewer training to their editorial boards, mentors when engaging their trainees in co-reviews, and authors to document and share their evaluations of their own research (e.g., when drafting their paper, to make sure their claims are well-calibrated). Researchers and labs could also use this tool when reviewing their own protocols prior to preregistration or data collection as an exercise to identify potential weaknesses in their studies.

Method and Results

The validity threats included in this app were developed using a reactive Delphi method, an iterative process of collecting and integrating the ratings and feedback of experts to determine the specific items for inclusion. This procedure was modeled after the process used in the development of the Transparency Checklist (Aczel et al., 2020), and allowed us to select a final list and wording of items based on expert consensus. To do so, we collected three waves of data. Study materials, code, and preregistration for the three waves are available at osf.io/6rfu9.

Expert Reviewers

We identified and invited 121 experts to review the initial items developed. These reviewers were selected based on their expertise in research methods and/or peer review and represent numerous fields and areas of expertise in psychology (social and personality, cognitive, quantitative, clinical, industrial-organizational, etc.) and beyond (e.g., statistics, metascience, philosophy of science, error detection). Many (if not most) also have extensive experience serving as editors of peer-reviewed journals. Reviewers who participated in this study were given a $25 Amazon gift card. See acknowledgements for a list of expert reviewers who participated and agreed to their name being listed.

Wave 1

Procedure

Reviewers were contacted by email and sent a link to the questionnaire. The questionnaire presented reviewers with a brief description of the purpose of the study and provided brief definitions of each of the four validities. Reviewers were then asked to read 51 items (presented in separate groups for each of the four validities) and indicate the degree to which they thought each item should be included (or excluded) in the app as a threat to that validity, using a Likert-type response scale from 1 (*Definitely exclude*) to 9 (*Definitely include*). When rating whether items should be included, reviewers were asked to consider how *common* and how *serious* each threat is to the validity of quantitative research in psychology. For each item, they were also invited to provide any feedback or suggestions for improving the wording of the item. At the end of the list of items for each of the four validities, reviewers were asked to provide any general comments on the items for that validity and any suggestions for additional items.

Results and Item Revision

59 experts completed Wave 1. Along with quantitative ratings, we received over 27,400 words of open-ended feedback, and we are grateful to the reviewers for the helpful comments and suggestions. After reviewing item ratings, wording suggestions and comments, recommendations for additional items, and general feedback about the project, the authors discussed and revised the items. As preregistered, we did not have specific inclusion criteria and item ratings were used in combination with open-ended feedback to generate and refine the item pool to be used in Wave 2. From the original list, we excluded 21 items, revised 28, and added 9, resulting in a final list of 49 items for Wave 2.

Wave 2

Procedure

The same 121 expert reviewers contacted in Wave 1 were invited to participate in Wave 2. As preregistered, we ended data collection 15 days after contacting and inviting reviewers to participate. Reviewers were provided a brief summary of the changes that had been made based on the feedback received in Wave 1 and the number of items that they would be asked to rate for each validity. As in Wave 1, items were rated using a Likert-type response scale from 1 (*Definitely exclude*) to 9 (*Definitely include*). Unlike Wave 1, reviewers were not asked to provide open-ended feedback or wording suggestions for each item. Rather, they were invited to note any feedback in a general comment box for each validity, or at the end of the survey.

Results and Item Revision

56 experts completed Wave 2. As preregistered, we defined consensus among reviewers for including an item in the final version as median item rating of 7 or higher (on the 9-point response scale) and interquartile ranges of 2 or smaller. Overall, 29 of the 42 items (69.5%) met these criteria: 4 items (of 8; 50%) on construct validity items, 11 items (of

19; 57.89%) on statistical validity, 8 items (of 9; 88.88%) on internal validity, and all 6 items (100%) on external validity.

We examined items that did not meet the criteria for inclusion and reviewed all comments provided by the reviewers. Based on the feedback collected, the authors discussed and agreed upon a set of revisions. Of the 13 items that did not meet the criteria, 9 items were excluded (2 from construct validity, 6 from statistical validity, and 1 from internal validity) 3 items were reworded (1 from construct and 2 from statistical validity), and 1 item was retained in its original wording (from construct validity). In Wave 3, we presented reviewers with only the 4 remaining items to evaluate.

Wave 3

Procedure

Reviewers that completed Wave 2 were invited to participate in Wave 3 to evaluate the four revised items. As preregistered, we ended data collection 21 days after contacting and inviting reviewers to participate. Reviewers were provided a brief summary of revisions and asked to rate the revised items. For each of the four items, reviewers were presented the original and revised wording (or the unrevised wording for one item), along with a histogram displaying the distribution of ratings for the original item in Wave 2. They were then asked to rate each revised item from 1 (*Definitely exclude*) to 9 (*Definitely include*) and note any comments at the end of the survey.

Results

46 experts from Wave 2 completed Wave 3. Using the same preregistered criteria as in the previous wave, 3 of the 4 items reach consensus (1 from construct and 2 from statistical validity). The list of items ultimately retained and their corresponding final ratings from expert reviewers are reported in Table 1.

Table 1

Final List of Potential Threats to Validity Included in Seaboat.io

Item	Median	IQR	Mean
Construct Validity			
C1. Construct(s) were poorly defined (e.g., inconsistent and/or unclear definitions).	9	1	8.46
C2. Insufficient information provided about how the constructs were operationalized.	9	1	8.16
C3. Reliability of measures is not considered or issues with reliability are overlooked.	8	2	7.79
C4. Measures or manipulations (or how they were administered) likely introduced error (e.g., demand characteristics, social desirability, inattentive responding).	8	2	7.38
C5. Insufficient evidence of, or attention to, the validity of measures or manipulations.	9	1	8.58
Statistical Validity			
S1. Insufficient information provided about the analyses to evaluate (or reproduce) the results.	9	1	8.64
S2. Low statistical power or precision to detect the effect of interest (e.g., small sample size/number of observations).	9	2	7.91
S3. Unknown or unclear stopping rules for data collection.	8	2	7.44
S4. Data exclusions were not sufficiently justified, outliers were treated inconsistently between similar studies, an unusually large number of observations were excluded, or it was otherwise unknown how exclusions impacted the results.	9	2	8.05
S5. Insufficient safeguards against flexible analysis decisions (e.g., dropping items or measures, transforming variables, haphazard inclusion of controls variables). Safeguards could include detailed preregistrations, direct replications, or robustness checks. Reporting results as exploratory with no confirmatory statistics or interpretations could also mitigate these concerns.	9	2	7.85
S6. Poor match between substantive hypothesis and statistical test.	9	1	8.23

Table 1 (Continued)

Final List of Potential Threats to Validity Included in Seaboat.io

Item	Median	IQR	Mean
Statistical Validity			
S7. Overinterpreted statistically ambiguous results (e.g., "marginally significant", "trending towards significance") or statistically unlikely results (e.g., a series of statistically significant results across studies, with none or few of the p-values below .01. True effects should produce heavily skewed distributions of p-values, with most p-values below .01.)	8	2	7.42
S8. Treated dependent observations (e.g., data clustered within persons, groups, countries) as independent; did not account for interdependence.	8	2	7.57
S9. Overinterpreted statistically ambiguous or uncertain results as significant/meaningful (e.g., "marginally significant", "trending towards significance").	7	2	7.38
S10. Interpreted results as evidence of "no difference" or "no effect" based only on non-significant p-values without directly testing for evidence of absence (e.g., using equivalence testing or Bayesian statistics).	8.5	2	7.67
S11. Failed to directly test or estimate the effect of interest (e.g., reporting that two effects differ because one is statistically significant and the other is not, rather than testing the interaction directly).	9	2	7.81
S12. Interpreted results as support for a hypothesis even though the pattern of results was not as predicted (e.g., hypothesized an interaction where cell A1 would be lower than cells A2, B1, and B2, but the pattern of results does not look like that).	8	2	7.73
S13. HARKing (Hypothesizing After the Results are Known): The results were presented as "predicted", without documentation of a priori predictions (e.g., preregistration) and there is reason to doubt this (e.g., theory is vague, result focuses on subgroup analysis).	9	2	7.93
Internal Validity			
I1. There are likely confounding variables (i.e., variables that might cause both the presumed predictor and the outcome) that were not accounted for.	9	1	8.29
I2. Selective or differential attrition: Missing data were likely related to the measures or manipulations.	8	2	7.91

Table 1 (Continued)

Final List of Potential Threats to Validity Included in Seaboat.io

Item	Median	IQR	Mean
Internal Validity			
I3. Findings about change over time were reported without a control or comparison group, for which alternative explanations (e.g., maturation, history, testing effects, regression to the mean) are likely.	8	1.25	7.86
I4. The outcome could have caused the predictor (i.e., reverse causality).	8	1	7.91
I5. The comparison groups may have differed in unintended ways (e.g., concerns about whether participants were randomly assigned to experimental conditions, groups were poorly matched).	8	2	7.89
I6. Researchers and/or participants were aware of which condition participants were in.	8	2	7.25
I7. Did not rule out order effects in within-person designs.	7	2	7.07
I8. The procedures for the control and experimental groups were not well-matched (i.e., differed in ways other than levels of the IV).	8	2	7.49
External Validity			
E1. The authors did not make it clear to what range of people, settings, measures, etc. they believe their findings do and do not generalize.	9	2	7.84
E2. Important sample characteristics were not reported (or measured).	8.5	2	7.86
E3. Claimed—or strongly implied—that their results generalize to populations that the sample did not adequately represent.	8	1	8.05
E4. The sample recruited did not match the population of interest (e.g., sampled only college students when the research question was about psychiatric populations).	9	2	7.84
E5. Claimed—or strongly implied—that effects generalize beyond the specific measures, manipulations, or settings sampled when the design does not allow for such generalization.	8	2	7.66
E6. Claimed implications for real-world phenomena far beyond what was studied.	8.5	2	7.84

Acknowledgements

I am very grateful to Dave Kenny who provided extensive help in developing the original validity threats used in Wave 1 and my co-authors Kimberly Quinn and Simine Vazire. We thank the following expert reviewers (and all anonymous reviewers) for participating in the process of developing and refining the specific threats to validity included in Seaboat.io: Anna Alexandrova[1], Ruben Arslan[2], Marjan Bakker[3], Adrian Barnett[4], Rene Bekkers[5], Jan R. Boehnke[6], Violet Brown[7], Nick Brown[8], Benjamin Brown[9], Rickard Carlsson[8], Felix Cheung[10], Brian S. Connelly[11], Katherine S. Corker[12], Pam Davis-Kean[13], Brent Donnellan[14], Phoebe C Ellsworth[13], R Chris Fraley[15], Willem E. Frankenhuis[16], Antonio Freitas[17], Eiko Fried[18], Andrew Gelman[19], Roger Giner-Sorolla[20], Gregory R. Hancock[21], James Heathers[22], Emma Henderson[23], Alex O. Holcombe[24], Maxwell Hong[25], Yoel Inbar[10], Chick Judd[26], Laura A. King[27], Kevin King[28], Daniël Lakens[29], Steve Lindsay[30], Rich Lucas[14], Thomas E. Malloy[31], Scott Maxwell[25], Cynthia Mohr[32], Don Moore[33], Beth Morling[34], Dominique Muller[35], Brian Nosek[36], Michèle B. Nuijten[3], Amy Orben[1], Fred Oswald[37], Cort W. Rudolph[38], Cristian Larroulet Philippi[1], Kate Ratliff[39], Don van Ravenzwaaij[40], Mijke Rhemtulla[41], Brent Roberts[15], Julia M. Rohrer[42], Cort Rudolph[38], Liam Satchell[43], Victoria Savalei[44], Laura Scherer[45], Patrick E. Shrout[46], Joseph Simmons[47], Leonard Simms[48], Julia Strand[49], Moin Syed[50], Louis Tay[51], Elina Vessonen[52], Gregory D. Webster[39], Jelte M. Wicherts[3], Brenton M. Wiernik[53], and Ethan Young[16].

[1]University of Cambridge, Cambridge, England. [2]University of Leipzig, Leipzig, Germany. [3]Tilburg University, Tilburg, Netherlands. [4]Queensland University of Technology, Queensland, Australia. [5]Vrije Universiteit Amsterdam, Amsterdam, Netherlands. [6]University of Dundee, Dundee, UK. [7]Washington University in St. Louis, MO, USA.

[8]Linnaeus University, Växjö, Sweden. [9]Georgia Gwinnett College, GA, USA. [10]University of Toronto, Toronto, Ontario, Canada. [11]University of Toronto Scarborough, Toronto, Ontario, Canada. [12]Grand Valley State University, MI, USA. [13]University of Michigan, MI, USA. [14]Michigan State University, MI, USA. [15]University of Illinois at Urbana-Champaign, IL, USA. [16]Utrecht University, Utrecht, Netherlands. [17]Stony Brook University, NY, USA. [18]Leiden University, Leiden, Netherlands. [19]Columbia University, NY, USA. [20]University of Kent, Canterbury, UK. [21]University of Maryland, Maryland, USA. [22]Cipher Skin, CO, USA. [23]University of Surrey, Guildford, England. [24]University of Sydney, Sydney, Australia. [25]University of Notre Dame, IN, USA. [26]University of Colorado Boulder, CO, USA. [27]University of Missouri, MO, USA. [28]University of Washington, WA, USA. [29]Eindhoven University of Technology, Eindhoven, Netherlands. [30]University of Victoria, British Columbia, Canada. [31]Rhode Island College, RI, USA. [32]Portland State University, WA, USA.[33]University of California, Berkeley, CA, USA. [34]University of Delaware, DE, USA. [35]Université Grenoble Alpes, Gières, France. [36]Center for Open Science, VA, USA. [37]Rice University, TX, USA. [38]Saint Louis University, MO, USA. [39]University of Florida, FL, USA. [40]University of Groningen, Groningen, Netherlands. [41]University of California, Davis, CA, USA. [42]Leipzig University, Leipzig, Germany. [43]University of Winchester, Winchester, UK. [44]University of British Columbia, Vancouver, Canada. [45]University of Colorado, CO, USA. [46]New York University, NY, USA. [47]University of Pennsylvania, PA, USA. [48]University at Buffalo, NY, USA. [49]Carleton College, MN, USA. [50]University of Minnesota, MN, USA. [51]Purdue University, IN, USA. [52]Finnish Institute for Health and Welfare, Helsinki, Finland. [53]University of South Florida, FL, USA.

Chapter 3: Researchers' Perceptions of the State of Social and Personality Psychology

"There is a crisis in social psychology. It is in a state of profound intellectual disarray and there is little sense of progress."

This quote reads as something that could easily have been published in 2022, but it is in fact an item from a survey sent to social psychologists in 1979 (Nederhof & Zwier, 1983, p. 264; the item was adapted from an earlier essay by Ring, 1967). Nederhof and Zwier aimed to understand how social psychologists perceived their own field after a decade of debate about whether or not the discipline was in crisis (for a summary of the crisis of the 1970s, see Faye, 2012). Our aim here is similar: to understand how social and personality psychologists perceive the field now, and how they perceive it has changed in the last ten years.

Why does it matter what researchers think about their own discipline? Surely we would not simply take researchers' own word about how solid their literature is, the adequacy of their norms and practices, or their relevance or importance in society. These are matters that are best addressed by measuring these qualities more directly, as reflected in the published literature and other artifacts that can speak to the functioning of the field. Just as self-reports from research participants about their own traits on highly evaluative (desirable or undesirable) characteristics are likely to be biased (John & Robins, 1994), researchers' own perceptions of their discipline on dimensions core to the field 's values (e.g., validity, importance) should not be taken at face value.

Nevertheless, researchers' perceptions of their own field are interesting in their own right, even if they are not necessarily taken to be an accurate reflection of the state of the discipline. Just as self-reports are interesting to study to better understand how people make sense of themselves, understanding a field's self-assessment helps us understand the

discipline better, and puts other findings about the field in context. For example, as the metascience literature begins to shed light on how social and personality psychology has changed over time by examining its published literature (Schiavone & Vazire, 2022), these trends can be compared to how researchers within the discipline perceive the field to have changed. What researchers tell themselves about what is happening in their field is an important part of the field's norms and culture.

Nederhof and Zwier (1983) put a great deal of weight on their respondents' views of the discipline. In their view, "the opinion of the scientific community of social psychologists as a whole should be decisive" as to whether there 'actually' is a crisis in social psychology (p. 272). Our aim is not nearly as ambitious. First, we recognize the threat that self-selection can have on the representativeness of samples such as ours, and so caution generalizing our estimates to the entire discipline. Second, we are not primarily interested in researchers' bottom-line verdict as to whether or not there is a crisis (we suspect this rests in part on uninteresting semantic issues having to do with the specific word, 'crisis', as Nederhof and Zwier also point out, p. 272). Instead, we are more interested in understanding *what aspects* of the discipline (e.g., statistical methods, theory, causal inference, applied value, etc.) researchers perceive to be stronger vs. weaker, and how researchers perceive these strengths and weaknesses to have changed over the last ten years. In other words, we seek to understand where researchers think our discipline has improved in recent years, where we have been strong all along, and what weaknesses remain.

Despite a growing literature on public perceptions of, and trust in, science, little is known about how scientists themselves view their fields and trust the work being published in it. Similarly, although discussions exist as to how the replication crisis in psychology (and beyond) could impact how the general public views research in psychology (Ebersole et al., 2016; Fetterman & Sassenberg, 2015; Hendriks et al., 2020; Mede et al., 2020), relatively

little attention has been paid to how the crisis has shaped (or could shape) researchers' perceptions within the field.

The attitudes of those who make up a field are sure to impact how they engage and interact within it. Thus, understanding how researchers perceive their field may provide some valuable insight into and clues as to where a field is at, and where they may be headed. Such data could, in a sense, provide a brief (albeit limited) finger on the pulse of those producing the research that carries the field onwards.

In this study, we sought to capture researchers' perceptions of the state of the field of social and personality psychology. We separate these findings into three parts. In Part I, we explore researchers' overall perceptions of the field, of the published literature, and how their own work fits within it. In Part II, we examine what researchers perceive to be important in evaluating research quality, and how that compares to their perceptions of the published literature and their own work. In Part III, we seek to better understand the variance among researchers' perceptions by examining the individual differences that may correlate with researchers' perceptions including self-reported intellectual humility, views on open science, and career stage.

Understanding Perceptions of the Field Overall, the Published Literature, and Researchers' Own Work (Part I)

Perceptions of Field Overall

First, we sought to capture researchers' general perceptions of the state of the field. For example, do researchers think the field is experiencing a crisis? There are many similarities between the crisis of the 70s and the crises of the 2010s. As occurred in the years that followed the emergence of discussion of a crisis in the 70s, much has been written in the last decade (not only in the published literature, but on blogs, social media, and in the popular press) ranging from debate over the existence of a crisis to discussion of what it

means for the field, what may have caused it, potential implications for the field and its public reputations, how serious of a threat it presents, and what (if anything) researchers should be doing about it.

To compare how researchers working in 2021 view the field of social and personality psychology to researchers working in the 1970s, we adapted several items from the survey sent by Nederhof and Zwier (1983). Such data can provide insight into whether the field feels they are still grappling with similar issues as in the 1970s. Moreover, by comparing these snapshots of researchers' opinions about their field, we can consider whether perceptions appear to have shifted (and in what ways; e.g., become more positive), or if the general feelings about the field remain similar to how they were nearly 50 years ago.

Responses to Nederhof and Zwier's (1983) survey indicated that 44.6% agreed with the statement, "There is a crisis in social psychology. It is in a state of profound intellectual disarray and there is little sense of progress.", 44.0% disagreed, while the rest were undecided (p. 264). Their findings also indicated that few believed that much (1.6%), or moderate (17.7%) progress had been made within their own field of research in the last two years. Nederhof and Zwier were not the only researchers to investigate perceptions of the field following the crisis in social psychology that emerged in the 70s. In 1980, Lewicki (1982) surveyed members of the Society of Experimental Social Psychology (SESP) about their concerns about and perceptions of the state of social psychology (1982). Researchers were split when asked about the impact of discourse about the crisis in social psychology, but few were neutral (4%). Although many perceived the impact to be positive (29% "partially profitable", 11% "profitable"), 30% believed it "has not only been unprofitable but it also has had some costs, e.g., it weakened the confidence in social psychology.", and the rest considered it too soon to say (26%). Most researchers saw a need for stronger theory to be the most important for the field, but nevertheless were optimistic about the future of the

field (including improvements in predictions and the field's importance to daily life). These studies, among others (see Lipsey, 1974, for example), offer fascinating accounts that document important periods in the history of the field and the attitudes and perceptions of those within it.

Perceptions of Published Literature

Moving from researchers' broader perceptions of the field, we narrowed in on how researchers perceive the qualities of the published literature. To capture perceptions of recently published work, we collected researchers' ratings of the typical article published in 2020 (our survey was conducted in 2021). By doing so, we can assess what researchers see as some of the strengths and weaknesses of the research currently being published. For example, how accurate, transparent, or interesting do researchers think the typical article is?

We investigated perceived change over time by collecting researchers' retrospective perceptions of the typical article published in 2010. By comparing perceptions of research published in 2010 vs. 2020, we can identify in what areas researchers perceive the published literature to have changed. Understanding perceived change over time can shed light on whether researchers believe progress has been made. What areas do they think have improved the most? What has remained the same? And, has anything gotten worse?

These timepoints (2010 and 2020) span what has been a period of exceptional self-scrutiny in social and personality psychology, and of frequent debates about whether (and how) research practices and norms should change (perhaps similar to the 1970s). This period saw not only the emergence of the replication crisis, but also arguments about the existence of a generalizability crisis (Yarkoni, 2022), theory crisis (Oberauer & Lewandowsky, 2019), and an overall crisis of confidence (Earp & Trafimow, 2015; Pashler & Wagenmakers, 2012). These concerns arose and were fueled by a run of failed replications

(Doyen et al., 2012; Shanks et al., 2013; for a review, see Nosek et al., 2021), high-profile cases of fraud (Crocker & Cooper, 2011; Stroebe et al., 2012), a smattering of issues around the statistical validity of psychological research (e.g., lack of power, p-hacking; see Button et al., 2013; Masicampo & Lalande, 2012; Simmons et al., 2011; Vankov et al., 2014; Wagenmakers et al., 2011), concerns about WEIRD samples (Henrich et al., 2010), and a general lack of transparency (Alsheikh-Ali et al., 2011; Wicherts et al., 2006, 2011).

The replicability crisis has revealed that, as a field, we were in the habit of rarely taking stock of the state of our published literature. We argue that we similarly know little about what researchers who make up the field think of the published literature. We agree with the proposal put forth in what is believed to be the earliest published metascience study in social psychology that "From time to time it is well that we pause in our study of specific problems and attempt to achieve a synoptic view of things psychological" (Smoke, 1935, p. 537). Given the events of the past decade and what has been brought to light, we ask how researchers within the field perceive this period of upheaval. How dubious do researchers think the early 2010s were for research published in social and personality psychology? How do researchers perceive the published literature a decade later? If much is thought to have changed, what has changed the most? Has anything gotten worse? Do researchers think the field has self-corrected? Or that things were not (and still are not) that bad? In what areas do researchers think the field is doing well? What weaknesses are believed to remain?

We can also consider how perceived changes in the published literature map onto or reflect the discussions around problems in the field and the various reform efforts to improve psychological research (i.e., the Credibility Revolution; Vazire, 2018) over the last decade (for a review, see Nelson et al., 2018). Many of the calls for reform have focused on increasing transparency (e.g., preregistration, open science badges, 21 word solution,

selective reporting; Kidwell et al., 2016; Nosek et al., 2015; Simmons et al., 2012; van 't Veer & Giner-Sorolla, 2016), reducing questionable research practices (e.g., p-hacking, selective reporting, optional stopping, HARKing; see Munafò et al., 2017; Simmons et al., 2011), and encouraging practices that increase replicability and statistical validity (e.g., Asendorpf et al., 2013; Benjamin et al., 2018; Chambers & Tzavella, 2022; Cumming, 2014; Lakens, 2017; Lakens et al., 2018; Rohrer, 2018; Wagenmakers et al., 2012).

In the wake of such efforts, researchers have sought to estimate rates of the adoption of certain practices within the published literature. For example, Several surveys have also sought to capture researchers' self-reported attitudes towards such reforms (Fuchs et al., 2012), such as conducting replication studies (Agnoli et al., 2021; Buttliere & Wicherts, 2018), and researchers' willingness to engage in particular research practices, ranging from questionable research practices (Agnoli et al., 2017; Fiedler & Schwarz, 2016; John et al., 2012) to transparency related practices (Christensen et al., 2022; Toribio-Flórez et al., 2021; Washburn et al., 2018).

In response to the debate and discourse about the quality of research in psychology in the 2010s, Moytl et al., (2017) surveyed researchers in social and personality psychology to better understand how they viewed the state of the field. They sought to examine whether researchers considered the field "rotten to the core" and whether they expected it to improve in the future. Researchers were asked about their perceptions of the field and of questionable research practices (QRPs), including how acceptable they considered them, how often they used them, and whether the 'state of our science discussion' about the validity of research was likely to impact whether they engaged in QRPs in their future work.

Their findings provide some insight into how researchers working in 2015 thought of their field and how they saw it moving forward. Overall, researchers perceived that both research in the field and their own research had moderately improved as a result of the

recent discussions about the state of the field. However, they were largely uncertain whether the discussions had been more positive or negative for the field. Moytl et al. (2017) acknowledged that researchers expected that only ~50% of studies would replicate and that many reported having used QRPs. Nevertheless, they reach an optimistic conclusion that the researchers surveyed had "strong intentions to embrace higher standards of science going forward." (p. 10). This conclusion was based on researchers' responses as to whether the discussions in the field would change their engagement in various QRPs in the future (which they report in Figure 4). Most researchers, however, reported—across all the QRPs—that their likelihood of engaging in them had not changed. For eight of the ten QRPs included, less than 35% of researchers reported that their likelihood of engaging had decreased.

Whereas Motyl et al. (2017) argue their findings indicate that many researchers view the field to be getting better, the current study speaks instead to whether researchers think the field has gotten better. In other words, where do researchers in social and personality think the field was just before the beginning of the replication crisis and how does that compare to a decade later? How rotten do researchers think it was, if at all? And, importantly, how much progress do they think has been made?

Perceptions of Researchers' Own Work

In addition to collecting researchers' perceptions of the field and published literature, we examine their perceptions of their own work. Although it is commonplace for researchers' work to be evaluated by others (e.g., editors, reviewers, tenure committees, funders), little research has considered how researchers evaluate and view their own work. Thus, we investigated what researchers perceive to be the strengths and weaknesses of their own work, and how their perceptions of their own work compare to their perceptions of the articles published in their field.

Although surveys of researchers have asked them to self-report on particular behaviors or practices used in the research and publications, research on scientists' general perceptions of their own research is scarce. However, one example of such work includes a study that recruited authors of commentaries that had been highly cited in their field to ask them why they thought their most cited paper had been cited so much (Small, 2004). Content analysis of authors' responses suggested four main themes around the interest, novelty, utility, and significance of their work. In an extension of this research, Small et al. (2008) also categorized authors' open-ended responses explaining what they perceived to be the social and political implications of their work (78% of the sample reported their work had such implications). Among the most common included health implications, the advancement of science, policy implications, followed by economic, technological, and environmental.

How do researchers perceive their own work compared to the work of others in their field? Several studies have, however, explored researchers' beliefs and self-reported behaviors around scientific norms (e.g., the four Mertonian norms; Merton, 1942), and how they compare to their perceptions of other researchers. In a survey of 3,247 NIH funded scientists conducted in 2002 (Anderson et al., 2007), scientists self-reported their personal endorsement of scientific norms (e.g., organized skepticism, disinterestedness) and counter norms (e.g., self-interestedness, secrecy), and their perceptions of how much their own behavior and the behavior of other scientists reflects those norms. Scientists' reported beliefs were strongly aligned with the scientific norms. Their reported behavior also reflected these norms, but to a lesser degree suggesting that they do not always live up to the ideals they endorse. However, scientists' perceptions of the behavior of other scientists were much more negative. The behavior of other scientists was perceived to not only align less with the scientific norms than their own, but to be more aligned with the counter norms

than norms. Stronger endorsement of the scientific norms was associated with more negative perceptions of others' behavior. Anderson (2000) described such differences in scientists' own ideals and how they perceive other scientists to behave as the "disappointment gap." It is worth noting that these findings describe the results for the sample overall, which included a range of disciplines. It is unknown how consistent these patterns of results would be between disciplines, or more specifically in the field of social and personality psychology.

Understanding What Researchers Value When Evaluating Research (Part II)

We next examine which characteristics researchers consider most important when evaluating research quality. Understanding what characteristics (e.g., methodological rigor, theoretical significance) researchers consider most important for evaluating research quality can help us understand the ideals in the field. The characteristics that are considered most important arguably reflect researchers' priorities - where they believe we should focus our efforts. This can help guide decision making about possible reforms, about graduate training, and about standards for reviewing journal submissions, grant applications, job candidates, etc. Moreover, this snapshot of researchers' priorities in 2021 will provide a benchmark for future work examining whether priorities and norms are shifting in the field. For example, if we knew what researchers in the 70s and 80s believed was most important for evaluating the quality of social and personality psychology studies back then, we could examine whether those priorities have shifted.

How researchers in cell biology process and evaluate the credibility and impact of research was explored through qualitative data collected through semi-structured interviews (Harney et al., 2021). When discussing assessing credibility, researchers often expressed that they sometimes relied on information such as journal and author reputation and prestige, perceived quality of the peer review it underwent, perceived quality of figures and

data presented, and their assessment of the quality of the methods, analyses, and conclusions drawn. Evaluations of quality (beyond just credibility) included judgements of the research being well-written, compelling, and substantial, as well as information about the journal mentioned above. When assessing impact, researchers often looked to metrics such as journal impact factor, citations, perceived selectivity, as well as their perceptions of how the research impacted future work.

The use of proxies to help assess research quality—such as where something was published, journal impact factor, number of downloads—was also observed in researchers' responses to a survey about how they decide what to read and cite (Nicholas et al., 2015). However, researchers reported that reading the abstract, assessing the credibility of the data, and evaluating the soundness of the arguments presented were more important to them than this information. What information researchers rely upon in the absence of any cues related to where something was published was explored in a survey about how researchers evaluate credibility in preprints (Soderberg et al., 2020). Characteristics rated ranged from information about authors' affiliations, metrics such as downloads, open science practices, to citations. Open science practices (i.e., open materials, data, and analysis code) were rated the most important, followed by information about whether independent sources had been able to reproduce or assess the robustness of the results. Characteristics rated the least important included feedback such as user comments and endorsements, download metrics, and information about the authors.

Another approach of attempting to gain insight into what qualities researchers may value is by looking at the characteristics of the research that is getting published. For example, a text-analysis of abstracts indexed in the PubMed database published between 1974 and 2014 found a relative increase in the use of positive words by 880% (compared to negative words which increased by 257% (Vinkers et al., 2015). Terms including "robust",

"unprecedented", "novel", "innovative", and "groundbreaking" saw relative increases ranging from 2500 to 15000%. What contributes to shifts in language use like these are questions ripe for the field of metascience to explore. Do these changes reflect characteristics that the peer review process has selected for? Are researchers' perceptions of their work becoming more positive over time? Or, do researchers' descriptions reflect what they expect editors, reviewers, and journals to value?

One of the many benefits of open peer review is that it allows for metascientific research to be conducted to understand what concerns, criticism, and praises reviewers and editors commonly raise during the peer review process. Such work can serve as behavioral measures that complement self-report research and allow for comparisons between what researchers say that they value and what they demonstrate during peer review. An analysis of the content of peer reviews (N = 1716) of papers published in the *British Medical Journal* (BMJ) provides a glimpse into the evaluation that occurs during the peer review process (Falk Delgado et al., 2019). They found that positive words and phrases often used by reviewers included "strong", "clear", "well-written", "important issue", "important question", "relevance to a general readership". Negative words and phrases included "bias", "confounding", "risk", "not clear", "risk of bias". They observed reviews often focused on the quality of writing, the methodology used, and how well it appealed to a general audience. Comments about the results and specific findings were less common.

Measuring researchers' views about what characteristics are most important for evaluating quality provides additional context for interpreting researchers' perceptions of the current state of the field. Are the areas that are perceived as most important for evaluating research quality the areas that are perceived as strong, or those perceived as relatively weaker? For example, in Part I we may learn that researchers perceive the applied value of the typical article published in the field in 2020 to be quite low, and this may

indicate a perceived weakness. However, if the applied value of research is not considered especially important for evaluating research quality, this suggests that the published literature may be strong in the areas that researchers believe matter most.

Finally, we explore whether the characteristics researchers think are important for evaluating research quality are related to their perceptions of their own research. Are the qualities they rate themselves the most highly on considered the most valuable when evaluating research quality? Perhaps just as people tend to overvalue the products that they themselves own, perhaps the qualities that researchers perceive their work to exemplify are also thought to be more important to evaluating research quality. In a study exploring perceptions of authorship contributions, authors of manuscripts submitted to the *Croatian Medical Journal* were asked to complete a questionnaire where they rated how much they had contributed to the manuscript in various categories (e.g., conception and design, drafting the article; Ivaniš et al., 2011). Authors also rated how important they considered each category of contribution. Perhaps unsurprisingly, the greater the contribution researchers reported having to a certain contribution category, the more important they perceived that contribution category to be.

Understanding Individual Differences Among Researchers' Perceptions (Part III)

Researchers are not a homogenous group. We know from psychological research that individual differences matter, and we expect researchers know from working within the field that there are many disagreements and debates within the field. As such, we expect researchers' perceptions of the field to vary and that those variations may be associated with characteristics of the researchers themselves, such as their career stage. To explore the relationship between researchers' perceptions and researchers' characteristics, we examined a) intellectual humility, b) views on open science, and c) career stage.

Intellectual Humility

Researchers recognizing the limits of their knowledge and openness to being wrong, or being intellectually humble, is critical to a functioning science (Hoekstra & Vazire, 2021). Past research on intellectual humility (for a review, see Porter et al., 2022) suggests that intellectual humility is related to scrutinizing and identifying weakness in arguments and information (Koetke et al., 2022; Leary et al., 2017), being open to feedback and opposing positions (Porter & Schumann, 2018; Stanley et al., 2020), and flexibility in beliefs and positions (Leary et al., 2017; Zmigrod et al., 2019). It would follow then, that researchers who are more intellectually humble may be more open to criticism and willing to consider weaknesses in their own work and in the field at large. We investigate whether self-reported intellectual humility is associated with researchers' perceptions of the published literature and their own work.

Support for the Open Science Movement

Efforts to increase the transparency of research in psychological science grew substantially in the 2010s. While not limited to the field of psychology, these efforts and those advocating for them have often been labeled as the "open science movement." Despite this being a phrase commonly used in conversations to refer to efforts to improve psychological research over the last decade, there is no wide-spread consensus on what exactly the open science movement includes and represents—or even whether 'open science' should or should not be capitalized. Popular perceptions of the open science movement, depending on who you ask, could encompass a range of efforts to increase the credibility of research (e.g., registered reports, error detection, replication), a push for greater transparency (e.g., the adoption of preregistration, open data, and open materials), and the people engaged or interested in such efforts (e.g., the Center for Open Science, the Society for the Improvement of Psychological Science).

The open science movement in psychology is likely seen by many to have emerged in response to the issues uncovered during the replication crisis of the 2010s. Here, we explore the relationship between self-reported support for the open science movement and researchers' perceptions of the published literature, their own work, and how their own work compares to the published literature. For example, is support for the open science movement associated with more negative perceptions of published research and whether it has changed over time?

Career Stage

Do researchers at various stages in their careers perceive the field and published literature similarly? Do their perceptions of their own work differ? How much do early career researchers and senior researchers agree in their perceptions of the state of the field? Previous research has identified a number of differences in the attitudes and behaviors of researchers of various career stages, much of which has focused on perceptions and experiences related to research integrity and transparency related practices. This includes research examining how scientists of different career stages perceive the climate around research integrity within their departments. A survey of researchers in Amsterdam (Haven et al., 2019), for example, found that compared to full and associate professors, early career researchers had more negative perceptions of the fairness of departmental expectations (e.g., publishing, obtaining grants), how much the department socialized around topics related to research integrity, the relationships between supervisors and supervisees, and the resources available related to the responsible conduct of research. PhD students perceived there to be more integrity inhibitors (e.g., pressure to publish, competition, suspicion) that negatively affected the research climate in their departments than senior researchers. While previous findings too have found that researchers' perceptions of research integrity climate

vary by career stage, the exact patterns of results sometimes differ (see Martinson et al., 2006; Wells et al., 2014).

Differences in the views of early and later career researchers were also observed in a qualitative study aimed to understand how biomedical scientists perceive the publication process and the culture around it (Tijdink et al., 2016). Compared to post-docs and professors, PhD candidates focused more on research quality and expressed more idealistic views about science as a means of seeking out truths. More senior researchers, on the other hand, were said to be more cynical and "more sympathetic to the somewhat dubious elements in the scientific process." (p. 6).

Beyond perceptions of departmental climate and the larger culture of publishing, early career researchers may face unique challenges as they attempt to establish themselves in a field that many consider to be in the midst of a crisis. A survey of 517 early-career researchers (i.e., researchers post-PhD, but less than 10 years so) in STEMM working at Australian universities or research institutes revealed that between 30.7 to 41.4% reported that their job satisfaction and/or career progression had been impacted-or-strongly impacted by questionable research practices in their institution, and between 28.9 - 33.6% by questionable research practices outside their institution (Christian et al., 2021). Open-ended responses highlighted experiences such as being pressured to engage in what they considered questionable or unethical research practices by more senior researchers.

Several studies have also explored early career researchers' attitudes towards open science (Toribio-Flórez et al., 2021) and perceptions of questionable research practices (Stürmer et al., 2017). Given the scope and focus of these studies on early career researchers, they often do not include samples of more senior researchers allowing for direct comparisons. However, a study on data sharing practices in the field of animal biotelemetry compared how authors at differing career stages responded to requests for

data. They found that corresponding authors who were early-career researchers were far more likely to share data upon request (72%) than senior researchers (11%; Campbell et al., 2019).

Perhaps the largest attempt to recruit and compare researchers' attitudes and behaviors relating to open science was reported in Wave I of the State of Social Science (3S) Survey conducted in 2018 (Christensen et al., 2022). This study included four fields in the social sciences (psychology, sociology, economics, and political science) and compared a sample of PhD students at top-20 PhD programs in North America and a sample of authors published in the 10 most cited journals within their discipline. Results found that self-reported awareness of and attitudes towards open science practices to be similarly positive in the sample of PhD students and the sample of authors. Curiously, however, fewer PhD students reported engaging in open science practices than published authors. These results are consistent in the overall sample (N = 2801) and in the subsample from the field of psychology alone (n = 598).

The public project page (osf.io/zn8u2/) for the State of Social Science Survey describes a follow-up wave that collected data in 2020 that included the same items assessing support for open science. Data from Wave 2 are publicly available and include a sample of N = 2068 researchers, of which 1609 had participated in Wave 1 and the remaining were newly recruited. By accessing their open data, we tested whether similar results would be found in Wave 2. Given the scope of our research, we only analyzed data from participants in the field of psychology (n = 461). The results of Bayesian t-tests indicate that while published authors reported engaging in more open science practices than PhD students in the 2018 sample (BF_{10} = 19.51), the data from 2020 suggest that PhD students and published authors reported similar levels of awareness, attitudes, and behaviors related to open science practices. Compared to Wave 1, researchers in psychology who participated

in Wave 2 overall reported greater awareness of open science practices (increasing from M = 0.90 in Wave 1 to M = 0.97 in Wave 2; BF_{10} = 5.28 x 10^8) and more engagement in open science practices (increasing from M = 0.39 in Wave 1 to M = 0.49 in Wave 2; BF_{10} = 4929.28). Attitudes towards open science did not appear to meaningfully differ between waves.

While not discussed in the main text of Christensen et al. (2022), the State of Social Science Survey also included four items asking participants about their confidence in the replicability of research in their field (see Christensen et al., 2022, Appendix Table 8, p. 40). Specifically, researchers reported how confident they were in the replicability of influential research findings, canonical research findings, recent research findings, and studies reported in the latest issue of their field's top journal, rating them from 1 (*Very low confidence*) to 5 (*High confidence*). Analyzing data from both waves, we compared PhD students' and published authors' perceptions of replicability in the field of psychology. In Wave I, published authors were more optimistic than PhD students about the replicability of canonical research findings. However, no other evidence of differences in perceptions of the replicability of research between the samples was found. Overall, perceptions of the replicability of the various categories of research findings were similar between Waves, with one exception being that participants in Wave 2 were less confident that canonical research findings would replicate (M = 3.28) than in Wave 1 (M = 3.53; BF_{10} = 7175.39).

In the current study, we compare researchers' perceptions in social and personality psychology across career stages. Do early and later career researchers share similar perceptions of the state of the published literature and how it has changed over the last decade? Is career stage related to how positively researchers perceive their own work?

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the studies (Simmons et al., 2012). Our data, analysis code, materials, and preregistration for this project are available at osf.io/pfu5y/. The data reported are part of a larger project exploring researchers' views of research in psychology. The complete survey materials (including measures not included in these analyses) and the general preregistration of the larger project is available at osf.io/wbhpq/. All measures used in the current analyses are available in the Supplemental Materials. This research was approved by the Institutional Review Board at the University of California, Davis.

Participants and Recruitment Strategy

Participants included 724 researchers who completed an online survey. We began recruitment by inviting (via email) first and last authors of empirical articles published in social and personality psychology between 2017 and 2021 in seven peer-reviewed journals: *Collabra Psychology*, *Journal of Experimental Social Psychology* (JESP), *Journal of Personality and Social Psychology* (JPSP), *Personality and Social Psychology Bulletin* (PSPB), *PLOS One*, *Psychological Science*, and *Social Psychological and Personality Science* (SPPS). Social and personality articles published in *Collabra* and *Psychological Science* were identified by the first author by reading the title and abstract of each article. *PLOS One* articles included articles listing social or personality psychology in the subject area. We also sampled ~300 articles publicly shared as preprints on PsyArXiv that were tagged with the subject social and personality psychology.

Our preregistered target sample size was 1,500 researchers, which was determined based on the number of unique authors of articles published between 2017 and 2021 and the funds we had available. The publication period (2017 - 2021) was selected to target

researchers that are most likely to be active in the field and the articles most likely to contain valid and up-to-date email addresses for contacting authors, which tend to rapidly decay over time (Wren et al., 2006). We attempted to collect contact information for authors whose email addresses were not listed in their published articles by searching the internet for publicly available contact information (e.g., personal websites, university profiles, other publications). To ensure data quality, all email addresses were double coded by trained undergraduate research assistants. The first author resolved disagreements and cases for which research assistants had difficulty confirming authors' identities or contact information. Authors whose email addresses we were unable to find were excluded, along with authors who were discovered to be deceased during the search.

After sending the original invitations, we decided to expand our sample in an effort to reach our target sample size. We did so, deviating from our preregistration, in three ways. First, we invited first and last authors who had published articles in the aforementioned journals between 2010 and 2016 who a) were not already included in the original sample and b) whose email addresses were provided in the contact information reported in the article. We did not attempt to locate authors whose email addresses were not reported. Second, we expanded our list of target journals, and included first and last authors—not already included in our sample—whose email addresses were reported in articles published between 2017 and 2021 in *Frontiers in Psychology* (in the Personality and Social Psychology section) and *Perspectives on Psychological Science* (coded as social and/or personality based on title), as well as articles published between 2011 and 2021 in *Personality and Social Psychology Review* and *Social Psychology Quarterly*. The journals and years sampled were determined based on relevance and accessibility. Third, we posted an invitation to members of the *Society of Social and Personality Psychology* (SPSP) on the SPSP listserv.

There was, of course, overlap in the researchers included in these groups. We identified authors with multiple publications based on name (and email addresses) in effort to include only unique authors. For the sub-sample recruited through the SPSP listserv, we included a question in the survey asking if they had already participated to flag data from people who had already participated using a link in an email invitation. Overall, we sent over 17,000 email invitations to authors. Of those, we received around 3,000 undeliverable notices, and we suspect many other email addresses were no longer current as researchers shift positions and universities. We received a number of responses indicating the authors had left academia, retired, passed away, were on sabbatical, or were otherwise not active in the field.

A total of 724 researchers completed the survey (639 authors recruited via email and 85 SPSP members recruited via the SPSP listserv). As preregistered, we excluded participants who did not consider social or personality psychology to be one of their areas of research (n = 53) or who did not report their area of research (n = 3). We also excluded participants recruited through the SPSP listserv who reported they had taken the survey previously (n = 8). This resulted in a final sample of 660 researchers. Researchers' reported subfields are reported in Table 2 and career stages in Table 3.

Of the 72 participants who reported personality as their primary area, 41 also identified social psychology as one of their areas of research (see Table 2). Of the 361 who reported social psychology as their primary area of research, 132 also considered personality psychology one of their areas of research. The remaining 227 participants in the final sample reported another area as their primary area of research, but reported either social (n = 114), personality (n = 34), or both (n = 79) as one of their areas of research.

Table 2

Researchers' Reported Subfields

Secondary Area of Research	Primary Area of Research (N = 660)		
	Social Psychology n = 361	Personality Psychology n = 72	Something Else n = 227
Social Psychology	—	41	114
Personality Psychology	132	—	34
Both	—	—	79

Table 3

Researchers' Reported Career Stage

		Totals (N = 660)	% of Sample
Pre-PhD	Not students	8	1.21%
	Graduate students	72	10.91%
Post-PhD	< 8 years	242	36.67%
	8 - 20 years	227	34.39%
	> 20 years	111	16.82%

Procedure and Measures

Researchers participated by completing an online survey. In exchange for their participation, we donated US$40 to a scientific organization of the researchers' choice from a list of eight societies.[2] Upon opening the survey, researchers were presented with a

[2] Researchers selected one society to direct a donation from the following: the Association for Research in Personality (ARP), the Center for Open Science (COS), PLOS, the Society of Experimental Social Psychology (SESP), the Society for the Improvement of Psychological Science (SIPS), and the Society for Personality and Social Psychology (SPSP). As noted in our preregistration, the amount donated per participant was originally set as US$20. However, due to response rates, we opted to double this amount (with the approval from the funder). All surveys completed prior to this decision were retroactively increased to US$40.

consent form describing the purpose of the study as "to better understand researchers' views of the field of social and personality psychology, and how the field has changed over the last ten years." Authors who had not completed the survey were sent two reminder emails. Below, we describe the variables from a large project (see osf.io/pfu5y/) that we preregistered for these analyses.

Perceptions of The Field. Four items were adapted from a survey of psychologists reported by Nederhof and Zwier (1983; see Appendix I, p. 275 - 276). We included what they referred to as the 'crisis item,' which stated "There is a crisis in social and personality psychology. It is in a state of profound intellectual disarray and there is little sense of progress." We also included three additional items: "During the last decade, social and personality psychologists have shown much concern over the state of their discipline."; "Social/personality psychology can rightfully claim to be a science rather than an art."; and "It may be a good idea to halt all data collection until some of the fundamental difficulties which social and personality psychology face today have been overcome." The only changes we made to these items was replacing "social psychology" with "social and personality psychology." These items were rated using a Likert-type response scale from 1 (*Very strongly disagree*) to 7 (*Very strongly agree*). Although the original authors used a reversed scale, from 1 (*Very strongly agree*) to 7 (*Very strongly disagree*), we opted to keep our scale anchors consistent throughout the survey. Thus, we reverse-scored the results reported by Nederhof and Zwier (1983) when comparing our samples.

Researchers were asked to indicate how confident they are in published research in social and personality psychology, in their own published work, and to rate how self-correcting is research in psychology, from 1 (*Not at all*) to 5 (*Extremely*). After doing so, researchers rated how much the public should trust research in psychology, from 1 (*Not at all)* to 5 (*A great deal*).

Perceptions of Published Articles. Researchers were asked to imagine the typical social or personality psychology article published in 2010, and the typical social or personality article published in 2020 and rate each on 15 characteristics (see Figure 6) using a Likert-type response scale from 1 (*Not at all*) to 5 (*Extremely*). We included various characteristics relating to the quality (e.g., methodologically rigorous, trustworthy, accurate), significance (e.g., important, groundbreaking), and engagingness (e.g., interesting, boring, creative) of research. For each item, researchers rated how they imagined the typical article published for each year side-by-side. To examine researchers' perceptions of the validity of published research, researchers were similarly asked to imagine the typical social or personality psychology article published in 2010 and 2020 and rate each on the Four Validities (i.e., construct, statistical conclusion, internal, and external validity; Shadish et al., 2002) from 1 (*Very low*) to 5 (*Very high*). Next, researchers estimated roughly what percent of empirical studies reported in published research articles in social and personality in 2010 and 2020 that they think are replicable (if we collect new data, we would get similar results), reproducible (if we re-analyze original data, we would get same results), report p-hacked results, and report fraudulent results/data, from 0% (*none of them*) to 100% (*all of them*).

Perceptions of Their Own Work. After rating how they imagined the typical social or personality psychology articles for each year, researchers were asked to consider their own work, published or in progress, over the last 10 years and to rate it on the same 15 characteristics, from 1 (*Not at all*) to 5 (*Extremely*).

Evaluating Research Quality. Researchers were then presented with a subset of seven out of the 15 characteristics previously rated along with the four validities items and asked, "when evaluating the quality of research in social and personality psychology, how important do you think each of these qualities is?", from 1 (*Not at all*) to 5 (*Extremely*).

Intellectual Humility. Intellectual humility was measured using the self-report Intellectual Humility Scale (Leary et al., 2017). Researchers rated themselves on six items (e.g., "I accept that my beliefs and attitudes may be wrong."), from 1 (*Not at all like me*) to 5 (*Very much like me*). Researchers' scores were calculated as their mean rating of the six items. The internal consistency of scores was acceptable $\alpha = 0.78$, 95% CI [0.75, 0.81].

Support for the Open Science Movement. Researchers' views on the open science movement were measured using two developed items asking how much they agree or disagree with the principles behind the open science movement in psychology, and with the practices and policies proposed by advocates of the open science movement in psychology, from 1 (*Disagree strongly*) to 5 (*Agree strongly*). The correlation between items was $r = .56$, (95% CI [0.50. 0.61], $BF_{10} = > 1{,}000$). As preregistered, researchers' responses to these items were averaged.

Career Stage. To assess career stage (see Table 3), researchers were first asked if they had a PhD. If not, they were then asked if they were a student, selecting graduate student, undergraduate student, or no they are not a student. If researchers had a PhD, they were asked to describe their current career stage by selecting either < 8 years post-PhD, 8 to 20 years post-PhD, or > 20 years post-PhD. As preregistered for the analyses of career stage, we included the three post-PhD groups and graduate students. Researchers who indicated they did not have a PhD and were not a student were excluded for these analyses ($n = 8$). To help protect the anonymity of participants, we opted to use this categorical approach to measure career stage (rather than a more precise measure).

General Analytic Approach

Data were analyzed and visualized using R (Version 4.2.0; R Core Team, 2022)—with the help of the following packages: tidyverse (Wickham et al., 2019), psych (Revelle, 2022), waffle (Rudis & Gandy, 2017), bayestestR (Makowski et al., 2019), correlation

(Makowski et al., 2019), BayesFactor (Morey et al., 2021), ggridges (Wilke, 2021), corrplot (Wei & Simko, 2021), and hrbrthemes (Rudis, 2020)—and JASP (Version 0.16.2; JASP Team, 2022).

Our analyses relied on Bayesian hypothesis tests. Bayesian statistics combine information from the observed data and a prior (probability distribution) selected by the researcher to produce a posterior distribution, which represents the probability of a hypothesis given the data observed (Dienes, 2011; Wagenmakers et al., 2018). Inferences from Bayesian hypothesis tests rely on Bayes factors, which represent the strength of evidence for one hypothesis relative to the other given the data collected (Rouder et al., 2009). Bayes factors compare two hypotheses: H_0 (the null hypothesis) and H_1 (the alternative hypothesis). The larger the Bayes factor, the stronger the evidence in favor of that hypothesis. Unlike frequentist statistics which test only against the null hypothesis, Bayes factors allow evidence for the null hypothesis to be quantified as they calculate relative predictive evidence. Bayesian statistics also offer the advantage of being conditional only on what is observed in the current data rather than dependent upon how a model performs over an infinite set of hypothetical samples (Wagenmakers et al., 2018).

In the results below, we conduct Bayesian paired-samples t-tests, correlations, and ANOVAs. We report BF_{10}, which represents the Bayes factor favoring the alternative hypothesis and BF_{01}, favoring the null hypothesis. As preregistered, we interpret all Bayes factors using Lee and Wagenmakers's (2013) proposed classifications. As preregistered, priors used in all Bayesian t-tests were defined as Cauchy distributions centered around zero and width parameters of 0.707. As analyses were largely exploratory, this value was selected as it is often relied upon as a default (JASP Team, 2022; Morey et al., 2022). All Bayesian correlation analyses were conducted, as preregistered, using stretched beta priors with widths of 1 were used to assign all values between -1 and 1 equal prior probabilities. For

all Bayesian ANOVAs, models were assigned equal prior probabilities (i.e., model odds of .5). When alternative hypotheses overperformed the null models, pairwise comparisons were conducted and considered the posterior odds corrected for multiple testing and uncorrected Bayes factors.

As can be seen from our preregistration, this paper includes an ambitious number of variables and analyses. To provide some structure, we have separated these results into three main parts below. For the sake of readability, we do not describe the results of every statistical test in-text, although these details are available in tables and figures (with additional details in the Supplemental Materials). Instead, we aimed to provide a general summary of the findings and to highlight specific results that we consider to be the most important and to have the strongest evidence. There is, of course, some subjectivity that came with doing so, and we encourage interested readers to explore these results in greater detail.

Part I. Overall Perceptions of the Field of Social and Personality Psychology, the Published Literature, and Researchers' Own Work

In Part I, we first describe researchers' general perceptions of the state of the field of social and personality psychology. In doing so, we compare researchers' responses to the items adapted by Nederhof and Zwier (1983) to the samples they originally recruited. Second, we analyze researchers' more specific perceptions of the typical article published in social and personality psychology in 2010 and in 2020 and investigate perceived change over time. Third, we explore researchers' perceptions of their own research and how they compare to their perceptions of the published literature.

General Perceptions of the State of the Field and Comparisons Between the Current Sample and Nederhof and Zwier (1983)

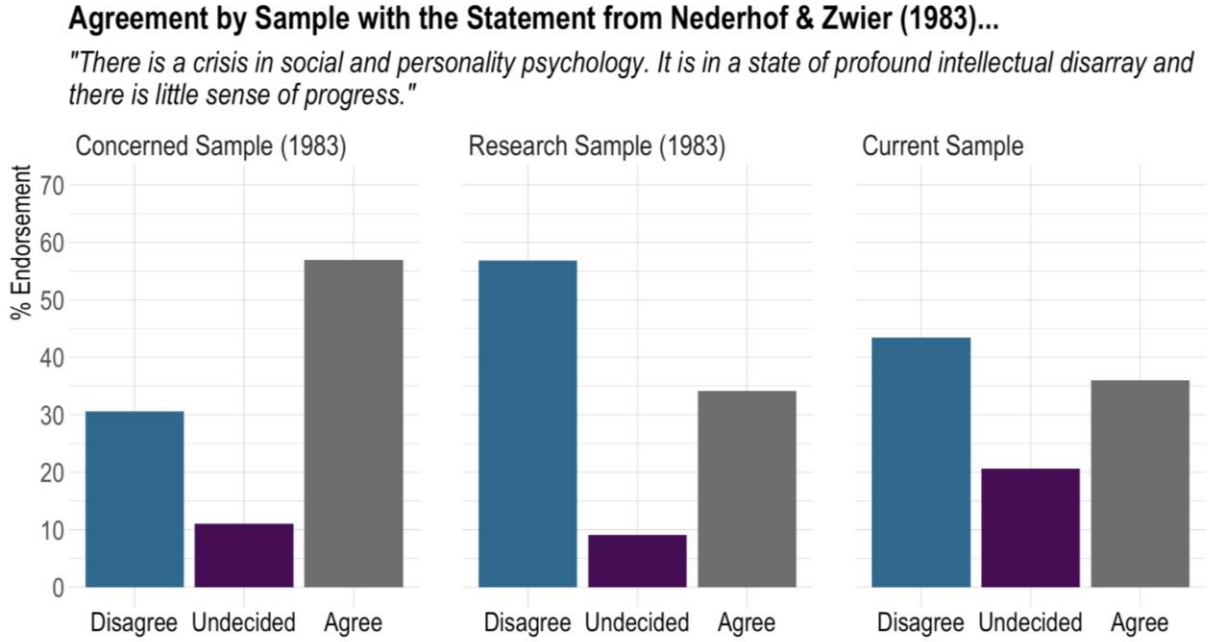
Researchers' responses to the items adapted from Nederhof and Zwier's (1983) survey of psychologists are presented in Figures 3 and 4 and reported in Table 4, where we compare the current sample to both the "Research" Sample and "Concerned" Sample recruited by Nederhof and Zwier (1983). The "Research" Sample included first authors of articles that had been recently published in two of the journals that are also included in our own sample, the *Journal of Experimental Social Psychology* and the *Journal of Personality and Social Psychology*, as well as the *European Journal of Social Psychology*. The "Concerned" Sample included first authors of work published (not limited to the journals above) on topics dealing "with (some aspect of) the 'crisis' in social psychology" (p. 262). Given the similar recruitment strategies used and overlap in journals sampled, we consider the "Research" Sample to be the best comparison with our own sample.

In our sample, the statement "There is a crisis in social and personality psychology. It is in a state of profound intellectual disarray and there is little sense of progress" elicited a wide range of responses (Figures 3 and 4), suggesting little agreement among researchers on this topic. More participants in our sample disagreed (43.36%; selected 1 – 3) than agreed (36.03%; selected 5 – 7). Compared to the researchers recruited by Nederhof and Zwier (1983), the current sample overall agreed slightly more with the statement than those in the "Research" Sample and less than those in the "Concerned" Sample, though we did not conduct inferential tests of these differences. As shown in Figure 3, the percentage of researchers in our sample who agreed (36.03%) was quite similar to that of the "Research" Sample (34.1%). However, roughly twice as many researchers were undecided in the current sample (20.61%; selected 4) than researchers in both samples reported in Nederhof and Zwier (10.7% across samples). The double-barreled nature of this item introduces some ambiguity as to which of the statements within the item researchers were agreeing or disagreeing with (e.g., the existence of a crisis and/or the state of the field) that should be

considered. Nevertheless, the inclusion of this item offers a historical comparison which provides some insight into how the perceptions of researchers working in the field in 2021 compare to researchers working ~40 years ago.

Figure 3

Comparison of Researchers' Perceptions of the Existence of a Crisis in the Current Sample and the Samples Reported by Nederhof and Zwier (1983)



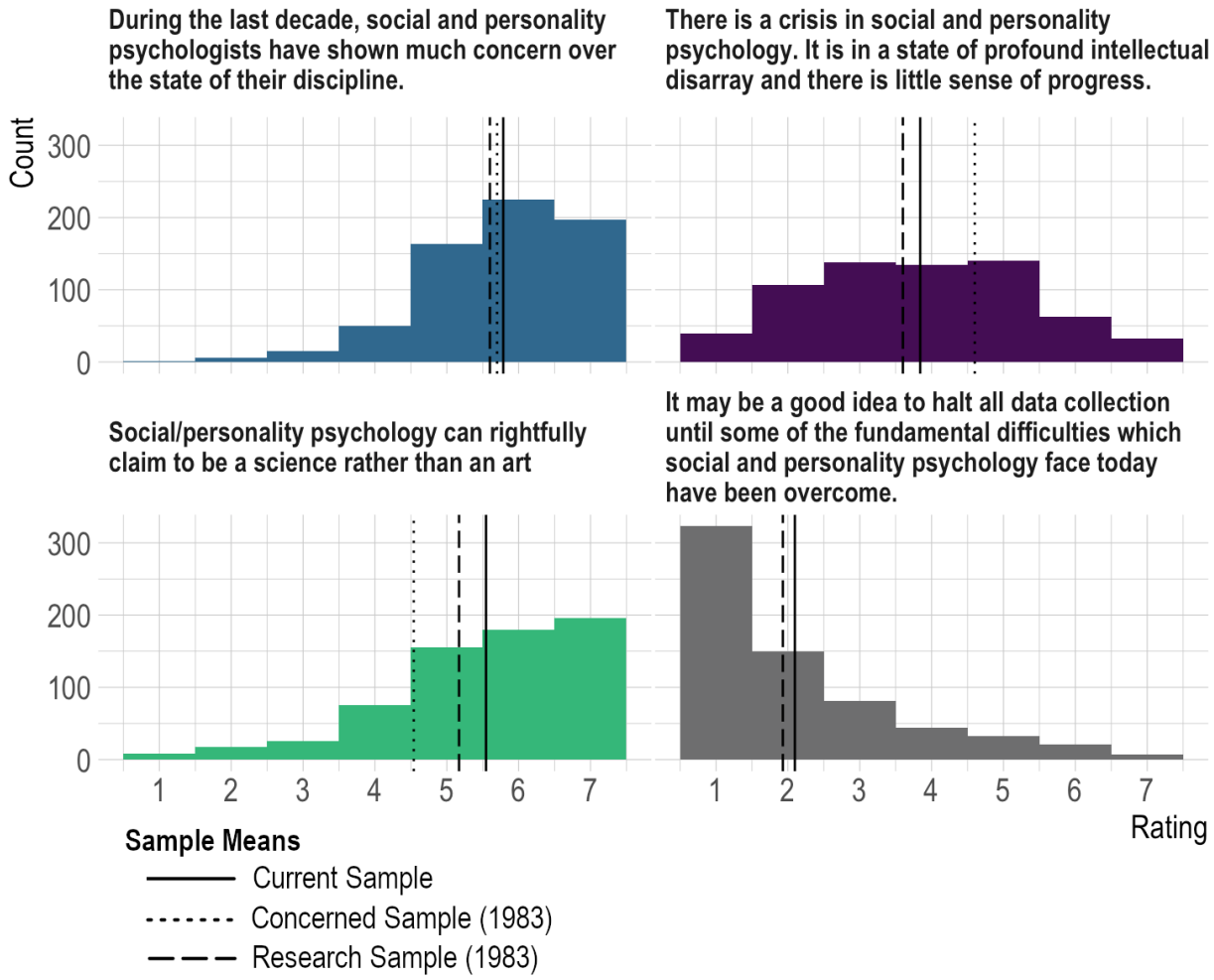
Note. Values for the Concerned and Research Samples were reported in Figure 1 and in Table 1 of Nederhof and Zwier (1983). Groupings for the current sample were created based on researcher responses: Disagree group = ratings from 1 - 3; Undecided = 4, Agree = 5 - 7.

Agreement that "During the last decade, social and personality psychologists have shown much concern over the state of their discipline" was high across all three samples (ranging from 87.6 – 89%; see Table 4). As seen in Figure 4, researchers in the current sample agreed more with the statement that "Social/personality psychology can rightfully claim to be a science rather than an art." than the "Concerned" Sample and to a similar extent as the "Research" Sample. There was widespread disagreement that "It may be a

good idea to halt all data collection until some of the fundamental difficulties which social and personality psychology face today have been overcome." among researchers in both the current sample and the “Research” Sample.

Figure 4

Researchers’ Ratings of Items Adapted from Nederhof and Zwier (1983)



Note. Bars represent responses from the current sample. Means for the “Concerned” and “Research” samples reported by Nederhof and Zwier (1983) in Tables 1, 2, and 3 are reverse scored in these figures, as the response scale used was opposite from that used in the current sample (e.g., 1 (*Strongly agree*) to 7 (*Strongly disagree*)). The “Concerned” sample’s mean was not reported for the item in the bottom right panel.

Table 4

Researchers' Perceptions of Their Field: Comparisons of the Current Sample and Samples Reported in Nederhof an Zwier (1983)

Items Rated from 1 (*Strongly disagree*) to 7 (*Strongly agree*)	Nederhof & Zwier (1983)		Current Sample $N = 660$
	"Concerned" Sample $N = 73$	"Research" Sample $N = 89$	
	M (*SD*)	*M* (*SD*)	*M* (*SD*)
During the last decade, social and personality psychologists have shown much concern over the state of their discipline.	5.7 (1.2)*	5.6 (1.1)*	5.79 (1.09)
There is a crisis in social and personality psychology. It is in a state of profound intellectual disarray and there is little sense of progress.	4.6 (1.6)*	3.6 (1.4)*	3.84 (1.56)
Social/personality psychology can rightfully claim to be a science rather than an art.	4.54 (1.3)*	5.17 (1.1)*	5.55 (1.37)
It may be a good idea to halt all data collection until some of the fundamental difficulties which social and personality psychology face today have been overcome.	—	1.93 (1.1)*	2.10 (1.45)

Note. *Means for the "Concerned" and "Research" samples reported by Nederhof and Zwier (1983) in Tables 1, 2, and 3 are reverse scored in this table, as the response scale used was opposite from that used in the current sample (e.g., 1 (*Strongly agree*) to 7 (*Strongly disagree*)).

Overall, researchers' perceptions of the state of the field in 2021 were not so dissimilar to those of researchers in 1979. Neither sample demonstrated widespread consensus among researchers on whether the field was experiencing a crisis. Nevertheless, a significant minority of researchers across samples agreed that there was a crisis in the field. Moreover, they not only agreed that there was a crisis, but agreed that the field was in a

state of profound intellectual disarray with little sense of progress. The severity of this statement should be considered when interpreting these results, as it casts the field as being in a rather disturbing state. What does it mean for the field that roughly one third of researchers agreed with such a worrisome statement? What should we expect from a healthy science? There is no clear or simple answer. However, we expect many—including the general public—would find one third to be an unsettling and uninspiring percentage.

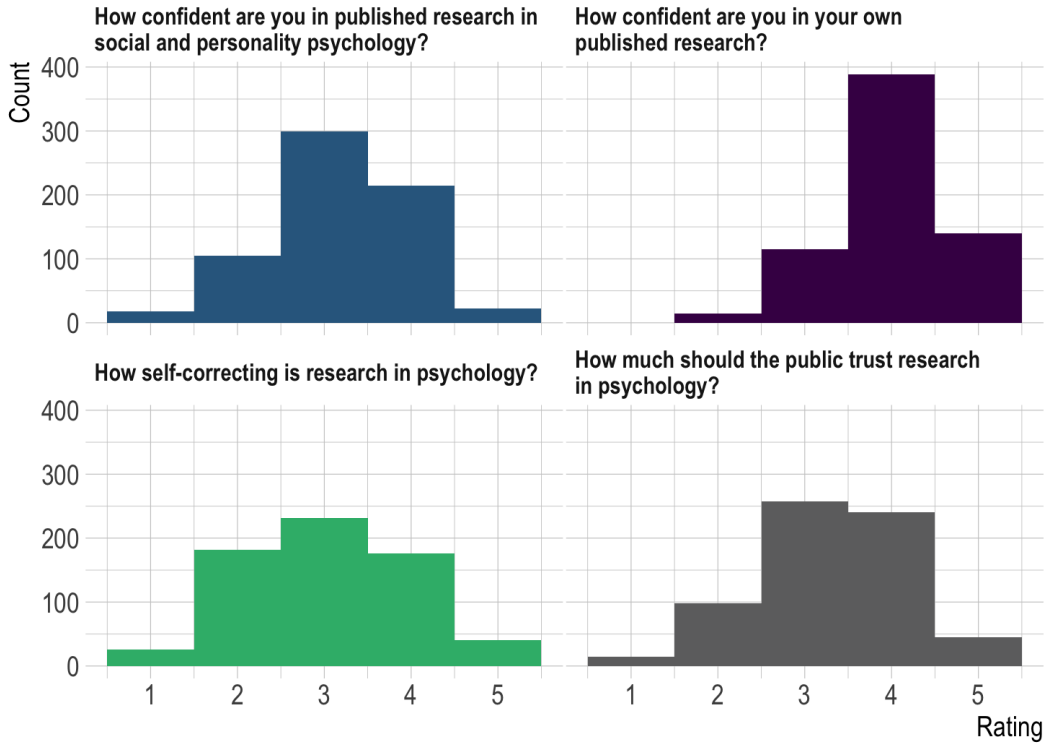
Nederhof and Zwier (1983) suggested their results may indicate that the researchers who were arguing that there was a crisis had been unsuccessful in convincing the field at large. They also pointed out that some of the researchers who disagreed "may merely have loathed the use of the word 'crisis'" (1983; p. 272) but may nevertheless have recognized and shared similar concerns about the state of the field. This could also be true of the current sample. It should also be noted that these data cannot speak to whether researchers believed that there was *ever* a crisis in their field. It is possible researchers in our sample may indeed agree that there was a crisis in psychology (e.g., the replication crisis), but believe that improvements have been made and that the field is no longer in crisis.

Additional measures of researchers' general perceptions of the field can be seen in Figure 5. Researchers' overall confidence in published research in social and personality psychology had a mean rating only slightly above the midpoint (M = 3.18, SD = 0.84) of the scale (ranging from 1 (*Not at all)* to 5 (*Extremely*)). Nearly half the sample (45.4%) selected the midpoint. While 35.9% of researchers had confidence in published research in the field, 18.7% indicated a lack of confidence. Unlike published research in the field more generally, researchers were, unsurprisingly, more confident in their own research (M = 4.00, SD = 0.69 on the same 5-point scale). The majority indicated confidence in their work (80.4% selecting 4 or 5). However, 17.5% of researchers selected the midpoint. When asked how much the public should trust research in psychology, most researchers selected either 3

(39.2%) or 4 (36.7%; M = 3.31, SD = 0.89). Finally, there was not much of a consensus as to how self-correcting research in psychology is (M = 3.04, SD = 0.98), as responses produced a fairly wide and symmetric distribution around the midpoint.

Figure 5

Researchers' General Perceptions of the Field



These results offer some insight into how researchers think about the general state of the field and how their perceptions compare to those of researchers who came before them. Next, we examine in greater depth how researchers perceive the state of published research in social and personality psychology, and what they think has—and has not—changed between 2010 and 2020.

Perceptions of Published Research

Perceptions of the Typical Article and Perceived Change Over Time

Researchers' perceptions of the typical article published in social and personality psychology in 2010 and 2020 are shown in Figures 6 and 7. In both figures, characteristics are presented in order from those with the highest mean rating to the lowest. As seen in Figure 6, the typical 2010 article was rated the highest on having exaggerated findings, being creative, and interesting and the lowest on being boring, statistically rigorous, and transparently reported. Unlike the typical article published in 2010, researchers rated the typical 2020 article the highest on being transparently reported, statistically and methodologically rigorous, scientific, and trustworthy and the lowest (but all with a median rating of 3) on how groundbreaking, boring, applied they were.

Figure 6

Perceptions of the Typical Article Published in 2010

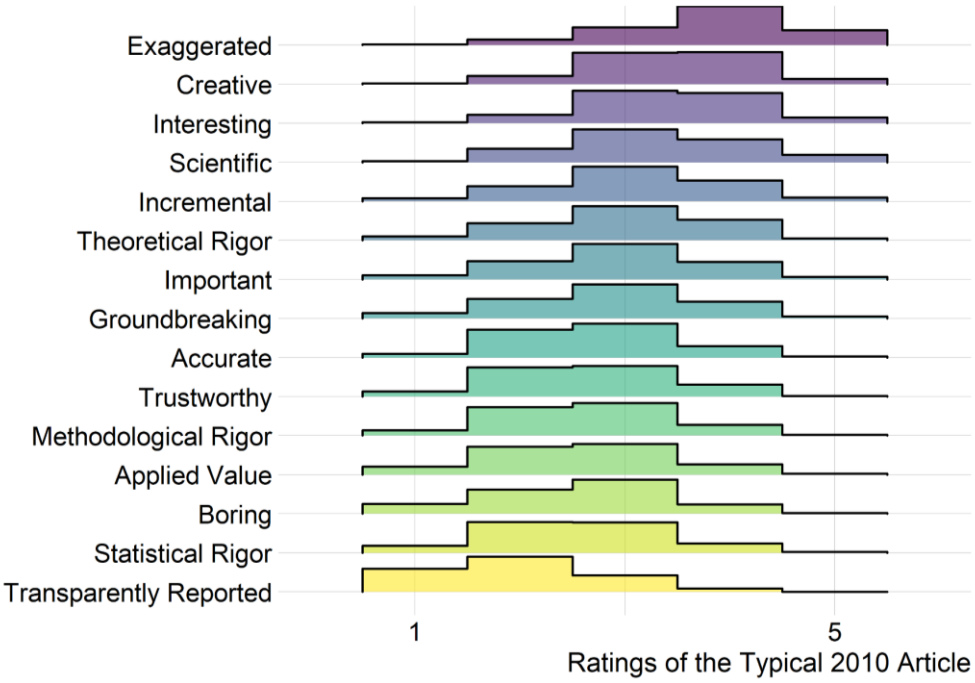
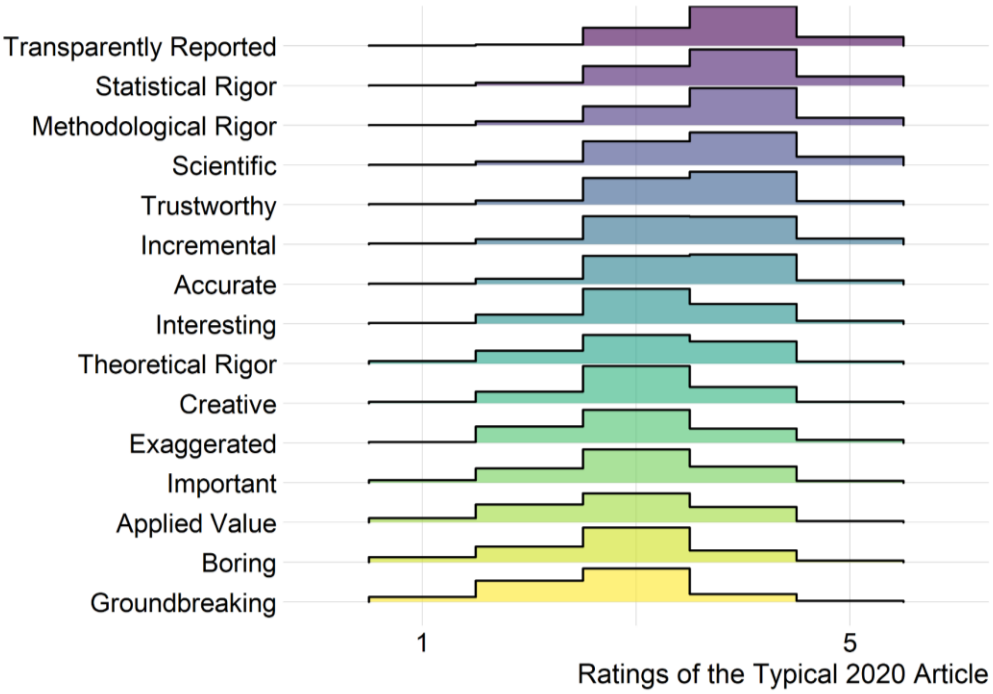


Figure 7

Perceptions of the Typical Article Published in 2020



To explore perceived change over time, we conducted Bayesian paired-samples t-tests comparing researchers' ratings of the typical article published in 2010 vs. 2020. As seen in Table 5, we found very strong-to-extreme evidence favoring the alternative hypotheses (i.e., that researchers' perceptions of the typical 2010 and 2020 article differed) for 14 out of the 15 characteristics rated. Overall, researchers viewed the typical 2020 article more positively than the typical 2010 article on nearly every characteristic (e.g., more accurate, methodologically rigorous, scientific, less boring; see Figure 8). However, researchers also saw the typical 2020 article as less creative, groundbreaking, and interesting compared to the typical 2010 article.

The largest perceived change between 2010 and 2020 articles was seen in how transparently reported the typical article was, which increased by nearly two points on the five-point scale. After transparency, the largest shifts in perceptions were of how statistically rigorous, methodologically rigorous, and trustworthy the typical articles were. Although evidence that perceptions differed between typical 2010 and 2020 articles was found for almost all characteristics, many of the accompanying estimated effect sizes (i.e., posterior medians) were quite small. Researchers perceived the least change in how important, theoretically rigorous, groundbreaking, and interesting articles were. Shifts in how individual researchers rated each characteristic in 2010 vs. 2020 can be seen in Figure 9.

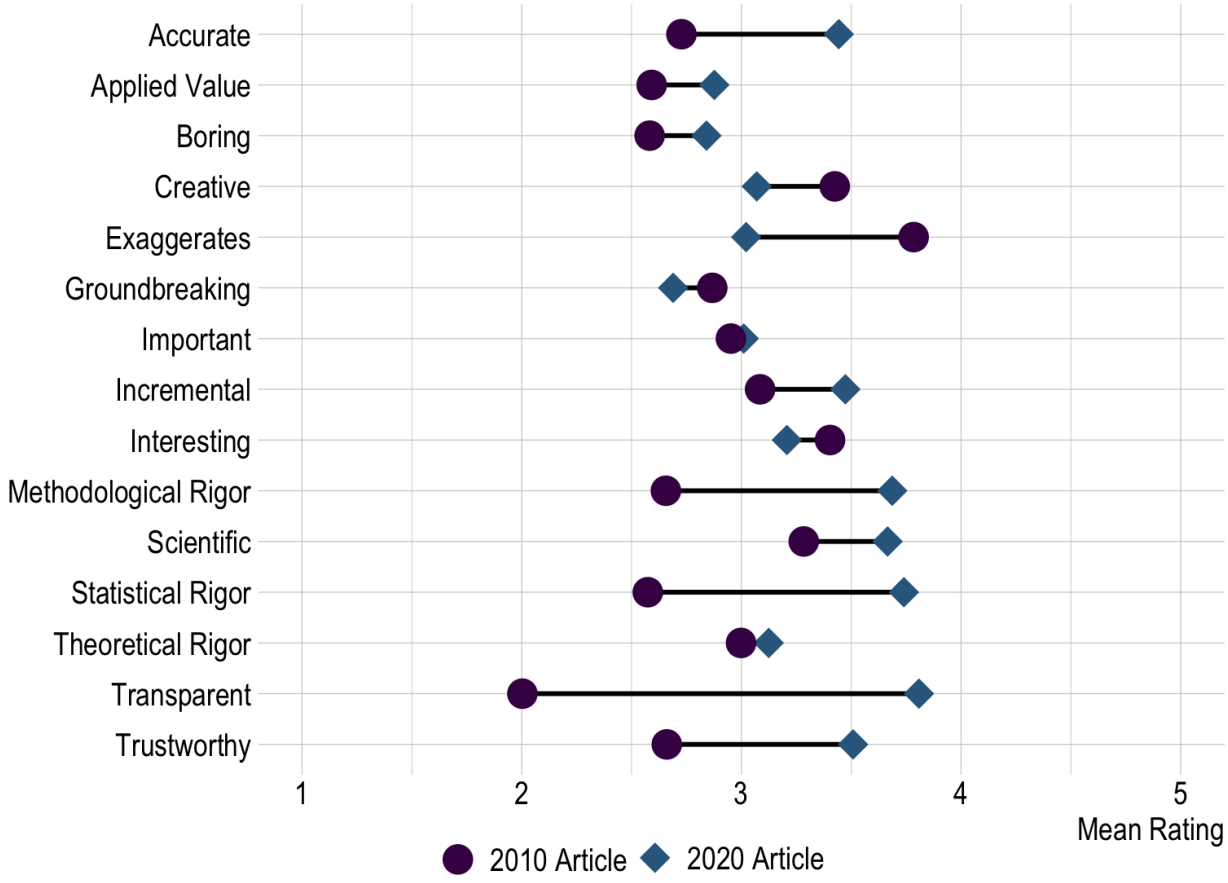
Overall, researchers' perceptions of the typical article published in social and personality psychology were not particularly positive or inspiring. As seen in Figure 8, most characteristics were, on average, rated around the midpoint of the bipolar scale. How researchers conceptualized the meaning of the midpoints presents an interesting measurement question. We suspect a rating of 3 was likely viewed as, for example, somewhat trustworthy, neutral, or neither untrustworthy nor trustworthy. Neither the typical article published in 2010 or 2020 had a single characteristic that reached a mean rating of 4 (when scored in the positive direction) on the 5-point scale. For example, perceptions of how accurate the typical article was had a mean rating of 2.73 in 2010 and 3.44 in 2020. While 2020 articles were viewed more positively than 2010 articles, neither ratings suggest that researchers view the typical article published as all that accurate. Similar conclusions can be reached for most—if not all—the other characteristics.

Some of the characteristics evaluated are arguably not essential for the typical article in a healthy science to exemplify (e.g., not all research need be creative, groundbreaking, or theoretically rigorous). However, it is difficult to argue with the importance of published research being accurate, methodologically rigorous, scientific, statistically rigorous,

transparently reported, and trustworthy. What does it mean for a field to view the typical research published within it so unimpressively? Are perceptions of the validity of research similarly unremarkable?

Figure 8

Mean Ratings of the Typical Article Published in 2010 and 2020



Note. This figure presents researchers' mean ratings of the typical article published in 2010 (i.e., circle) and 2020 (i.e., diamond) in social and personality psychology. The line connecting the two represents the difference between the two means.

Table 5

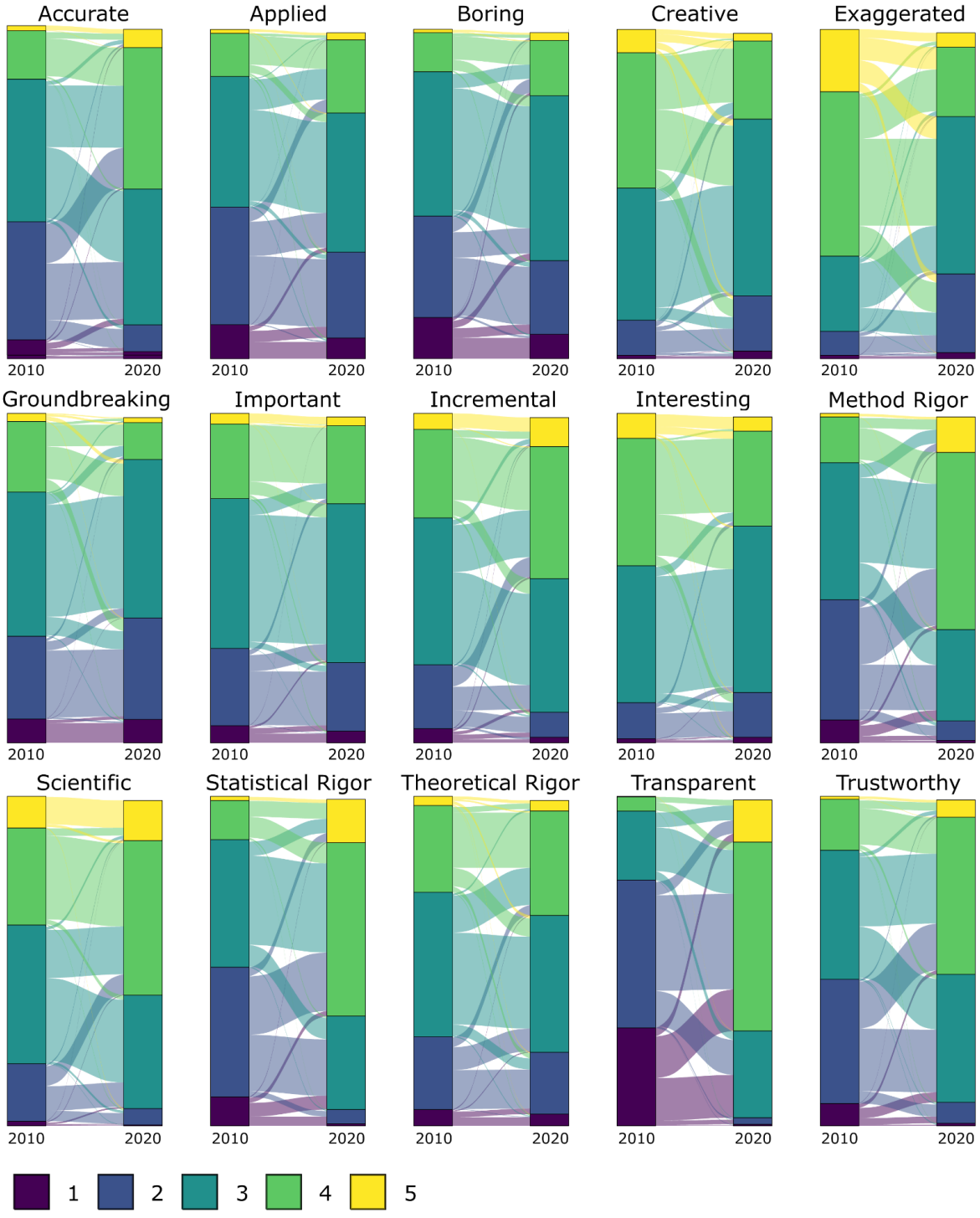
Comparisons of Researchers' Perceptions of the Typical Article Published in 2010 and 2020

	Mean Article Ratings (*SD*)		Bayesian Paired Samples t-tests		
	2010	2020	BF_{10}	BF_{01}	Posterior Median [95% CI]
Accurate	2.73 (0.82)	3.44 (0.77)	7.33×10^{75}	—	0.72 [0.66, 0.79]
Applied value	2.59 (0.89)	2.88 (0.90)	1.30×10^{16}	—	0.28 [0.22, 0.34]
Boring	2.58 (0.89)	2.84 (0.87)	1.12×10^{14}	—	0.26 [0.20, 0.33]
Creative	3.43 (0.82)	3.07 (0.77)	5.72×10^{21}	—	-0.36 [-0.42, -0.29]
Exaggerates Findings	3.78 (0.87)	3.02 (0.84)	3.18×10^{71}	—	-0.76 [-0.84, -0.69]
Groundbreaking	2.87 (0.92)	2.69 (0.82)	4.36×10^{6}	—	-0.18 [-0.24, -0.12]
Important	2.95 (0.89)	3.01 (0.84)	—	1.51	0.06 [0.01, 0.11]
Incremental	3.08 (0.91)	3.47 (0.83)	2.77×10^{25}	—	0.38 [0.33, 0.45]
Interesting	3.40 (0.83)	3.21 (0.79)	4.62×10^{8}	—	-0.19 [-0.25, -0.14]
Methodologically Rigorous	2.66 (0.84)	3.69 (0.77)	5.95×10^{116}	—	1.04 [0.97, 1.11]
Scientific	3.28 (0.91)	3.67 (0.77)	3.23×10^{29}	—	0.38 [0.32, 0.44]
Statistically Rigorous	2.57 (0.86)	3.74 (0.76)	7.32×10^{157}	—	1.17 [1.11, 1.23]
Theoretically Rigorous	3.00 (0.89)	3.12 (0.88)	90.17	—	0.13 [0.06, 0.19]
Transparently Reported	2.00 (0.83)	3.81 (0.69)	1.51×10^{220}	—	1.81 [1.74, 1.88]
Trustworthy	2.66 (0.85)	3.51 (0.73)	2.54×10^{95}	—	0.85 [0.79, 0.92]

Note. This table reports descriptives and the results of Bayesian paired samples t-tests: BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis); Posterior Median [95% CI] = Median and 95% credible interval of the posterior distribution.

Figure 9

Mapping Change in Perceptions of the Typical Article Published in 2010 vs. 2020



Note. This figure connects researchers' ratings of the typical article published in 2010 (the bar on the left) and the typical article published in 2020 (the bar on the right) separated by characteristic. The colours represent the value selected from the Likert-type response scale from 1 (*Not at all*) to 5 (*Extremely*). The thin lines connecting the bars represent how an individual rated the typical 2010 article relative to how they rated the typical 2010 article.

The Four Validities

Researchers' perceptions of the validity of the typical research published in social and personality psychology can be seen in Figure 10 and Table 6. The typical 2010 article was, on average, rated below the midpoint (1 = *Very low* to 5 = *Very high*) across all four validities. Of the four validities, the typical 2010 article was rated the lowest on external validity, followed by statistical, construct, and internal validity. Researchers' perceived validity to have improved over time, as mean ratings of each validity were higher for the typical article published in 2020 than 2010 (all $BF_{10s} > 1.79 \times 10^{46}$, indicating extreme evidence for the alternative hypotheses). The largest perceived change over time was in perceptions of statistical validity, which increased by nearly one point (see Figure 11 for change in individual ratings between timepoints). Unlike the typical 2010 article, the validity of the typical 2020 article was, on average, rated above the midpoint across the four validities—with the exception of external validity, which fell slightly below. The typical 2020 article was perceived to be the strongest in statistical validity and the weakest in external validity.

While researchers appear to believe progress has been made over the last decade in improving the validity of published research, their ratings also suggest the field still has much room left to grow. For an overview of recent reform efforts to improve the four validities in psychology, see Vazire et al. (2022). Much attention has been directed towards increasing the statistical validity of research in psychology over the past decade (e.g., to

decrease QRPs, increase statistical power, develop tools for identifying statistical errors). Thus, it is perhaps unsurprising that researchers expected the most improvements to be in statistical validity.

External validity, in particular, appears to be seen as an area of weakness by researchers in our sample. Concerns about the external validity of findings in psychology—or the lack thereof were highlighted in 2010 by Henrich et al. who criticized research in psychology (and other fields) for largely relying only on WEIRD samples (Western, Educated, Industrialized, Rich, and Democratic). This article has been highly cited in the years that followed (accruing ~11,309 by 2022). Despite discussions about the problem of WEIRD samples persisting, little has changed on this front over the last decade (Pollet & Saxton, 2019). So little, moreover, that Apicella et al. (2020) argue that "the needle hasn't moved" (p. 322) in their review discussing the lack of sample diversity in research published in the last decade. Thus, researchers' perceptions of the external validity of the typical article may overestimate the extent to which improvements have actually been made during this period.

Figure 10

Perceptions of the Validity of the Typical Article Published in 2010 and 2020

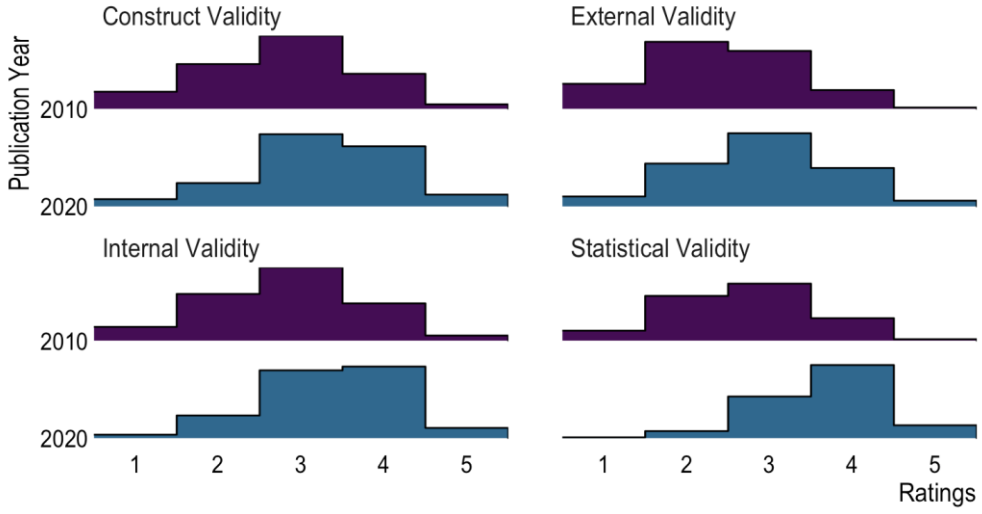


Table 6

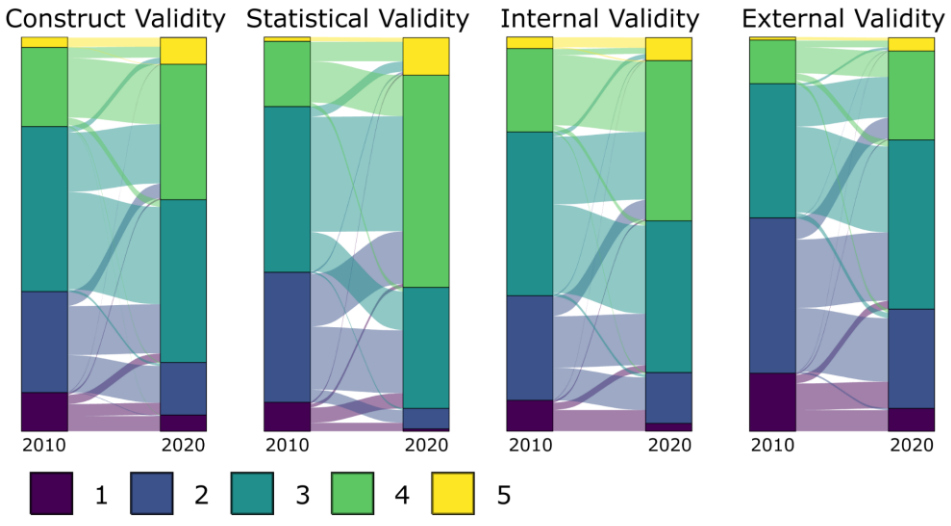
Comparisons of Perceptions of the Validity of the Typical Article Published in 2010 and 2020

	Article Ratings *M* (*SD*)		Bayesian Paired Samples t-tests		
	2010	2020	BF_{10}	BF_{01}	Posterior Median [95% CI]
Construct Validity	2.80 (0.96)	3.26 (0.92)	1.79×10^{46}	—	0.46 [0.41, 0.53]
Statistical Validity	2.71 (0.87)	3.67 (0.75)	1.40×10^{128}	—	0.96 [0.90, 1.01]
Internal Validity	2.85 (0.94)	3.35 (0.85)	9.86×10^{52}	—	0.51 [0.45, 0.57]
External Validity	2.44 (0.90)	2.92 (0.92)	1.044×10^{47}	—	0.49 [0.43, 0.54]

Note. This table reports descriptives and the results of Bayesian paired samples t-tests: BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis); Posterior Median [95% CI] = Median and 95% credible interval of the posterior distribution.

Figure 11

Mapping Change in Perceptions of the Validity of the Typical Article in 2010 vs. 2020



Note. This figure connects researchers' ratings of the typical article published in 2010 (the bar on the left) and the typical article published in 2020 (the bar on the right) separated by characteristic. The colours represent the value selected from the Likert-type response scale from 1 (*Not at all*) to 5 (*Extremely*). The thin lines connecting the bars represent how an individual rated the typical 2010 article relative to how they rated the typical 2010 article.

Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in the Published Research

Perceived improvements over time were also observed in researchers' estimates of the replicability and reproducibility of empirical studies reported in published research articles in social and personality psychology (see Figure 12 and Table 7). Researchers estimated that fewer than half (42.47%) of empirical studies published in 2010 would replicate, compared to just over half (54.58%) in 2020. Perceptions of reproducibility were more optimistic, increasing from 62.52% for 2010 studies to 75.06% for 2020 studies. Overall, researchers perceived p-hacking to have occurred in less than half of 2010 studies, down to around a third of 2020 studies (46.07% to 30.75%). Fraudulent data or results were thought to be rare, and to have also decreased very slightly over time (7.01% in 2010 to 5.74% in 2020).

Researchers' estimates of replicability, reproducibility, and p-hacking appeared to vary greatly, as seen in Figure 12. Estimates of the prevalence of p-hacking seemed to elicit the most disagreement among researchers, particularly for empirical studies published in 2010 which resulted in a fairly wide and flat distribution. While many researchers disagreed on just how many studies would fare on each estimate, they seemed largely in agreement that the published literature had improved on these qualities over time.

How do researchers' estimates compare to previous replication efforts? The Open Science Collaboration replicated 100 studies published in psychology journals in 2008 and

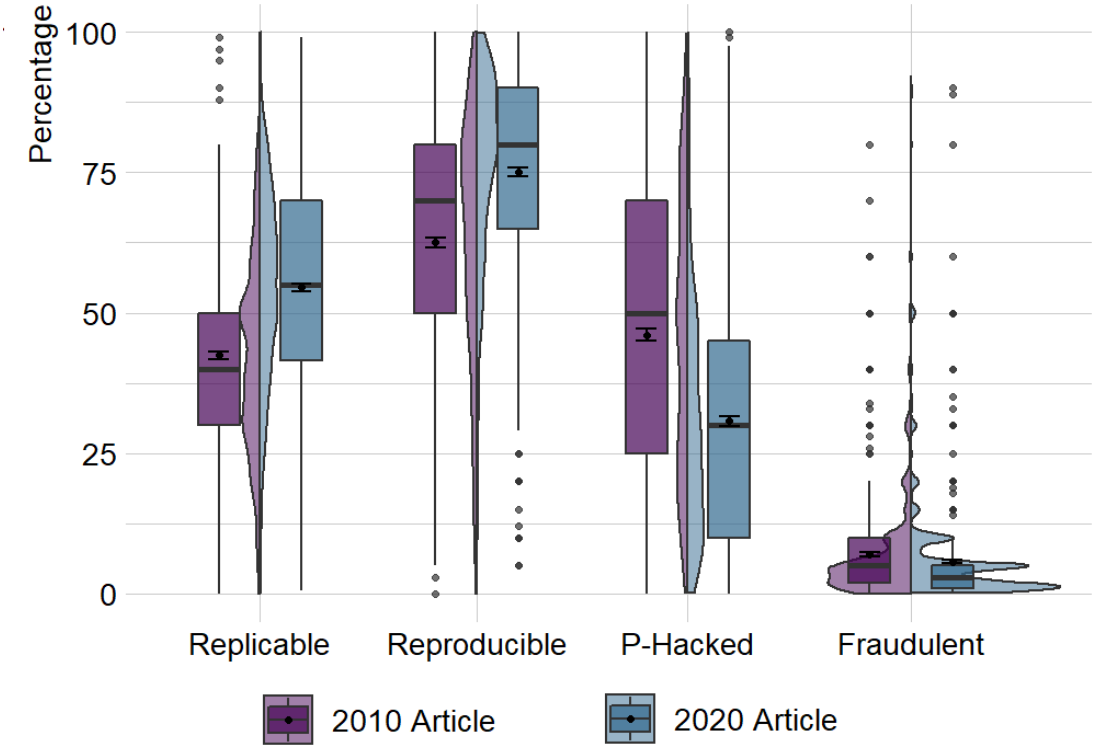
found significant results in only 36% of the studies conducted (Open Science Collaboration, 2015). Reviewing the outcomes of 307 replication studies in psychology, Nosek et al. (2021) estimated that 64% replicated effects similar to the original studies. Estimating replicability rates is complex, and it should be noted that many replications in psychology have selected studies with fairly straightforward designs (e.g., correlational studies, comparing two groups), and much less is known about the replicability of studies with more complicated designs. What constitutes a replication study (e.g., direct, conceptual) and what criteria should be for a "successful" replication continue to be debated in psychology. The variability in researchers' estimates of replicability may reflect some of these uncertainties. Researchers' perceptions were similar to those reported in Motyl et al. (2017), where researchers expected that slightly more than 50% of 2015 studies published in four popular journals in social and personality psychology would replicate compared to ~45% of studies in 2005.

Endeavors to test the analytic reproducibility of key findings in published articles, in cases when data were accessible and without assistance from the original authors, have revealed much lower rates than those estimated by our sample including 33% in economics articles (Chang & Li, 2015), 31% of articles published in *Cognition* (Hardwicke et al., 2018), and 36% published in *Psychological Science* (Hardwicke et al., 2021). However, the item used to estimate reproducibility did not state whether or require that the original data were publicly available, so researchers' response should not be taken to mean that they believe the rates of open data to be so high. We also did not specify whether "the same results" meant that the same general pattern of results would be found, the same statistical conclusions would be reached, or the precisely identical values would be observed. Instead, researchers were free to use their own definition of what they believe it means for results to be reproducible.

Past research exploring researchers' perceptions of the prevalence of questionable research practices and self-reports on their own research practices have generally found low rates of perceived and admitted fraud (Fanelli, 2009; Fiedler & Schwarz, 2016; John et al., 2012). Researchers in our sample similarly estimated fraud to be fairly rare in 2010 and 2020. Some outliers can be seen in Figure 12, but researchers agreed far more on this estimate than in previous estimates.

Figure 12

Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 and 2020



Note. The black dots within the boxplots represent the means with 95% CI bands. The thicker horizontal lines within the boxplots represent the medians.

Table 7

Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 and 2020

	% of Articles *M* (*SD*)		Bayesian Paired Samples t-tests		
	2010	2020	BF_{10}	BF_{01}	Posterior Median [95% CI]
Replicable	42.47 (17.75)	54.58 (17.36)	3.69×10^{104}	—	-12.03 [-12.85, -11.17]
Reproducible	62.52 (22.46)	75.06 (18.19)	4.68×10^{95}	—	-12.51 [-13.48, -11.51]
P-hacked results	46.07 (25.09)	30.75 (21.27)	4.48×10^{74}	—	15.19 [13.89, 16.61]
Fraudulent results/data	7.01 (9.26)	5.74 (9.36)	2.17×10^{5}	—	1.26 [0.82, 1.71]

Note. This table reports descriptives and the results of Bayesian paired samples t-tests: BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis); Posterior Median [95% CI] = Median and 95% credible interval of the posterior distribution.

Researchers' Perceptions of Their Own Work

Researchers' ratings of their own work can be seen in Figure 13. Compared to the typical article published in 2020, researchers generally viewed their work more positively (see Figure 14 and Table 8 where nine out of fifteen tests produced $BF_{10} > 100$). The largest differences (over half a point on the five-point scale) were in researchers viewing their own work as less exaggerated and more accurate and trustworthy than the typical article published in 2020. Researchers also considered their work to be more creative, interesting, scientific, methodologically and theoretically rigorous, and less boring. While not observed across every variable, researchers perceiving their own work to be higher quality (e.g., more trustworthy, more accurate) than the typical article is in line with previous findings

suggesting a “disappointment gap” between scientists’ own beliefs and reported behaviors and how they perceive other scientists to behave (Anderson, 2000; Anderson et al., 2007).

Figure 13

Researchers’ Ratings of Their Own Work

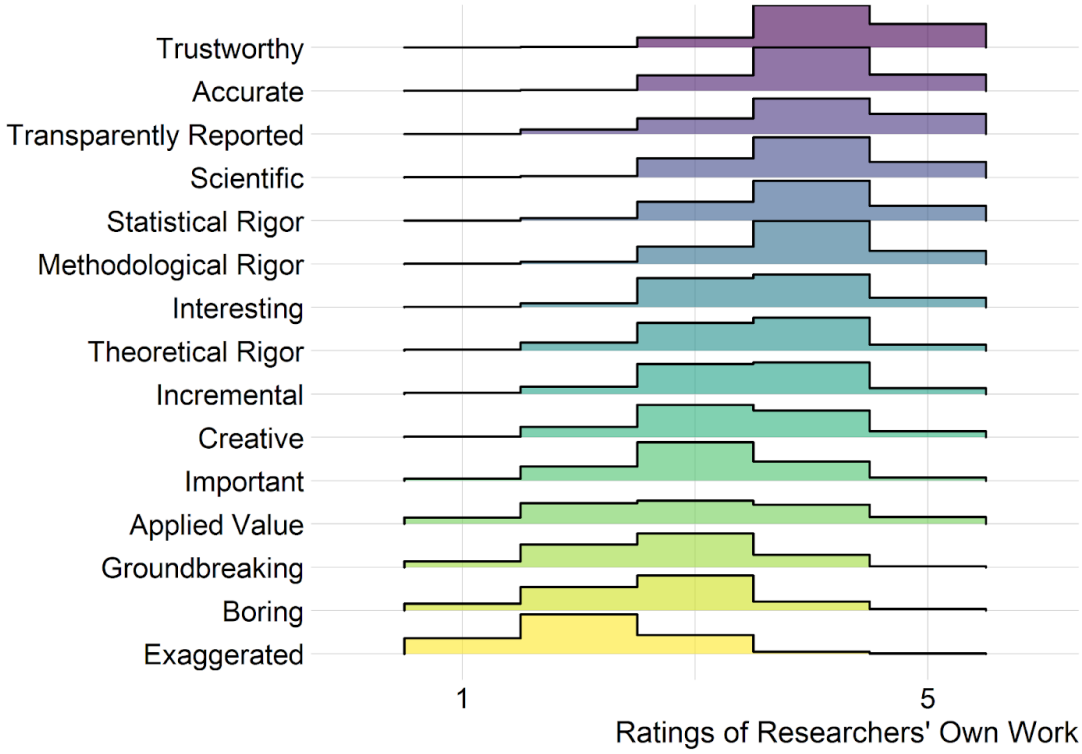
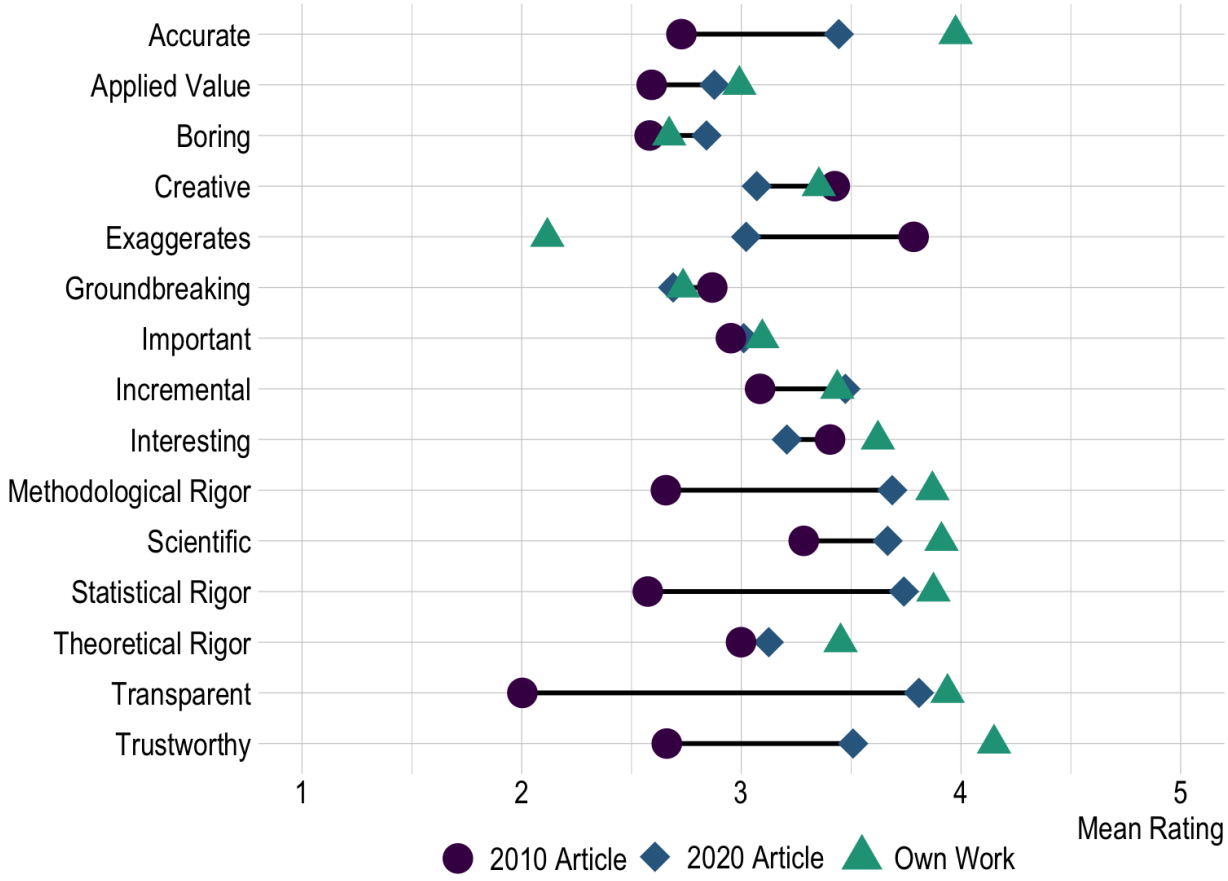


Figure 14

Researchers' Perceptions of Their Own Work and the Typical Article Published in 2010 and 2020



Note. This figure presents mean ratings of the typical article published in 2010 and 2020 in social and personality psychology and researchers' ratings of their own work. The line connecting the two represents the difference between means for 2010 and 2020 articles.

Table 8

Comparisons of Researchers' Perceptions of Their Own Work and the Typical Article Published in 2020

	Mean Ratings (*SD*)		Bayesian Paired Samples t-tests		
	Own Work	2020 Article	BF_{10}	BF_{01}	Posterior Median [95% CI]
Accurate	3.98 (0.70)	3.44 (0.77)	7.22×10^{42}	—	-0.53 [-0.60, -0.46]
Applied value	2.99 (1.11)	2.88 (0.90)	—	1.38	-0.11 [-0.19, -0.02]
Boring	2.67 (0.88)	2.84 (0.87)	104.58	—	0.17 [0.08, 0.25]
Creative	3.35 (0.85)	3.07 (0.77)	8.09×10^{6}	—	-0.27 [-0.36, -0.19]
Exaggerates Findings	2.12 (0.78)	3.02 (0.84)	3.85×10^{79}	—	0.90 [0.82, 0.98]
Groundbreaking	2.73 (0.88)	2.69 (0.82)	—	15.36	-0.04 [-0.11, 0.04]
Important	3.10 (0.83)	3.01 (0.84)	—	1.71	-0.08 [-0.14, -0.01]
Incremental	3.44 (0.84)	3.47 (0.83)	—	15.31	0.04 [-0.05, 0.11]
Interesting	3.62 (0.79)	3.21 (0.79)	1.90×10^{22}	—	-0.41 [-0.48, -0.34]
Methodologically Rigorous	3.87 (0.74)	3.69 (0.77)	1.77×10^{3}	—	-0.18 [-0.26, -0.10]
Scientific	3.91 (0.74)	3.67 (0.77)	1.47×10^{9}	—	-0.24 [-0.31, -0.17]
Statistically Rigorous	3.87 (0.76)	3.74 (0.76)	10.09	—	-0.13 [-0.21, -0.06]
Theoretically Rigorous	3.45 (0.85)	3.12 (0.88)	7.72×10^{9}	—	-0.32 [-0.40, -0.23]
Transparently Reported	3.94 (0.85)	3.81 (0.69)	4.49	—	-0.12 [-0.20, -0.04]
Trustworthy	4.15 (0.68)	3.51 (0.73)	1.32×10^{53}	—	-0.64 [-0.72, -0.57]

Note. This table reports descriptives and the results of Bayesian paired samples t-tests: BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis); Posterior Median [95% CI] = Median and 95% credible interval of the posterior distribution.

Part II. Understanding What Researchers Value When Evaluating Research

Whereas in Part I we explored researchers' perceptions of the characteristics of the published literature in social and personality psychology and of their own work, in Part II we examine how important researchers believe those characteristics to be when assessing research quality. This adds to our understanding of what researchers say they value in the evaluation of the quality of research—which is currently limited—and allows for greater contextualization of the results presented in Part I. Below, we describe how important researchers consider certain characteristics to be for evaluating research quality and test whether what they consider important is related to their perceptions of these characteristics in the published literature and their own work.

What Do Researchers Consider Important for Evaluating Research Quality?

Researchers' ratings of the importance of various characteristics when evaluating the quality of research in social and personality psychology can be seen in Figure 15, ordered from most to least important. As shown in the distributions of responses, researchers considered statistical validity the most important, followed closely by construct validity, internal validity, methodological rigor, and transparency. Although most rated research being interesting, creative, and having applied value on the important side, there was greater variability in these responses suggesting less agreement among researchers on the importance of these qualities. Of all the characteristics, how groundbreaking the research is was the only one that researchers, on average, rated below the midpoint of the scale (see Table 9).

Figure 15

Researchers Ratings of the Importance of Characteristics in Evaluating Research Quality

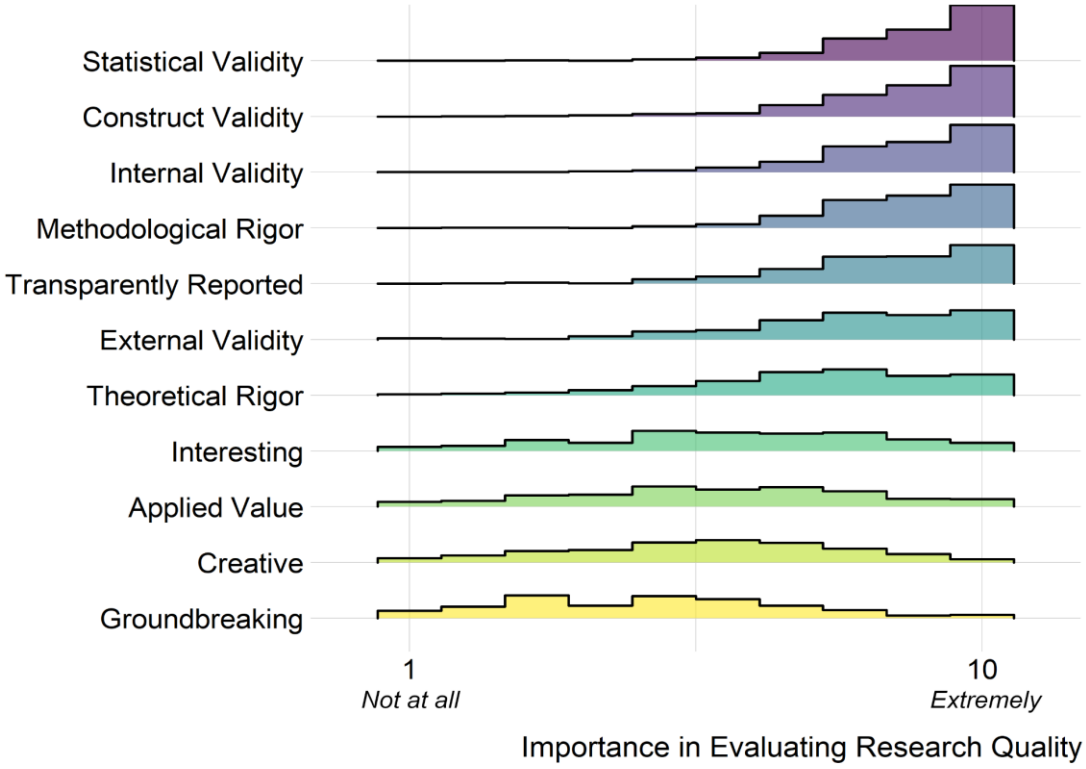


Table 9

Descriptives of Researchers' Ratings of the Importance of Characteristics for Evaluating Research Quality

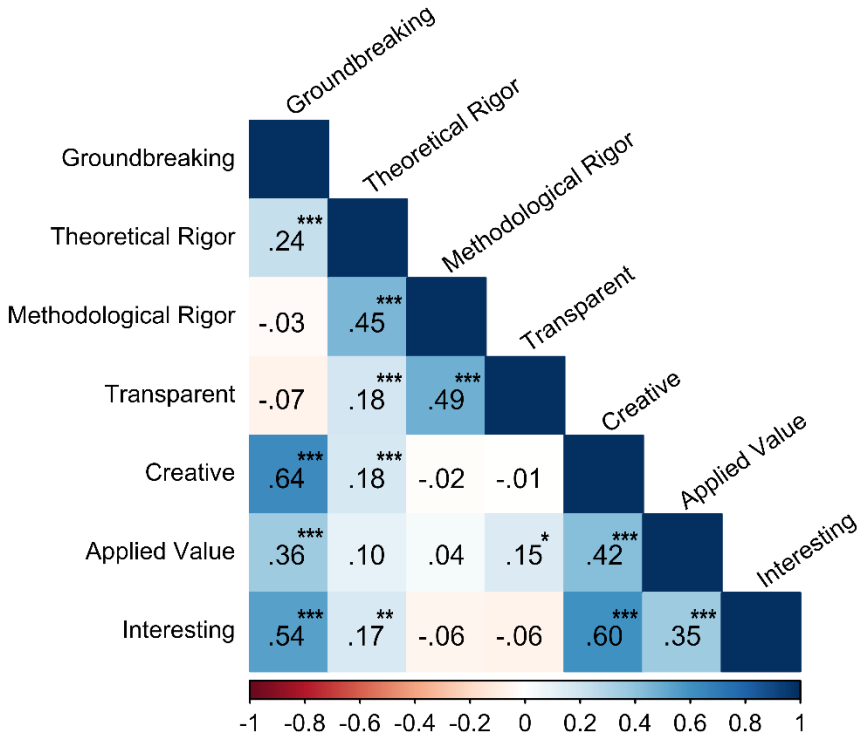
Characteristic	Mean (*SD*)	Median	Characteristic	Mean (*SD*)	Median
Applied Value	5.83 (2.32)	6	Construct Validity	8.81 (1.40)	9
Creative	5.65 (2.18)	6	Statistical Validity	8.98 (1.26)	9
Groundbreaking	4.79 (2.21)	5	Internal Validity	8.75 (1.39)	9
Interesting	6.10 (2.33)	6	External Validity	7.93 (1.86)	8
Methodologically Rigorous	8.73 (1.30)	9	—	—	—
Theoretically Rigorous	7.44 (1.98)	8	—	—	—
Transparently Reported	8.41 (1.59)	9	—	—	—

Correlations between researchers' ratings of the importance of each characteristic for evaluating research quality are reported in Figure 16a. The strongest correlations are highlighted in Figure 16b, which shows positive correlations between ratings of the importance of research being groundbreaking, interesting, creative, and having applied value. Valuing methodological rigor was associated with valuing both transparency and theoretical rigor.

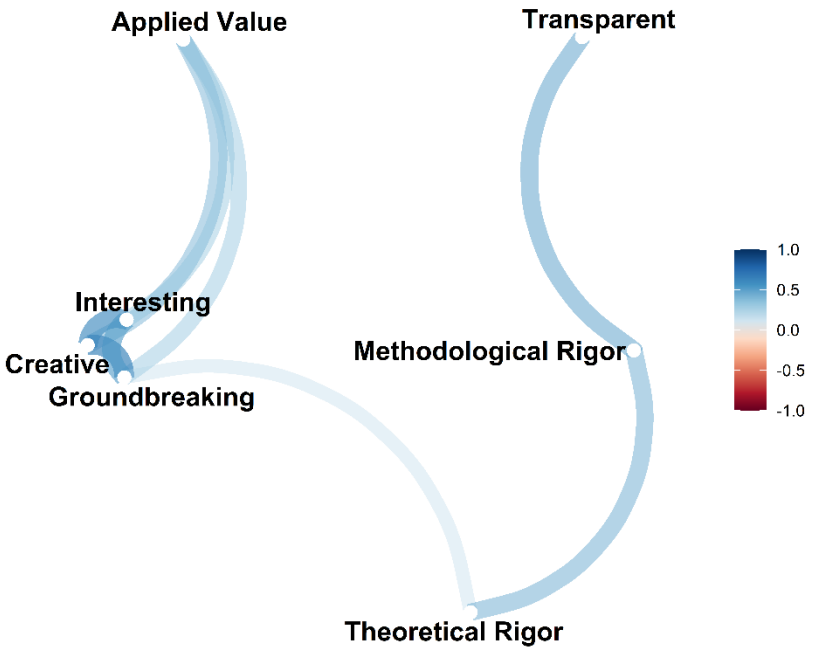
Figure 16

Heatmap and Correlation Network of Researchers' Ratings of the Importance of Characteristics in Evaluating Research Quality

16a)



16b)



Note. * = BF_{10} = > 30; ** = BF_{10} > 100; *** BF_{10} = > 1000. Figure 16b presents only correlations of a minimum of r = .2.

Evaluating Research Quality and Perceptions of Published Research

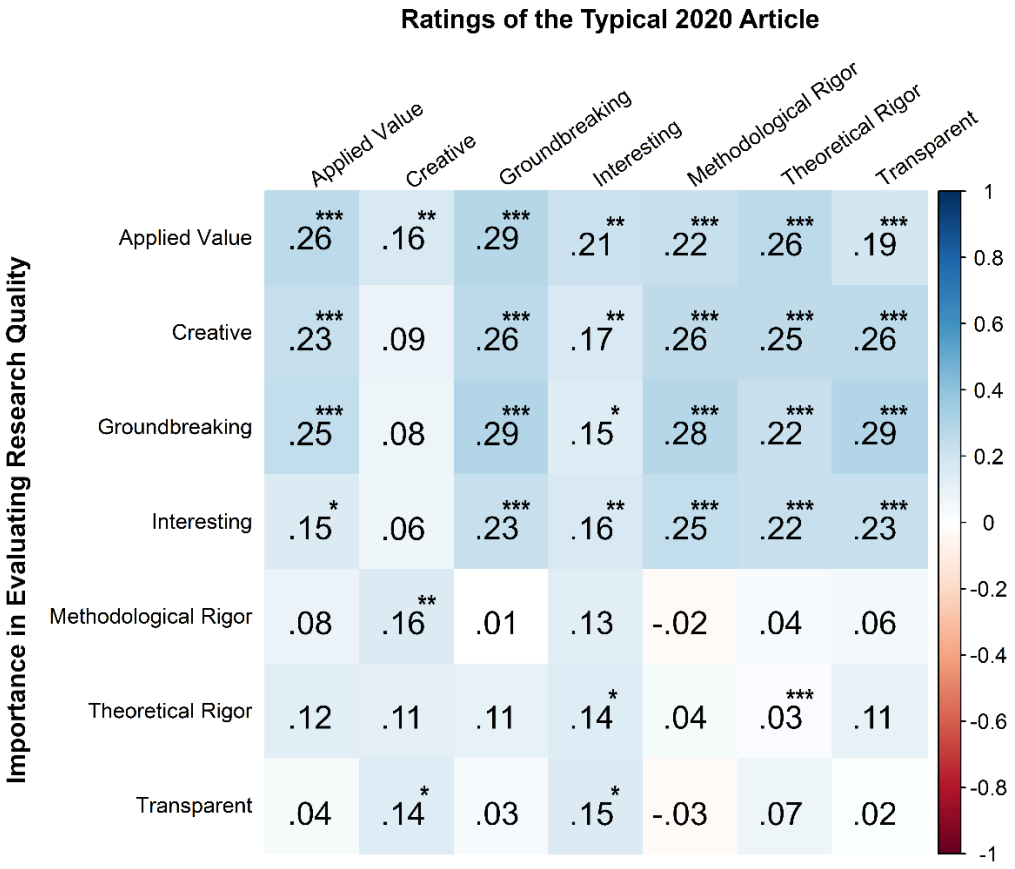
Next, we examined how researchers' views about what qualities are (or are not) important when evaluating research quality relate to their perceptions of the typical article published in 2020 in social and personality psychology. Figure 17 shows these results. Generally, the correlations on the diagonal do not seem to be noticeably stronger than the off-diagonal correlations, suggesting that believing that a characteristic is important does not consistently predict having a more positive perception of the literature on that characteristic. Instead, believing that certain characteristics are important (creative, groundbreaking, interesting, and applied value), was associated with having a more positive view of the typical article published in 2020.

Although researchers' ratings of the importance of each of the four validities in evaluating the quality of research were correlated with one another, none were associated

with how researchers rated the typical article published in 2020 on each validity (see supplemental materials). For the four validities, believing that a validity was important did not predict having more positive perceptions of the literature in that domain. The validity rated the most important (statistical validity), however, was also rated the highest amongst the four for the typical 2020 article. Similarly, the typical 2020 article was rated the lowest in external validity of the four and external validity was also, on average, rated as less important than the other three validities for evaluating quality.

Figure 17

Heatmap of the Correlations Between Ratings of the Importance of Characteristics in Evaluating Research Quality and Perceptions of the Typical 2020 Article



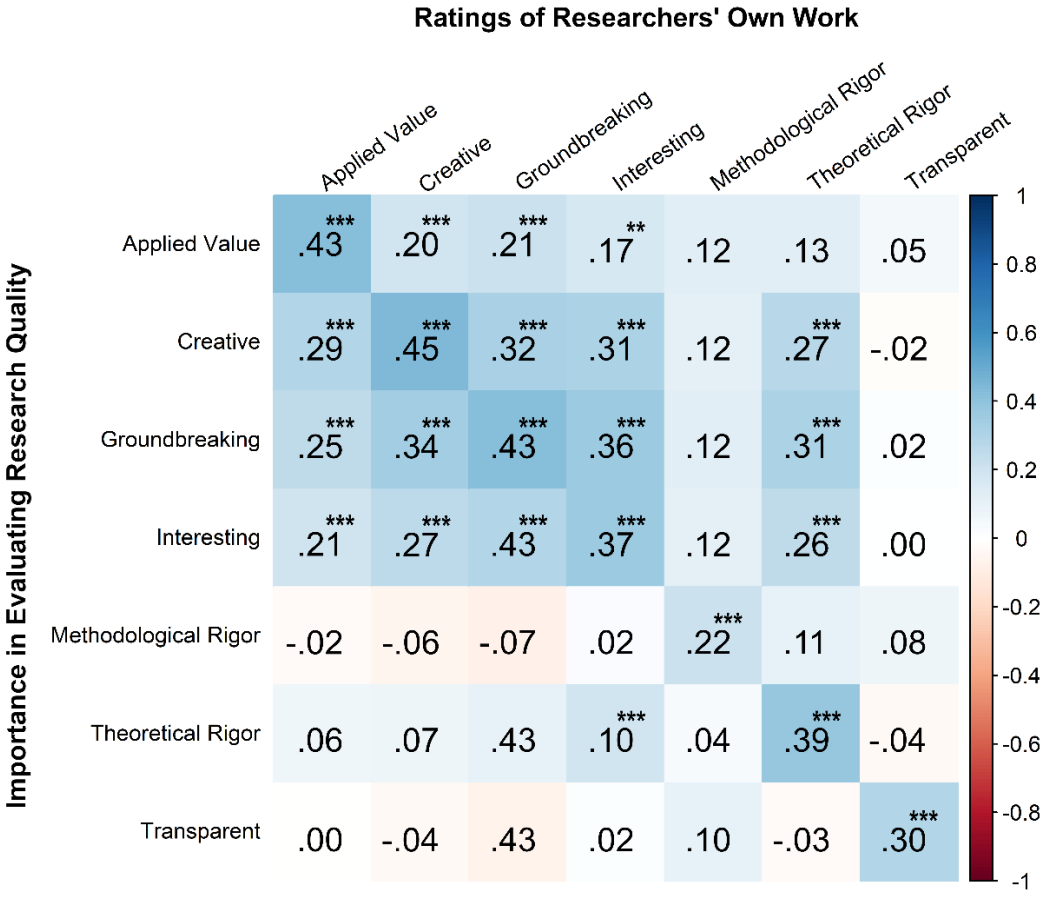
Note. * = BF_{10} = > 30; ** = BF_{10} > 100; *** BF_{10} = > 1000.

Relationship Between What Researchers Consider Important for Evaluating Research and Their Perceptions of Their Own Work

Overall, the more important researchers perceived a characteristic to be for evaluating research quality, the more highly they rated their own work on that characteristic (see Figure 18, diagonal).

Figure 18

Heatmap of the Correlations Between Researchers' Ratings of the Importance of Characteristics in Evaluating Research Quality and Their Own Work



Note. * = BF_{10} = > 30; ** = BF_{10} > 100; *** BF_{10} = > 1000.

Part III. Understanding Individual Differences Among Researchers' Perceptions

To identify potential individual differences in how researchers view work in their field, we tested whether self-reported intellectual humility, support for open science, and career stage were related to perceptions of research published in 2010 and 2020, perceptions of their own work, and perceptions of how their own work compares to the typical research article published in social and personality psychology.

Self-Reported Intellectual Humility

Are researchers who self-report being more intellectually humble more critical of the published literature—and of their own work? Or, do researchers who view themselves as more intellectually humble see their field and own research more positively?

Self-Reported Intellectual Humility and Perceptions of the Published Literature

Self-Reported Intellectual Humility and Perceptions of the Typical Published Article. Overall, researchers' scores on self-reported intellectual humility did not appear to be related to their perceptions of the typical article published in 2010 or 2020, or the difference between the two timepoints (see Table 10). Of the 15 characteristics evaluated, only one test result provided moderate evidence of a positive, but weak, correlation between self-reported intellectual humility and ratings of how scientific the typical article published in 2010 was. Otherwise, most tests found moderate-to-strong evidence favoring the null hypotheses that self-reported intellectual humility is not associated with researchers' perceptions of the typical article published in 2010 or 2020 or the difference between the two (see Table 10). Perceptions of the validity of the typical articles were also not associated with self-reported intellectual humility (see Table 11).

Self-Reported Intellectual Humility and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in the Published Research. Self-reported intellectual humility did not appear to be related to researchers' estimates of replicability and reproducibility rates and the prevalence of p-hacking and fraud in empirical studies reported in articles in the published literature (in 2010 and 2020) or whether they changed over time (see Table 12).

Self-Reported Intellectual Humility and Perceptions of Researchers Own Work

Whereas self-reported intellectual humility appeared largely unrelated to researchers' perceptions of the typical published research, extreme evidence was found for the alternative hypotheses that self-reported intellectual humility is related to researchers' ratings of their own work on six of the characteristics evaluated (out of 15), along with strong evidence for two others (Table 13). Overall, the more intellectually humble researchers said they were, the more positively they perceived their work (i.e., more trustworthy, scientific, accurate, statistically and methodologically rigorous, less exaggerated), though the associations were weak-to-moderate in magnitude.

Self-Reported Intellectual Humility and How Researchers Perceive Their Own Work Compared to the Typical Published Article. To test how intellectual humility relates to how researchers view their own work compared to the typical published article, we calculated difference scores by subtracting researchers' ratings of the typical 2020 article from their ratings of their own work. As seen in Table 13, we found moderate-to-very-strong evidence in favor of the alternative hypotheses that self-reported intellectual humility is associated with seeing their own work more positively than the typical published article for 7 out of the 15 characteristics. The strongest evidence was found for an association between seeing oneself as intellectually humble and rating one's own work as

less exaggerated than the typical article published in 2020. While the evidence for these associations is strong, the magnitude of the effects is quite weak.

Table 10

Relationship Between Self-Reported Intellectual Humility and Researchers' Perceptions of the Typical Article Published in 2010 and 2020

Self-Reported Intellectual Humility	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Accurate	0.02 [-0.06, 0.09]	—	18.52	0.01 [-0.07, 0.09]	—	19.23	-0.01 [-0.08, 0.07]	—	19.88
Applied Value	0.04 [-0.04, 0.11]	—	11.76	0.05 [-0.03, 0.12]	—	9.43	0.02 [-0.06, 0.09]	—	18.78
Boring	0.002 [-0.08, 0.07]	—	20.41	-0.05 [-0.12, 0.03]	—	10.20	-0.06 [-0.13, 0.02]	—	6.44
Creative	-0.004 [-0.08, 0.07]	—	20.20	0.004 [-0.07, 0.08]	—	20.04	0.01 [-0.06, 0.09]	—	19.66
Exaggerates Findings	0.05 [-0.03, 0.12]	—	10.10	0.06 [-0.02, 0.14]	—	6.67	0.01 [-0.07, 0.08]	—	19.67
Groundbreaking	0.04 [-0.03, 0.12]	—	11.63	0.03 [-0.06, 0.10]	—	15.63	-0.01 [-0.10, 0.06]	—	18.91
Important	0.007 [-0.07, 0.09]	—	20.00	0.02 [-0.05, 0.10]	—	16.67	0.0| [-0.04, 0.11]	—	14.60
Incremental	0.11 [0.04, 0.19]	2.75	—	0.08 [0.01, 0.16]	—	2.16	-0.04 [-0.11, 0.04]	—	13.01
Interesting	0.06 [-0.02, 0.14]	—	6.17	0.02 [-0.06, 0.10]	—	18.18	-0.04 [-0.12, 0.03]	—	11.60
Methodologically Rigorous	0.07 [-0.01, 0.15]	—	3.60	-0.008 [-0.08, 0.07]	—	19.61	-0.07 [-0.15, 0.00]	—	3.59
Scientific	0.12 [0.05, 0.20]	7.69	—	0.04 [-0.03, 0.12]	—	10.75	-0.10 [-0.17, -0.02]	—	1.09
Statistically Rigorous	0.06 [-0.01, 0.14]	—	7.14	0.003 [-0.07, 0.08]	—	20.13	-0.06 [-0.13, 0.02]	—	7.72
Theoretically Rigorous	0.04 [-0.03, 0.12]	—	11.11	-0.05 [-0.13, 0.03]	—	9.01	-0.10 [-0.17, -0.02]	1.05	—
Transparently Reported	0.02 [-0.06, 0.09]	—	18.52	0.06 [-0.02, 0.13]	—	6.58	0.03 [-0.05, 0.11]	—	16.39
Trustworthy	0.05 [-0.03, 0.12]	—	9.17	0.04 [-0.04, 0.12]	—	10.53	-0.01 [-0.09, 0.07]	—	19.44

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 11

Relationship Between Self-Reported Intellectual Humility and Researchers' Perceptions of the Validity of the Typical Article Published in 2010 and 2020

Support for Open Science	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Construct Validity	0.11 [0.03, 0.19]	2.21	—	0.05 [-0.02, 0.13]	—	8.70	-0.07 [-0.14, 0.01]	—	3.92
Statistical Validity	0.05 [-0.02, 0.13]	—	8.20	0.05 [-0.03, 0.13]	—	9.43	-0.009 [-0.09, 0.07]	—	20.00
Internal Validity	0.06 [-0.01, 0.14]	—	5.85	0.09 [0.01, 0.16]	—	1.34	0.03 [-0.04, 0.11]	—	15.38
External Validity	0.04 [-0.03, 0.12]	—	10.99	0.04 [-0.03, 0.12]	—	11.76	0.002 [-0.07, 0.08]	—	20.41

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 12

Relationship Between Self-Reported Intellectual Humility and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 and 2020

Self-Reported Intellectual Humility	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Replicable	0.05 [-0.03, 0.12]	—	9.71	0.06 [-0.02, 0.14]	—	7.09	0.01 [-0.06, 0.09]	—	18.87
Reproducible	0.03 [-0.05, 0.11]	—	14.93	0.04 [-0.05, 0.11]	—	13.16	-.002 [-0.08, 0.07]	—	20.00
P-hacked results	0.007 [-0.07, 0.08]	—	19.61	-0.03 [-0.11, 0.05]	—	14.93	-0.04 [-0.12, 0.04]	—	11.11
Fraudulent results/data	0.0005 [-0.08, 0.08]	—	20.00	-0.03 [-0.11, 0.05]	—	15.38	-0.05 [-0.12, 0.04]	—	10.11

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 13

Relationship Between Self-Reported Intellectual Humility and Researchers' View of Their Own Work and the Difference Between Their Work Compared to the Typical 2020 Article

Self-Reported Intellectual Humility	Researchers' Own Work			Researchers' Own Work – 2020 Article		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Accurate	0.19 [0.11, 0.26]	3779.72	—	0.13 [0.06, 0.21]	17.13	—
Applied Value	0.08 [0.01, 0.16]	—	2.21	0.04 [-0.04, 0.11]	—	12.82
Boring	0.21 [-0.06, 0.09]	—	17.86	0.06 [-0.02, 0.13]	—	7.52
Creative	0.16 [0.09, 0.24]	289.10	—	0.12 [0.04, 0.19]	3.80	—
Exaggerates Findings	-0.14 [-0.21, -0.06]	25.74	—	-0.15 [-0.23, -0.08]	90.44	—
Groundbreaking	0.07 [0.00, 0.15]	—	3.28	0.04 [-0.04, 0.11]	—	13.16
Important	0.13 [0.06, 0.21]	16.24	—	0.10 [0.02, 0.17]	1.03	—
Incremental	0.05 [-0.04, 0.12]	—	10.10	-0.03 [-0.10, 0.05]	—	16.39
Interesting	0.08 [0.00, 0.15]	—	3.27	0.04 [-0.03, 0.12]	1.03	—
Methodologically Rigorous	0.16 [0.09, 0.24]	432.71	—	0.13 [0.06, 0.21]	11.50	—
Scientific	0.19 [0.11, 0.26]	7643.10	—	0.12 [0.05, 0.19]	4.68	—
Statistically Rigorous	0.17 [0.09, 0.24]	580.46	—	0.13 [0.05, 0.20]	7.68	—
Theoretically Rigorous	0.08 [0.00, 0.16]	—	2.48	0.10 [0.02, 0.18]	1.04	—
Transparently Reported	0.11 [0.03, 0.18]	2.18	—	0.04 [-0.03, 0.12]	—	10.64
Trustworthy	0.22 [0.15, 0.29]	3.05 x 105	—	0.12 [0.05, 0.20]	5.59	—

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Support for the Open Science Movement

Support of Open Science and Perceptions of the Published Literature

Support of Open Science and Perceptions of the Typical Published Article.

First, we found that greater support for the open science movement was associated with more negative views of the typical 2010 article across several characteristics (as less accurate, transparently reported, scientific, methodologically, statistically, and theoretically rigorous, and more exaggerated; see Table 14). While the evidence for these associations is very strong, the magnitude of the effects is moderate. Second, unlike ratings of 2010 articles, these patterns of results were not observed for perceptions of the typical article published in 2020. Rather, Bayes factors mostly favored the null hypotheses (i.e., no correlation between support for the open science movement and ratings of the typical 2020 article). However, greater support was associated with viewing the typical 2020 article as less boring and more creative. Third, we examined how perceived change between the typical article published in 2010 and 2020 relates to support for the open science movement (Table 14). Overall, the more researchers support open science, the more positively they see changes between what was typically published in 2010 compared to 2020 (more accurate, interesting, scientific, trustworthy, statistically, methodologically, and theoretically rigorous, and less boring and exaggerated).

Perceptions of the validity of the typical articles published in 2010 and 2020 revealed similar patterns of association with support for the open science movement. The more support researchers reported, the more negatively they rated the typical article published in 2010 across all four validities (construct, statistical, internal, and external validity; see Table 15). Open science support was not associated with perceptions of the validity of 2020 articles but was associated with perceiving more improvements on the four validities over time.

Support of Open Science and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in the Published Research. Consistent with the findings reported above, support for the open science movement was associated with more negative views of empirical studies published in 2010, but not in 2020 (see Table 16). The more researchers supported open science, the less replicable, reproducible, and the more p-hacked they estimated empirical studies in articles published in 2010 to be. However, open science support was not associated with higher estimates of fraud in empirical studies in 2010. Supporting the open science movement was associated with greater perceived change over time (increases in replicability and reproducibility, decreases in p-hacking; see Table 16).

Support of Open Science and Perceptions of Researchers' Own Work

Researchers' support for the open science movement was mostly not associated with how they view their work, with a few exceptions (Table 17) including perceiving their work to be more transparently reported, less groundbreaking, and less theoretically rigorous.

Support of Open Science and How Researchers Perceive Their Own Work Compared to the Typical Published Article. Shifting to how researchers viewed their own work compared to the typical article, we found that greater support for the open science movement was associated with researchers viewing their work—compared to the typical 2020 article—in a mostly negative light (Table 17; more boring and less creative, interesting, groundbreaking, important, and theoretically rigorous), with the exception that supporting the open science movement did predict rating one's work as more transparently reported and more incremental than the typical 2020 article.

Table 14

Relationship Between Support for the Open Science Movement and Researchers' Perceptions of the Typical Article Published in 2010 and 2020

Open Science	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Accurate	-0.23 [-0.31, -0.16]	2.39×10^{6}	—	-0.03 [-0.11, 0.05]	—	14.67	0.20 [0.13, 0.27]	3.53×10^{4}	—
Applied Value	-0.12 [-0.20, -0.05]	5.47	—	-0.05 [-0.13, 0.03]	—	9.70	0.09 [0.02, 0.17]	—	1.41
Boring	-0.02 [-0.09, 0.05]	—	17.16	-0.17 [-0.24, -0.09]	492.68	—	-0.18 [-0.25, -0.10]	1.34×10^{3}	—
Creative	0.0005 [-0.07, 0.08]	—	20.32	0.14 [0.07, 0.22]	48.05	—	0.14 [0.06, 0.21]	32.09	—
Exaggerates Findings	0.16 [0.08, 0.24]	185.76	—	-0.05 [-0.12, 0.03]	—	9.57	-0.20 [-0.27, -0.12]	1.89×10^{4}	—
Groundbreaking	-0.10 [-0.18, -0.03]	1.13	—	-0.007 [-0.08, 0.07]	—	19.86	0.13 [0.05, 0.20]	8.17	—
Important	-0.12 [-0.19, -0.04]	3.87	—	0.04 [-0.04, 0.11]	—	12.83	0.23 [0.16, 0.30]	1.09×10^{6}	—
Incremental	-0.14 [-0.21, -0.06]	19.68	—	-0.13 [-0.21, -0.05]	13.03	—	0.03 [-0.05, 0.11]	—	15.36
Interesting	-0.09 [-0.17, -0.02]	—	1.34	0.11 [0.03, 0.19]	2.24	—	0.23 [0.16, 0.31]	3.37×10^{6}	—
Methodologically Rigorous	-0.24 [-0.31, -0.17]	1.03×10^{7}	—	-0.009 [-0.09, 0.07]	—	19.68	0.22 [0.14, 0.29]	1.85×10^{5}	—
Scientific	-0.17 [-0.24, -0.10]	486.62	—	-0.02 [-0.10, 0.05]	—	16.66	0.19 [0.11, 0.26]	3.97×10^{3}	—
Statistically Rigorous	-0.23 [-0.30, -0.16]	1.55×10^{6}	—	-0.09 [-0.17, -0.01]	—	1.62	0.16 [0.08, 0.23]	229.52	—
Theoretically Rigorous	-0.21 [-0.28, -0.13]	4.96×10^{12}	—	-.001 [-0.08, 0.07]	—	20.21	0.23 [0.15, 0.30]	1.20×10^{6}	—
Transparently Reported	-0.30 [-0.37, -0.23]	1.63×10^{12}	—	-0.07 [-0.14, 0.01]	—	4.94	0.23 [0.16, 0.30]	2.05×10^{6}	—
Trustworthy	-0.28 [-0.35, -0.21]	1.54×10^{10}	—	-0.003 [-0.08, 0.08]	—	20.15	0.28 [0.21, 0.35]	1.43×10^{10}	—

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 15

Relationship Between Support for the Open Science Movement and Researchers' Perceptions of the Validity of the Typical Article Published in 2010 and 2020

Support for Open Science	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Construct Validity	-0.22 [-0.29, -0.14]	4.14×10^5	—	-0.09 [-0.17, -0.02]	—	1.13	0.17 [0.09, 0.24]	862.54	—
Statistical Validity	-0.23 [-0.30, -0.15]	1.67×10^6	—	-0.02 [-0.09, 0.06]	—	18.18	0.24 [0.17, 0.31]	9.39×10^6	—
Internal Validity	-0.20 [-0.28, -0.13]	4.78×10^4	—	-0.05 [-0.13, 0.03]	—	9.71	0.21 [0.13, 0.28]	6.81×10^4	—
External Validity	-0.19 [-0.26, -0.11]	5.81×10^3	—	-0.02 [-0.09, 0.06]	—	16.95	0.19 [0.12, 0.26]	1.45×10^4	—

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 16

Relationship Between Support for the Open Science Movement and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 and 2020

Support for Open Science	2010 Articles			2020 Articles			2020 Articles – 2010 Articles		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Replicable	-0.19 [-0.27, -0.12]	6740.17	—	-.001 [-0.08, 0.08]	—	20.00	0.30 [0.23, 0.37]	5.99×10^{11}	—
Reproducible	-0.14 [-0.22, -0.06]	22.18	—	-.009 [-0.09, 0.07]	—	19.61	0.24 [0.17, 0.31]	6.44×10^{6}	—
P-hacked results	0.14 [0.05, 0.21]	14.17	—	-0.01 [-0.09, 0.06]	—	19.23	-0.22 [-0.30, -0.15]	5.20×10^{5}	—
Fraudulent results/data	0.04 [-0.04, 0.12]	—	12.82	-0.004 [-0.08, 0.07]	—	19.61	-0.07 [-0.15, 0.01]	—	5.03

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Table 17

Relationship Between Support for the Open Science Movement and Researchers' Perceptions of Their Own Work and the Difference Between Their Own Work Compared to the Typical 2020 Article

Support for the Open Science Movement	Researchers' Own Work			Researchers' Own Work – 2020 Article		
	rho [95% CI]	BF_{10}	BF_{01}	rho [95% CI]	BF_{10}	BF_{01}
Accurate	-0.07 [-0.14, 0.01]	—	4.40	-0.04 [-0.12, 0.04]	—	12.53
Applied Value	-0.13 [-0.20, -0.05]	12.39	—	-0.08 [-0.16, -0.01]	—	2.43
Boring	0.07 [-0.01, 0.15]	—	3.07	0.19 [0.12, 0.26]	5373.26	—
Creative	-0.13 [-0.21, -0.05]	13.78	—	-0.20 [-0.27, -0.12]	2.21×10^4	—
Exaggerates Findings	-0.04 [-0.12, 0.03]	—	10.33	0.001 [-0.08, 0.08]	—	20.16
Groundbreaking	-0.19 [-0.27, -0.12]	8074.30	—	-0.16 [-0.23, -0.08]	211.37	—
Important	-0.11 [-0.19, -0.04]	3.14	—	-0.14 [-0.21, -0.06]	23.22	—
Incremental	0.09 [0.01, 0.16]	—	1.40	0.17 [0.10, 0.25]	583.64	—
Interesting	-0.11 [-0.19, -0.04]	3.23	—	-0.18 [-0.26, -0.11]	3031.53	—
Methodologically Rigorous	-0.005 [-0.08, 0.07]	—	20.08	-0.007 [-0.08, 0.07]	—	19.86
Scientific	-0.05 [-0.12, 0.03]	—	10.03	-0.02 [-0.10, 0.05]	—	17.49
Statistically Rigorous	-0.03 [-0.11, 0.04]	—	14.97	0.04 [-0.03, 0.12]	—	11.42
Theoretically Rigorous	-0.20 [-0.27, -0.13]	2.23×10^4	—	-0.15 [-0.23, -0.08]	101.32	—
Transparently Reported	0.14 [0.07, 0.22]	46.79	—	0.17 [0.09, 0.24]	439.63	—
Trustworthy	0.004 [-0.07, 0.08]	—	20.21	0.0005 [-0.07, 0.08]	—	20.16

Note. This table reports the results of Bayesian correlation analyses: rho [95% CI] = estimated correlation coefficient and 95% credible interval; BF_{10} = Bayes factor favoring the alternative hypothesis (relative to the null hypothesis); BF_{01} = Bayes factor favoring the null hypothesis (relative to the alternative hypothesis).

Career Stage

To test the relationship between career stage and researchers' ratings of published research, we conducted Bayesian Independent ANOVAs to compare the predictive performance of a null and alternative hypotheses, where H_1 described researchers' ratings while allowing them to differ by level of seniority, whereas H_0 described author's ratings using the grand mean across all levels of seniority. As preregistered, we excluded participants who did not have a PhD and were also not graduate students (n = 8) from these analyses.

Career Stage and Perceptions of the Published Literature

Career Stage and Perceptions of the Typical Published Article. We found little evidence of differences between researchers of various career stages in their perceptions of the typical article published in 2020. Indeed, most results favored the null hypotheses (see Table 18). Whereas researchers differed little in their perceptions of the typical 2020 article by career stage, differences were found in how they viewed the typical article published in 2010. Namely, researchers more than 20 years post-PhD viewed the typical 2010 article more positively on a number of characteristics (more accurate, trustworthy, transparently reported, methodologically and statistically rigorous), compared to graduate students and researchers fewer than 8 years post-PhD. Researchers 8 to 20 years post-PhD also viewed the typical 2010 article more positively on a few characteristics (more methodologically and statistically rigorous) compared to earlier career researchers.

Researchers also differed in their perceptions of the amount that the typical article changed over time on six characteristics, all focused on the credibility of the research (i.e., accurate, scientific, trustworthy, transparently reported, and statistically and methodologically rigorous; see Figure 19 and Table 18). These differences emerged between researchers who were more than 20 years post-PhD, who believed less had changed over time, compared to the rest of the sample. Perceptions of the amount of change over time in how boring, creative, groundbreaking, important, and interesting the typical article was did not appear to differ by career stage.

Similar patterns of results were also found in perceptions of the validity of research. Namely, researchers' perceptions of the validity of the typical article differ by career stage for the typical article published in 2010 but not in 2020 (see Figure 20 and Table 19). Overall, researchers more than 20 years post-PhD rated the typical 2010 article higher on all four validities than graduate students and research fewer than 8 years post-PhD. Perceived change in the validity of the typical article over time also differed by career stage. Across the four validities, these findings indicated that researchers more than 20 years post-PhD believed that less had changed over time than all (or nearly all) of the other groups. The largest disagreements were seen in perceived change in the external validity of the typical article over time. For example, researchers fewer than 20 years post-PhD reported roughly two-to-three times as much change in the external validity of the typical article over time compared to researchers more than 20 years post-PhD.

Career Stage and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in the Published Research. Estimates of the replicability, reproducibility, and prevalence of p-hacking in empirical studies published in 2020 were fairly similar across career stages (see Figure 21 and Table 20). However, researchers' perceptions tended to diverge about studies published in 2010. Researchers

more than 20 years post-PhD differed the most from the rest of the sample and estimated the least amount of p-hacking and the highest rates of replicability and reproducibility for studies published in 2010. Comparing 2010 and 2020 estimates, researchers more than 20 years post-PhD generally thought that the least had changed over time. For example, graduate students and researchers who were 20 or fewer years post-PhD estimated around twice as much change in replicability rates between 2010 and 2020 than researchers more than 20 years post-PhD. As seen in Figure 21, estimated rates of fraud were the exception as researchers overall tended to agree that very little had changed over time.

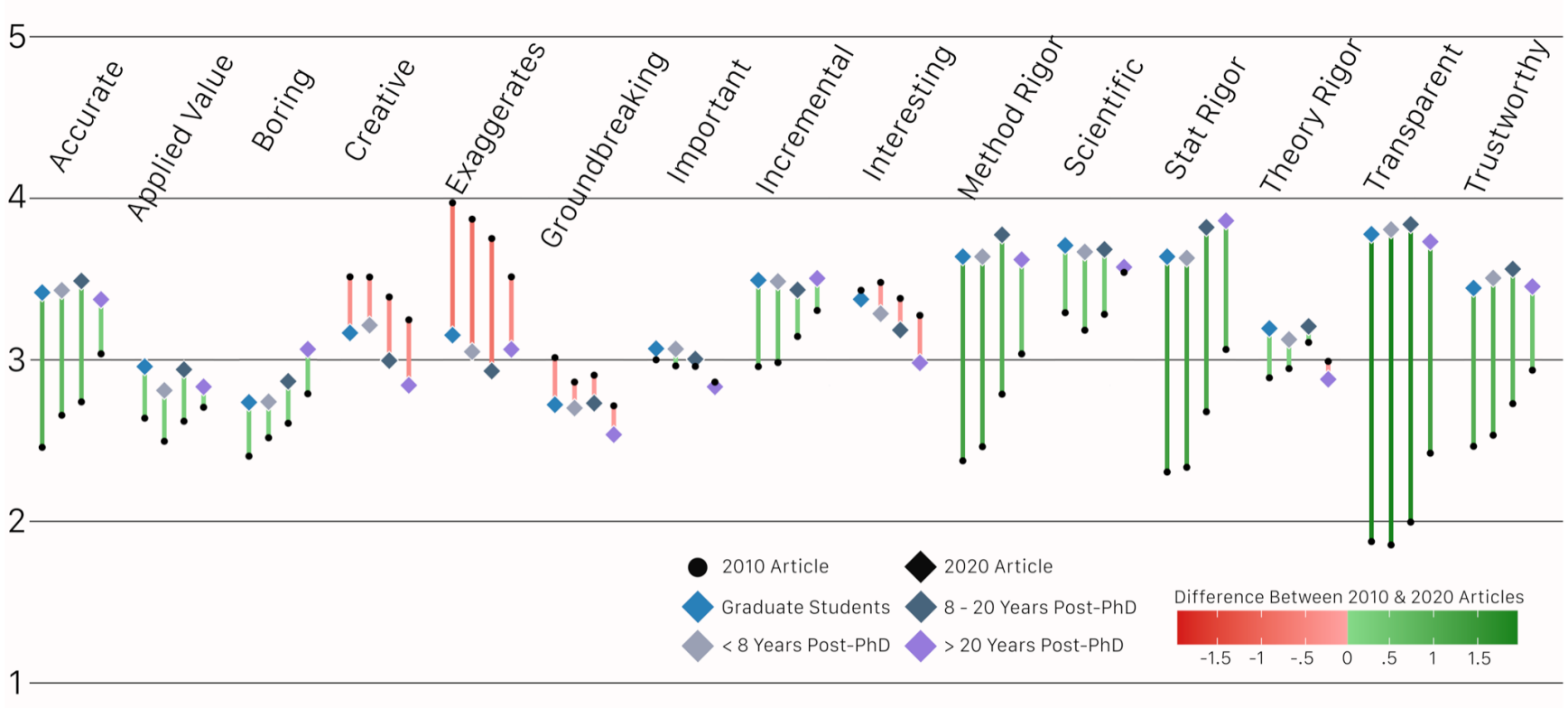
Career Stage and Perceptions of Researchers' Own Work

Overall, researchers more than 20 years post-PhD viewed their own work more positively than researchers who were fewer than 8 years post-PhD and graduate students—rating their own work as more groundbreaking, theoretically rigorous, and creative (see Figure 22 and Table 21). However, graduate students considered their work to be more transparently reported than researchers between 8 to 20 years post-PhD.

Career Stage and How Researchers Perceive Their Own Work Compared to the Typical Published Article. Researchers more than 20 years post-PhD viewed their own work more positively than they viewed the typical 2020 article, more so than researchers at other career stages (though this varied by characteristic of the research and the specific career stage comparison group; Table 21 and Figure 22). Researchers 8 to 20 years post-PhD also viewed their work as more groundbreaking than the typical 2020 article, more so than researchers fewer than 8 years post-PhD.

Figure 19

Mean Ratings of Perceptions of the Typical Article Published in 2010 vs. 2020 by Career Stage



Note. This figure presents researchers' mean ratings of the typical article published in 2010 (i.e., black circle) and 2020 (i.e., diamond) in social and personality psychology separated by career stage, represented by the colour of the diamond. The line connecting the two represents the difference between the two means for each group. The colour of the line indicates the extent to which ratings of the typical article published in 2020 increased relative to 2010 (i.e., green lines) or decreased (i.e., red lines).

Table 18

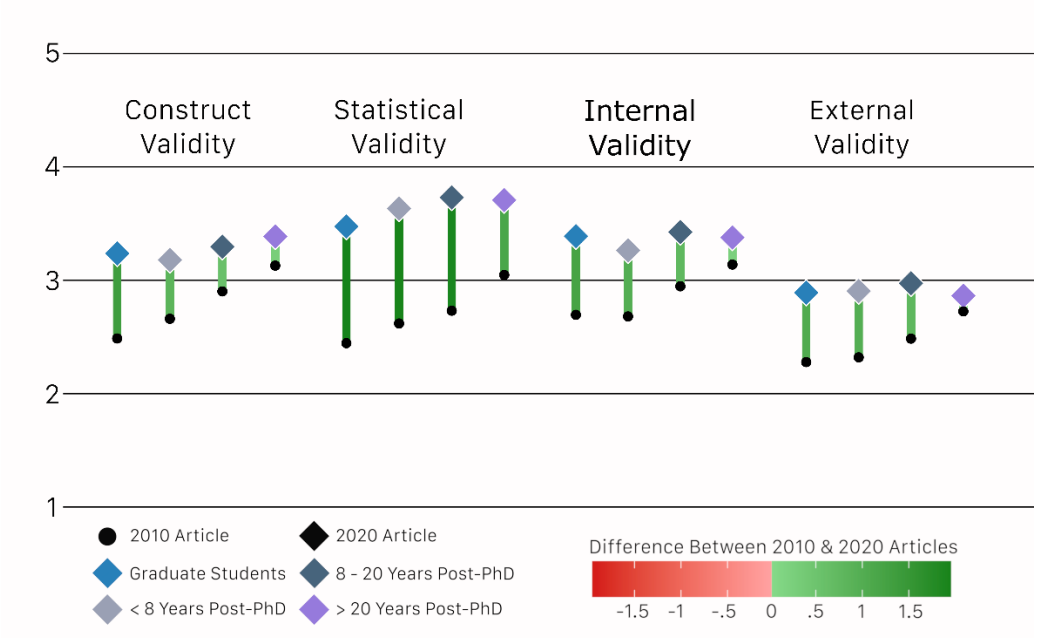
Comparisons of Null Models and Models including Career Stage and Researchers' Perceptions of the Typical Article Published in 2010 and 2020

Career Model	2010 Articles				2020 Articles				2020 Articles – 2010 Articles			
	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BF_M	BF_{10}	Error %
Accurate	0.998	580.54	—	0.007	0.015	—	64.32	0.022	1.00	8275.75	—	0.006
Applied Value	0.068	—	13.75	0.008	0.029	—	33.59	0.015	0.09	—	10.09	0.004
Boring	0.592	—	1.45	0.0001	0.580	1.38	—	0.0004	0.013	—	77.64	0.024
Creative	0.323	—	2.09	0.0003	0.993	148.16	—	0.001	0.018	—	53.45	0.021
Exaggerates Findings	0.926	12.51	—	0.001	0.060	—	15.80	0.008	0.882	7.50	—	0.001
Groundbreaking	0.065	—	14.45	0.004	0.049	—	19.412	0.008	0.016	—	62.67	0.018
Important	0.012	—	82.27	0.023	0.105	—	8.551	0.003	0.047	—	20.14	0.01
Incremental	0.592	1.45	—	0.0003	0.010	—	103.40	0.029	0.933	14.00	—	0.001
Interesting	0.06	—	15.65	0.008	0.835	5.06	—	0.001	0.074	—	12.44	0.003
Methodologically Rigorous	1.00	1.07×10^8	—	0.006	0.059	—	15.94	0.01	1.00	6.80×10^5	—	0.012
Scientific	0.591	1.45	—	0.043	0.015	—	64.04	0.019	1.00	5961.17	—	0.002
Statistically Rigorous	1.00	7.15×10^{11}	—	0.0004	0.564	1.29	—	0.002	1.00	5.54×10^4	—	0.01
Theoretically Rigorous	0.072	—	12.80	0.008	0.496	—	1.018	0.0002	0.798	3.94	—	0.0014
Transparently Reported	1.00	2.62×10^{11}	—	0.014	0.015	—	64.20	0.021	1.00	1.67×10^6	—	0.014
Trustworthy	0.994	178.18	—	0.0008	0.02	—	49.51	0.018	0.999	676.42	—	0.038

Note. This table reports the following information: P(M/D) = posterior probabilities for the career model; BF_M = Bayes factor favoring the career model compared to the null model; BF_{01} = Bayes factor favoring the null model; Error % = Error estimate for the Bayes factors.

Figure 20

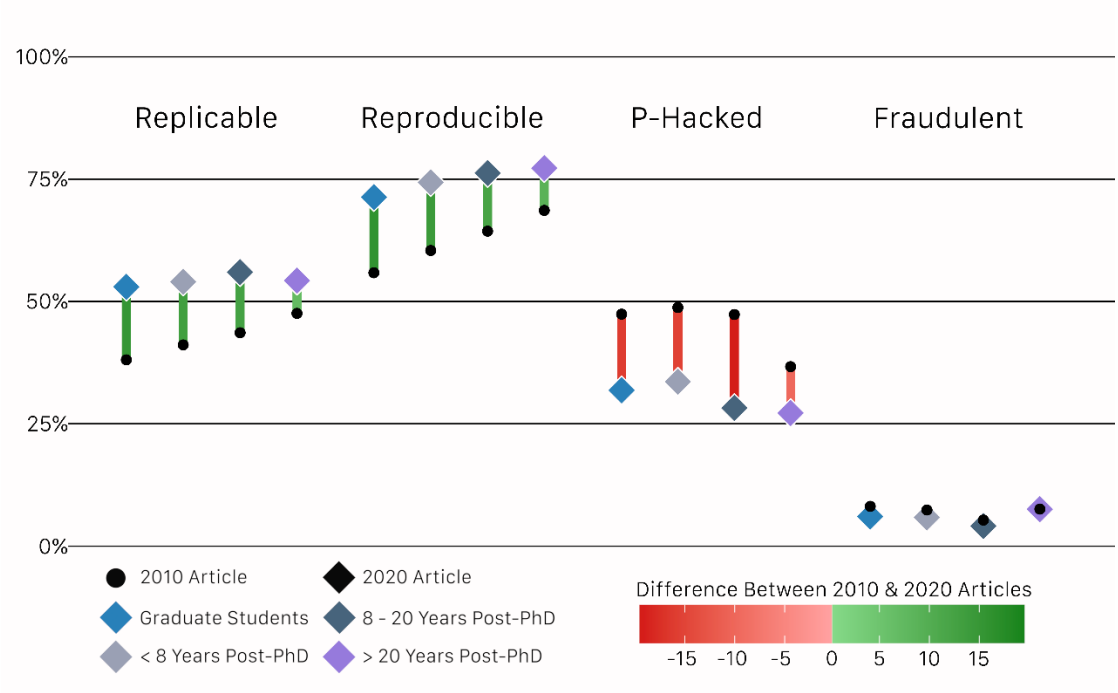
Mean Ratings of Perceptions of the Validity of the Typical Article Published in 2010 vs. 2020 by Career Stage



Note. This figure presents researchers' mean ratings of the typical article published in 2010 (i.e., black circle) and 2020 (i.e., diamond) in social and personality psychology separated by career stage, represented by the colour of the diamond. The line connecting the two represents the difference between the two means for each group. The colour of the line indicates the extent to which ratings of the typical article published in 2020 increased relative to 2010 (i.e., green lines) or decreased (i.e., red lines).

Figure 21

Mean Ratings of the Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 vs. 2020 by Career Stage



Note. This figure presents researchers' mean ratings of the typical article published in 2010 (i.e., black circle) and 2020 (i.e., diamond) in social and personality psychology separated by career stage, represented by the colour of the diamond. The line connecting the two represents the difference between the two means for each group. The colour of the line indicates the extent to which ratings of the typical article published in 2020 increased relative to 2010 (i.e., green lines) or decreased (i.e., red lines).

Table 19

Comparisons of Null Models and Models including Career Stage and Researchers' Perceptions of the Validity of the Typical Article Published in 2010 and 2020

Career Model	Typical 2010 Article				Typical 2020 Article				2020 Articles – 2010 Articles			
	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BF_M	BF_{01}	Error %
Construct Validity	1.00	3514.80	—	0.001	0.047	—	20.44	0.01	0.997	318.38	—	0.017
Statistical Validity	0.999	1167.47	—	0.003	0.160	—	5.24	0.001	0.973	36.63	—	0.001
Internal Validity	0.996	236.66	—	0.001	0.056	—	16.95	0.012	0.995	181.40	—	0.0006
External Validity	0.968	29.96	—	0.001	0.012	—	80.35	0.025	1.00	6951.41	—	0.0006

Note. This table reports the following information: P(M/D) = posterior probabilities for the career model; BF_M = Bayes factor favoring the career model compared to the null model; BF_{01} = Bayes factor favoring the null model; Error % = Error estimate for the Bayes factors.

Table 20

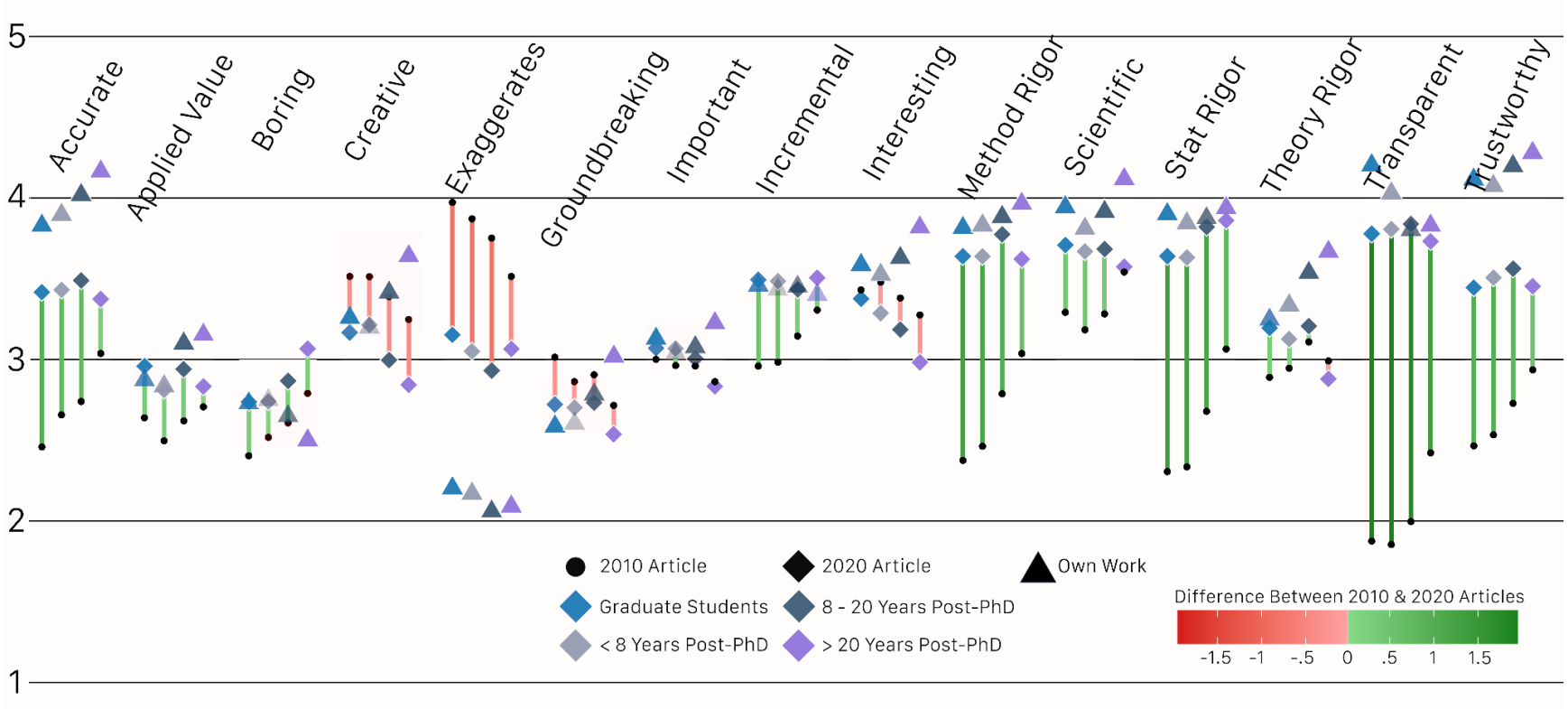
Comparisons of Null Models and Models including Career Stage and Estimated Replicability and Reproducibility Rates and the Prevalence of P-Hacking and Fraud in Empirical Studies in Articles Published in 2010 and 2020

Career Model	Typical 2010 Article				Typical 2020 Article				2020 Articles – 2010 Articles			
	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BFM	BF_{01}	Error %	P(M/D)	BF_M	BF_{01}	Error %
Replicable	0.86	6.04	—	0.0009	0.02	—	50.01	0.019	1.00	12214.91	—	0.007
Reproducible	0.95	17.67	—	0.0008	0.08	—	11.25	0.003	0.97	28.00	—	0.0009
P-hacked results	0.95	21.02	—	0.001	0.48	—	1.09	0.0003	0.99	208.95	—	0.0009
Fraudulent results/data	0.40	—	1.47	0.0004	0.52	1.09	—	0.00007	0.15	—	5.53	0.001

Note. This table reports the following information: P(M/D) = posterior probabilities for the career model; BF_M = Bayes factor favoring the career model compared to the null model; BF_{01} = Bayes factor favoring the null model; Error % = Error estimate for the Bayes factors.

Figure 22

Mean Ratings of Researchers' Perceptions of Their Own Work and the Typical Article Published in 2010 and 2020 by Career Stage



Note. This figure presents researchers' mean ratings of the typical article published in 2010 (i.e., black circle) and 2020 (i.e., diamond) in social and personality psychology separated by career stage, represented by the colour of the diamond. The line connecting the two represents the difference between the two means for each group. The colour of the line indicates the extent

to which ratings of the typical article published in 2020 increased relative to 2010 (i.e., green lines) or decreased (i.e., red lines). The triangle represents researchers' ratings of their own work.

Table 21

Comparisons of Null Models and Models including Career Stage and Perceptions of Researchers' Own Work and Their Work Compared to the Typical Article Published in 2020

Career Model	Researchers' Own Work				Researchers' Own Work – 2020 Article			
	P(M/D)	BF_M	BF_{01}	Error %	P(M/D)	BF_M	BF_{01}	Error %
Accurate	0.856	5.92	—	0.001	0.712	2.47	—	0.001
Applied Value	0.426	—	1.35	0.0004	0.102	—	8.85	0.003
Boring	0.112	—	7.95	0.004	0.993	150.50	—	0.001
Creative	0.997	312.47	—	0.001	1.00	5.84×10^7	—	0.014
Exaggerates Findings	0.030	—	32.58	0.015	0.012	—	83.42	0.024
Groundbreaking	0.984	61.51	—	0.001	0.999	1727.55	—	0.01
Important	0.043	—	22.20	0.01	0.933	13.85	—	0.001
Incremental	0.008	—	128.92	0.051	0.010	—	95.63	0.028
Interesting	0.468	—	1.14	0.00003	1.00	2.69×10^4	—	0.006
Methodologically Rigorous	0.025	—	39.43	0.014	0.051	—	18.79	0.008
Scientific	0.772	3.39	—	0.001	0.951	19.24	—	0.001
Statistically Rigorous	0.012	—	84.14	0.025	0.056	—	16.86	0.009
Theoretically Rigorous	0.966	28.75	—	0.001	1.00	4774.50	—	0.002
Transparently Reported	0.945	17.27	—	0.001	0.754	3.065	—	0.001
Trustworthy	0.232	—	3.31	0.002	0.099	—	9.12	0.004

Note. This table reports the following information: P(M/D) = posterior probabilities for the career model; BF_M = Bayes factor favoring the career model compared to the null model; BF_{01} = Bayes factor favoring the null model; Error % = Error estimate for the Bayes factors.

General Discussion

Overall, the data provided by researchers in the current sample offer a glimpse into how researchers in social and personality psychology profess to perceive their field, how the published literature has changed over time, and how they view their own work. Our findings indicate that researchers in social and personality psychology perceive the quality of the published literature to have improved over the last decade. Some of the largest improvements were believed to be in how transparently reported, statistically and methodologically rigorous, accurate, and exaggerated the typical published article was in 2020 compared to 2010. Researchers also believed that significant strides have been made in decreasing the prevalence of p-hacking and increasing the replicability and reproducibility of published findings.

An important question is how well the current sample represents the perceptions of researchers in the field of social and personality psychology more broadly. The response rate to our survey, overall, was quite low (around 5%). Many of the email addresses were no longer in existence and we received a flood of auto-replies indicating researchers were on parental leave, sabbatical, working remotely, or out of office dealing with covid related issues (among other things). Completing our survey took some time and was described as estimated to take ~25 minutes. This likely deterred some participation, but was a tradeoff made to allow us to include the depth of measures that we did. We suspected that offering to donate to a scientific organization in exchange for participation rather than paying researchers directly might increase interest and willingness to participate. However, whether this ultimately helped or hindered participation is unknown. Researchers who did not particularly identify with or care about the organizations on the list of where donations could be directed may have found this to be a deterrent, and others may have preferred to be directly compensated. The researchers who opted to participate in our survey may not be

representative of researchers in the field at large. Responses rates may have been higher among researchers interested in metascience and those who generally held more positive attitudes towards the authors of the survey (one of whom is widely recognized for their views on the replication crisis and credibility revolution, and for advocating for reform). It is difficult to estimate the extent to which these factors impacted the response rate and representativeness of our sample.

The researchers invited to participate made up a representative sample of authors of articles recently published in several popular journals in the field of social and personality psychology. However, the sample of researchers who actually participated likely do not, and those who selected to participate may differ from the larger population in important ways. We included very few demographic questions to help minimize potential risks of participants' anonymity being compromised. The distribution of participants across various career stages is encouraging, and suggests our sample was not skewed towards early-career researchers. Most participants (53.48%) were eight or more years post-PhD, and senior researchers (> 20 years post-PhD; 16.82%) were more represented in the sample than graduate students (10.91%).

What is a reasonable response rate to expect and how does the response rate of the current sample compare to previous surveys? In a survey of published authors across four fields in the social sciences, Christensen et al., (2022) found the lowest response rates among authors published in psychology journals. Authors in psychology had a response rate of ~26% compared to those in sociology, economics, and political science which ranged from ~38% to ~53%.[3] These response rates are notably higher than in the current study.

[3] Responses rates by discipline were presented in Figure 1 (Christensen et al., 2022). The exact response rates for published authors in each discipline were not reported. The response rates we included are based on examination of the figure and should not be taken as the precise values.

This is probably due to the sizable difference in compensation, with their participants being paid between US$300 to US$400 per hour, and differences in survey length (~15 vs. ~30 minutes). Christensen et al. (2022) acknowledge their response rates likely reflect an upper bound on what is reasonable or possible for researchers to obtain. Indeed, other surveys of researchers in psychology have found much lower rates of responding. A 2015 survey of members of professional societies in social and personality psychology elicited a collected response rate around ~15% (Motyl et al., 2017). Invitations to this survey were sent using email addresses of active members that had been provided by the societies themselves or were sent by the societies on the researchers' behalf. Given some of the societies themselves shared the survey and that the email addresses of active members were likely more up to date than those in published articles (membership information is often updated yearly during the process of renewing society membership), it is not surprising that the current study achieved a lower response rate. A 10-minute survey sent to members of the SPSP mailing list yielded a response rate of 15.1% (Inbar & Lammers, 2012). Very similar response rates to the current sample have been found in surveys using the email addresses of authors reported in published articles in psychology (4.99%; Houtkoop et al., 2018), geo and space sciences (~4.3%; Schmidt et al., 2016), and using a combination of snowball sampling and author email addresses in a general sample of scientists (~9% not excluding incomplete responses (Tenopir et al., 2011).

Overall, given past rates of responding, the length of our survey, and the added stressors that researchers were likely facing in 2021 due to the Covid-19 pandemic, we consider our response rate to be adequate. As the number of surveys targeting researchers and published authors continues to increase, the accessibility of researchers may decrease (e.g., societies restricting the use of mailing lists). Together, this creates an interesting problem. We need researchers to participate to better understand the state of our field, yet

the more surveys requests researchers are inundated with, perhaps the less willing they will be to participate, and the more resources will be required to convince them into participation. For example, Christensen et al. (2022) required at least US$140,670 (not including the pilot study) and paid graduate students between 46.7 and 75% less money to complete the survey than published authors. Few researchers have the resources available to support recruiting a large sample of researchers compensated at a level likely to elicit somewhat high rates of responding.

Another way of exploring potential biases in our sample is by examining the distribution of where researchers chose to direct donations. The amounts donated to each organization can be seen in Figure 23. The most popularly selected organizations were the Center for Open Science and SPSP. Do the organizations that researchers commonly chose to support suggest bias in our sample? Many chose SPSP, which is widely seen as the flagship professional society in social and personality psychology making this an unsurprising choice. The large number of donations directed towards the Center for Open Science and SIPS suggests that many in our sample support open science and the reform efforts both organizations are known to advocate for. It is difficult to say whether this suggests that researchers with more positive attitudes towards open science and related efforts self-selected to participate in our survey at a disproportionate rate. Support for open science among psychologists was found to be quite high in a recent survey that aimed to recruit a representative sample (Christensen et al., 2022). Perhaps then, it is reasonable to think that social and personality psychologists would likely opt to support organizations like the Center for Open Science.

Figure 23

Waffle Chart of the Where Researchers Selected to Direct Donations

Note. Each 👆 = US$100 donated, with donations rounded to the nearest 100th. ARP = Association for Research in Personality, COS = the Center for Open Science, PLOS = PLOS (formerly the Public Library of Science, SESP = the Society of Experimental Social Psychology, SIPS = the Society for the Improvement of Psychological Science, SPSP = the Society for Personality and Social Psychology.

The third most commonly selected organization was SIPS. The relative amount of donations directed towards SPSP vs. SIPS could indicate a potential overrepresentation of researchers who strongly support reform efforts given the differences in the size of these organizations. Whereas SPSP reports typical conference attendance to be between 3500 and 4000 (SPSP, 2022), the last SIPS conference prior to the Covid-19 pandemic included 521 registrants (SIPS, 2022). Given that SPSP is substantially larger than SIPS—and that SIPS members span many subfields of psychology—it is reasonable to suspect that there was

some self-selection that occurred in who participated in our survey. It is possible researchers were aware that one of the authors of the survey (who were all listed in the signature of the email invitations) co-founded SIPS and were thus more likely to select SIPS as a result. However, email invitations were not sent directly from this author, and we would venture that a number of participants paid little attention to this detail. It is also possible researchers selected SIPS to support what is a much younger society compared to the more well-established societies included, and one known to charge attendees very-little-to-nothing for conferences.

How does the potential overrepresentation of researchers with more positive attitudes towards open science impact the generalizability of our results? First, while we believe our invitations were sent to a representative sample of authors, we do not claim that is the sample we achieved. Readers should constrain their conclusions keeping in mind the biases and issues that plague survey research such as this. Researchers interested and motivated to take time out of their busy lives to complete a 25-minute survey are unlikely to represent the typical researcher in the field. They are, however, in many cases the best data that we have. Second, our analysis of the relationships between researchers' perceptions of the field and their self-reported support for the open science movement should not absolve concerns that our sample may be skewed towards supporters. However, these findings can shed some light on how patterns of results differ across differing levels of support. Overall, support for the open science movement tended to distinguish some researchers' perceptions of the published literature in 2010 but was largely unrelated to perceptions of the literature in 2020. Whether these patterns are observed in other samples should be explored in future research.

Overall, our findings should be interpreted keeping in mind that they relied on self-reports of what researchers think about themselves and their field—not how they actually

are or what the field is actually like. Researchers' self-reports about themselves are inevitably going to be biased. They, just like participants in their own research studies, are not immune to social desirability concerns. Moreover, they exist in a field where their careers depend greatly on their research being evaluated positively by others.

We expect that certain measures included in this study were likely to be more affected by biases in self-reporting, including the measure of intellectual humility. It is, for example, reasonable to expect a researcher who displays extraordinarily little intellectual humility to accurately self-report on their lack of humility? Recent findings comparing self and informant reports of intellectual humility have found little agreement between the two (Meagher, 2022). In addition to the general concerns surrounding the validity of self-reported intellectual humility, we are not aware of research examining the self-reports of scientists. Intellectual humility is a trait that is particularly valuable to the process of science and is reflected in the scientific norm of disinterestedness described by Merton (1942). Given that intellectual humility is both a generally desirable trait and one aligned with scientific ideals, researchers may be particularly vulnerable to social desirability bias when responding to these measures.

Do researchers' self-reports on intellectual humility differ from those of other samples? In reviewing 16 studies that also used the *General Intellectual Humility Scale* (Leary et al., 2017), we found that they all reported mean scores lower than that of the current sample (*M* = 4.32; *SD* = 0.51; *Mdn* = 4.33). Of 14 samples recruited online (through Amazon's Mechanical Turk, Qualtrics Panels, and Prolific), ten reported means ranging from 3.83 – 3.98 (*SDs* = 0.55 – 1.04) and only four reported means between 4.0 – 4.1 (*SDs* = 0.56 – 0.7; see Bowes & Tasimi, 2022; Drummond Otten & Fischhoff, 2022; Leary et al., 2017; Stanley et al., 2020; Zedelius et al., 2022). Similar means have been found in community (*M* = 3.78, *SD* = .66; Deffler et al., 2016) and college samples (*M* = 4.01, *SD* =

0.58; Meagher, 2022). Researchers in our sample self-reported, on average, being around a quarter-to-a-half a point higher on intellectual humility than participants across these 16 samples. Around ⅙ of the researchers gave themselves the highest possible ratings on intellectual humility. Are researchers in our sample more intellectually humble than samples from the general public? Again, our data cannot speak to how intellectually humble researchers *actually* are. However, compared to these past findings, our data suggest that researchers seem to at least describe themselves as such. Future research should seek to develop behavioral measures of intellectual humility to examine how intellectually humble (or arrogant) the claims expressed in the published literature commonly are, and how well they correspond to researchers' descriptions of themselves and their own work.

It is unknown to what extent biases in self-reporting such as overreporting on positive qualities and underreporting on negative qualities would extend to researchers' self-reports about their field. Are researchers motivated to present their field in a positive light? Do researchers feel a need to defend the field in response to the criticisms voiced over the past decade, or are they themselves quick to point out problematic trends that they see? We expect that researchers differ in how much they consider view criticisms of the field to be personally threatening and we encourage future research in this area. How strongly researchers identify with or feel they have contributed to a field may be an important factor. This could also potentially help to explain some of the differences in how researchers at various career stages perceive the field.

Early career researchers may feel little-to-no responsibility for the state of the field (and especially the state of the field in 2010) and no need or obligation to protect the field from criticism. On the other hand, senior researchers who are well respected in the field, have trained numerous PhD students, and served in official and unofficial leadership positions (e.g., journal editors, society leaders, experts in their area of research) may feel a

greater sense of investment and responsibility, and thus motivation to defend the state of the field. However, future work examining individual differences in responses within these groups could help us better understand the variables that predict how researchers respond to criticism and times of crises.

It is also worth emphasizing that most of the items and measures included in this survey were developed ad hoc. Many of the characteristics researchers rated were relatively straightforward (e.g., how interesting, methodologically rigorous) and had high face validity. However, several potential measurement-related limitations should be considered. For example, to explore researchers' perception of the published literature, they were asked to "Imagine the typical social and personality psychology article published in 2010 and 2020." This was designed to capture researchers' general perceptions about what is most common in the published literature without overcomplicating or adding burdensome constraints or qualifiers. A downside to this approach is that it does not clearly state what population of articles researchers should consider when deciding what is typical. For example, a researcher considering the journals they most often read and submit to may respond differently than a researcher considering all possible articles published in the field (including obscure and even predatory journals). We considered adding more specificity (e.g., imagine the typical social and personality psychology article published in X journals or the top Y% of journals on Z metric). Ultimately, we decided against such additions as they present their own problems (e.g., which journals and metrics should be used) and would likely result in an overly complicated and difficult to answer questions. Nevertheless, the ambiguity of the final wording should be considered when interpreting our results.

Other measures that should be considered and examined more closely in future research include the items concerning support for the open science movement. We did not measure (or attempt to measure) actual support for open science or self-reported

engagement in open science practices. Rather, we sought to measure researchers' self-reported beliefs about what they perceived to be the ideals and the practices and principles of the open science movement. Self-reported support for open science may have looked differently had we asked researchers how much they support specific ideals or practices associated with open science (e.g., sharing data, preregistering studies). Interpreting researchers' self-reports can be difficult without also knowing what they perceive the ideals and the practices and principles of the open science movement to be. An interesting approach that could help to better contextualize these self-reports would be to collect qualitative data to explore what researchers commonly associate with this movement.

Looking Forward

A compelling question that follows from this research is how well researchers' perceptions actually align with reality? How accurate are their self-reports about the state of their field and how it has changed over time, about their own work, and about themselves? For example, the researchers surveyed believed p-hacking decreased by ~33% over the last decade—and that replicability and reproducibility rates increased by ~28% and 20%, respectively. This would appear to be good news for the field, if it is indeed reflective of reality. However, as we have argued previously (see Schiavone & Vazire, 2022), the field must not be too quick to declare the time of crisis to have passed and that the problems brought to the surface in the 2010s have largely been addressed. Thus, we should not simply take researchers' self-reports as evidence that these concerns have been or are being resolved.

Does it matter if researchers' perceptions are accurate? Answers to this question likely depend on the specific variables being considered, and the potential impact of discrepancies between perceptions and reality. If researchers drastically underestimate how boring their work or the published literature is, this may be of little practical consequence.

However, the implications could be much more dire if researchers were discovered to wildly overestimate the statistical rigor of research in the field (or of their own research). Unrecognized problematic practices or behaviors threatening the statistical rigor of research within a field are unlikely to resolve themselves. Thus, understanding discrepancies between researchers' perceptions and characteristics observed in the literature can identify potential areas requiring the field's attention (if there are discrepancies about issues that researchers consider important to get right) and inform efforts to close these gaps. These could include efforts aimed at simply increasing the accuracy of researchers' perceptions to better match what was observed, and efforts to improve characteristics of the literature to bring them closer to where researchers perceived them to be, or some combination of the two.

To investigate the accuracy of researchers' perceptions, metascientific research is needed to empirically analyze the state of the published literature. Questions about the quality of research are undoubtedly complex. In effort to answer difficult questions about where the field is at, metascientific research combining approaches and sources of data is needed to piece together a clearer picture of the field. We describe potential approaches and propose an agenda for metascientific research in social and personality psychology in Schiavone and Vazire (2022). By taking stock of where we are as a field and inspecting how our beliefs correspond with our behaviors, we can endeavor to identify rose-coloured hues present or developing in our collective lenses. In doing so, perhaps we can better safeguard against, and course correct before finding ourselves once again in the throes of another crisis.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

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

Agnoli, F., Fraser, H., Singleton Thorn, F., & Fidler, F. (2021). Australian and Italian psychologists' view of replication. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211039216. https://doi.org/10.1177/25152459211039218

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLOS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis, J. P. A. (2011). Public availability of published research data in high-impact journals. *PLOS ONE*, *6*(9), e24357. https://doi.org/10.1371/journal.pone.0024357

Anderson, M. S. (2000). Normative orientations of university faculty and doctoral students. *Science and Engineering Ethics*, *6*(4), 443–461. https://doi.org/10.1007/s11948-000-0002-6

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science:

Results from a national survey of U.S. scientists. *Journal of Empirical Research on Human Research Ethics*, *2*(4), 3–14. https://doi.org/10.1525/jer.2007.2.4.3

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design Is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, *41*(5), 319–329. https://doi.org/10.1016/j.evolhumbehav.2020.07.015

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (20180118). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3. https://doi.org/10.1037/amp0000191

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

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. https://doi.org/10.1177/1745691612459060.

Bastian H. (2017, August 29). Bias in open science advocacy: the case of article badges for data sharing. PLOS Blogs Absolutely Maybe. https://perma.cc/ PL7S-R5W3

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

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

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences, 115*(19), 4887-4890. https://doi.org/10.1073/pnas.1719557115.

Bowes, S. M., & Tasimi, A. (2022). Clarifying the relations between intellectual humility and pseudoscience beliefs, conspiratorial ideation, and susceptibility to fake news. *Journal of Research in Personality*, *98*, 104220. https://doi.org/10.1016/j.jrp.2022.104220

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Buttliere, B., & Wicherts, J. M. (2018). *Opinions on the value of direct replication: A survey of 2,000 psychologists* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/z9kx6

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, H. A., Micheli-Campbell, M. A., & Udyawer, V. (2019). Early career researchers embrace data sharing. *Trends in Ecology & Evolution*, *34*(2), 95–98. https://doi.org/10.1016/j.tree.2018.11.010

Chalkia, A., Van Oudenhove, L., & Beckers, T. (2020). Preventing the return of fear in humans using reconsolidation update mechanisms: A verification report of Schiller et al. (2010). *Cortex, 129*, 510-525. https://doi.org/10.1016/j.cortex.2020.03.031.

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

Chang, A. C., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say "usually not." *Finance and Economics Discussion Series*, *2015*(83), 1–26. https://doi.org/10.17016/FEDS.2015.083

Christensen, G., Wang, Z., Paluck, E. L., Swanson, N., Birke, D. J., Miguel, E., & Littman, R. (2022). *Open science practices are on the rise: The State of Social Science (3s) Survey* [Preprint]. MetaArXiv. https://doi.org/10.31222/osf.io/5rksu

Christian, K., Johnstone, C., Larkins, J., Wright, W., & Doran, M. R. (2021). A survey of early-career researchers in Australia. *ELife*, *10*, e60613. https://doi.org/10.7554/eLife.60613

Cobo, E., Cortes, J., Ribera, J. M., Cardellach, F., Selva-O'Callaghan, A., Kostov, B., Garcia,

L., Cirugeda, L., Altman, D. G., Gonzalez, J. A., Sanchez, J. A., Miras, F., Urrutia, A., Fonollosa, V., Rey-Joly, C., & Vilardell, M. (2011). Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: Masked randomised trial. *BMJ*, *343*(nov22 2), d6783–d6783. https://doi.org/10.1136/bmj.d6783

Cobo, E., Selva-O'Callagham, A., Ribera, J.-M., Cardellach, F., Dominguez, R., & Vilardell, M. (2007). Statistical Reviewers Improve Reporting in Biomedical Articles: A Randomized Trial. *PLOS ONE*, *2*(3), e332. https://doi.org/10.1371/journal.pone.0000332

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science*, *334*(6060), 1182–1182. https://doi.org/10.1126/science.1216775

Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Davis, W. E., Giner-Sorolla, R., Lindsay, D. S., Lougheed, J. P., Makel, M. C., Meier, M. E., Sun, J., Vaughn, L. A., & Zelenski, J. M. (2018). Peer-Review Guidelines Promoting Replicability and Transparency in Psychological Science. *Advances in Methods and Practices in Psychological Science*, *1*(4), 556–573. https://doi.org/10.1177/2515245918806489

Deffler, S. A., Leary, M. R., & Hoyle, R. H. (2016). Knowing what you know: Intellectual humility and judgments of recognition memory. *Personality and Individual Differences*, *96*, 255–259. https://doi.org/10.1016/j.paid.2016.03.016

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE*, *7*(1), e29081.

https://doi.org/10.1371/journal.pone.0029081

Drummond Otten, C., & Fischhoff, B. (2022). Calibration of scientific reasoning ability. *Journal of Behavioral Decision Making*, e2306. https://doi.org/10.1002/bdm.2306

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2016). Scientists' reputations are based on getting it right, not being right. *PLOS Biology*, *14*(5), e1002460. https://doi.org/10.1371/journal.pbio.1002460

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition.

Environmental Engineering Science, *34*(1), 51-61.

Epskamp, S. & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute p values. Retrieved from http://CRAN.R-project.org/package=statcheck. (R package version 1.2.2)

The EQUATOR Network. (2018, June 16). *Peer reviewing research*. https://www.equator-network.org/toolkits/peer-reviewing-research/

Falk Delgado, A., Garretson, G., & Falk Delgado, A. (2019). The language of peer review reports on articles published in the BMJ, 2014–2017: An observational study. *Scientometrics*, *120*(3), 1225–1235. https://doi.org/10.1007/s11192-019-03160-6

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Fetterman, A. K., & Sassenberg, K. (2015). The reputational consequences of failed replications and wrongness admission among scientists. *PLOS ONE*, *10*(12), e0143723. https://doi.org/10.1371/journal.pone.0143723

Fidler, F., & Wilcox, J. (2021). Reproducibility of Scientific Results. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests – An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. https://doi.org/10.1016/j.jesp.2017.11.008

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52.

https://doi.org/10.1177/1948550615612150

Fisch, R., & Daniel, H.-D. (1982). Research and publication trends in experimental social psychology: 1971–1980—a thematic analysis of the Journal of Experimental Social Psychology, the European Journal of Social Psychology, and the Zeifschrift für Soziaipsychologie. *European Journal of Social Psychology*, *12*(4), 395–412. https://doi.org/10.1002/ejsp.2420120406

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flore, P. C., Mulder, J., & Wicherts, J. M. (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Fowler, J. H., & Aksnes, D. W. (2007). Does self-citation pay? *Scientometrics*, *72*(3), 427–437. https://doi.org/10.1007/s11192-007-1777-2

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Fried, S. B., Gumpper, D. C., & Allen, J. C. (1973). Ten years of social psychology: Is there a growing commitment to field research? *American Psychologist*, *28*(2), 155–156. https://doi.org/10.1037/h0034202

Fuchs, H. M., Jenny, M., & Fiedler, S. (2012). Psychologists are open to change, yet wary of rules. *Perspectives on Psychological Science*, *7*(6), 639–642. https://doi.org/10.1177/1745691612459521

Gardner, M. J., & Bond, J. (1990). An exploratory study of statistical assessment of papers

published in the British Medical Journal. *The Journal of the American Medical Association, 263*(10): 1355 –1357. https://doi.org/10.1001/jama.1990.03440100061010

Ghiasi, G., Mongeon, P., Sugimoto, C. R., & Larivière, V. (2018). *Gender homophily in citations*. In *Proceedings of the 23rd International Conference on Science and Technology Indicators (p.1520- 1525)*. Leiden, NL: Centre for Science and Technology Studies (CWTS). http://hdl.handle.net/1887/65291

Gross, K., & Bergstrom, C. T. (2019). Contest models highlight inherent inefficiencies of scientific funding competitions. *PLOS Biology, 17*(1): e3000065. https://doi.org/10.1371/journal.pbio.3000065

Han, S., Olonisakin, T. F., Pribis, J. P., Zupetic, J., Yoon, J. H., Holleran, K. M., Jeong, K., Shaikh, N., Rubio, D. M., & Lee, J. S. (2017). A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLOS ONE*, *12*(9), e0183591. https://doi.org/10.1371/journal.pone.0183591

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M. B., Peloquin, N., 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*:201494. http://doi.org/10.1098/rsos.201494

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Henry, M., 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*:180448. https://doi.org/10.1098/rsos.180448

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). https://doi:10.1146/annurev-statistics-031219-041104

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., Van den Akker, O., 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*. Advanced online publication. https://doi.org/10.1177/25152459211040837

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

Harney, J., Mayville, L., Hrynaszkiewicz, I., & Kiermer, V. (2021). *Researchers' goals When assessing credibility and impact in committees and in their own work*. OSF Preprints. https://doi.org/10.31219/osf.io/ryds4

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. https://doi.org/10.7717/peerj.1935

Haven, T. L., Tijdink, J. K., Martinson, B. C., & Bouter, L. M. (2019). Perceptions of research integrity climate differ between academic ranks and disciplinary fields: Results from a survey among academic researchers in Amsterdam. *PLOS ONE*, *14*(1), e0210599. https://doi.org/10.1371/journal.pone.0210599

Heathers, J. A., Anaya, J., van der Zee, T., & Brown, N. J. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative Techniques (SPRITE)*. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26968v1

Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis = trust crisis? The

effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science*, *29*(3), 270–288. https://doi.org/10.1177/0963662520902383

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

Higbee, K. L., & Wells, M. G. (1972). Some research trends in social psychology during the 1960s. *American Psychologist*, *27*(10), 963–966. https://doi.org/10.1037/h0033453

Hirst, A., & Altman, D. G. (2012). Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLOS ONE*, *7*(4), e35621. https://doi.org/10.1371/journal.pone.0035621

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

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

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

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science*, *7*(5), 496–503. https://doi.org/10.1177/1745691612448792

Ioannidis, J. P. A., 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

Ivaniš, A., Hren, D., Marušić, M., & Marušić, A. (2011). Less work, less respect: Authors' perceived importance of research contributions and their declared contributions to research articles. *PLOS ONE*, *6*(6), e20206. https://doi.org/10.1371/journal.pone.0020206

JASP Team. (2022). *JASP*. https://jasp-stats.org/

Jefferson, T., Smith, R., Yee, Y., Drummond, M., Pratt, M., & Gale, R. (1998). Evaluating the BMJ guidelines for economic submissions: Prospective audit of economic submissions to BMJ and The Lancet. *The Journal of the American Medical Association*, *280*(3), 275. https://doi.org/10.1001/jama.280.3.275

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

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*, 206–219. https://doi.org/10.1037/0022-3514.66.1.206

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., ... & Sirota, M. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159-169. https://doi.org/10.1038/s41562-020-01007-2

Kahalon, R., Klein, V., Ksenofontov, I., Ullrich, J., & Wright, S. C. (2021). Mentioning the sample's country in the article's title leads to bias in research evaluation. *Social Psychological and Personality Science*. Advanced online publication. https://doi.org/10.1177/19485506211024036

Kahneman, D. (2022, February 28). *Adversarial collaboration: An EDGE lecture by Daniel*

Kahneman. Edge. https://www.edge.org/adversarial-collaboration-daniel-kahneman

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J. E., Hall, C., Joy-Gaba, J. A., Legg, A. M., Levitan, C., … Ratliff, K. A. (2019). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement*. PsyArXiv. https://doi.org/10.31234/osf.io/vef2c

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: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Koetke, J., Schumann, K., & Porter, T. (2022). Intellectual humility predicts scrutiny of covid-19 misinformation. *Social Psychological and Personality Science*, *13*(1). https://doi.org/10.1177/1948550620988242

Krawczyk, M. (2015). The search for significance: A few peculiarities in the distribution of p values in experimental psychology literature. *PLOS ONE*, *10*(6), e0127872. https://doi.org/10.1371/journal.pone.0127872

Lakens, D. (2015). Comment: What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology*, *68*(4), 829–832. https://doi.org/10.1080/17470218.2014.982664

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), Article 3. https://doi.org/10.1038/s41562-018-0311-x

Lariviere, V., & Sugimoto, C. R. (2019). The journal impact factor: A brief history, critique, and discussion of adverse effects. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 3-24). Springer, Cham. https://doi.org/10.1007/978-3-030-02511-3_1

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10.1177/0146167217697695

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lewicki, P. (1982). Social psychology as viewed by its practitioners: Survey of SESP members' opinions. *Personality and Social Psychology Bulletin*, *8*, 409–416. https://doi.org/10.1177/0146167282083004

Lipsey, M. W. (1974). Research and relevance: A survey of graduate students and faculty in

psychology. *American Psychologist*, *29*(7), 541–553. https://doi.org/10.1037/h0036907

Makowski, D., Ben-Shachar, M., & Lüdecke, D (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541

Makowski, D., Ben-Shachar, M., Patil, I., & Lüdecke, D. (2020). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software, 5*(51), 2306. https://doi.org/10.21105/joss.02306

Martinson, B. C., Anderson, M. S., Crain, A. L., & De Vries, R. (2006). Scientists' perceptions of organizational justice and self-reported misbehaviors. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 51–66. https://doi.org/10.1525/jer.2006.1.1.51

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Meagher, B. R. (2022). An assessment of self and informant data for measuring intellectual humility. *Personality and Individual Differences*, *184*, 111218. https://doi.org/10.1016/j.paid.2021.111218

Mede, N. G., Schäfer, M. S., Ziegler, R., & Weißkopf, M. (2020). The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir)reproducibility of scientific research. *Public Understanding of Science*, 12.

Merton, R. K. (1942). A note on science and democracy. *Journal of Legal and Political Sociology*, *1*, 115.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2022). *BayesFactor: Computation of Bayes Factors for Common Designs*. (Version 0.9.12-

4.3) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

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*(1), 34–58. https://doi.org/10.1037/pspa0000084

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

Nederhof, A. J., & Zwier, A. G. (1983). The 'crisis' in social psychology, an empirical approach. *European Journal of Social Psychology*, *13*(3), 255–280. https://doi.org/10.1002/ejsp.2420130305

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

Nicholas, D., Jamali, H. R., Watkinson, A., Herman, E., Tenopir, C., Volentine, R., Allard, S., & Levine, K. (2015). Do younger researchers assess trustworthiness differently when deciding what to read and cite and where to publish? *International Journal of Knowledge Content Development & Technology*, *5*(2), 45–63. https://doi.org/10.5865/IJKCT.2015.5.2.045

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–

1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). *Replicability, robustness, and reproducibility in psychological science* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ksfvq

The NPQIP Collaborative Group (2019). Did a change in Nature journals' editorial policy for life sciences research improve reporting? *The British Medical Journal: Open Science*, *3*(1), e000035. https://doi.org/10.1136/bmjos-2017-000035

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

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

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

O'Donnell, M., Dev, A. S., Antonoplis, S., Baum, S. M., Benedetti, A. H., Brown, N. D., Carrillo, B., Choi, A. L., Connor, P., Donnelly, K., Ellwood-Lowe, M. E., Foushee, R., Jansen, R., Jarvis, S. N., Lundell-Creagh, R., Ocampo, J. M., Okafor, G. N., Azad, Z. R., Rosenblum, M., … Nelson, L. D. (2021). Empirical audit and review and an assessment of evidentiary value in research on the psychological consequences of

scarcity. *Proceedings of the National Academy of Sciences*, *118*(44), e2103313118. https://doi.org/10.1073/pnas.2103313118

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

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius*, *2*, 2378023115625071. https://doi.org/10.1177/2378023115625071

Peterson, D., & Panofsky, A. (2020). *Metascience as a scientific social movement*. SocArXiv. https://doi.org/10.31235/osf.io/4dsqa

Pollet, T. V., & Saxton, T. K. (2019). How diverse are the samples used in the journals 'Evolution & Human Behavior' and 'Evolutionary Psychology'? *Evolutionary Psychological Science*, *5*(3), 357–368. https://doi.org/10.1007/s40806-019-00192-2

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 524–536. https://doi.org/10.1038/s44159-022-00081-9

Porter, T., & Schumann, K. (2018). Intellectual humility and openness to the opposing view. *Self and Identity*, *17*(2), 139–162. https://doi.org/10.1080/15298868.2017.1361861

Quiñones-Vidal, E., Loźpez-García, J. J., Peñarañda-Ortega, M., & Tortosa-Gil, F. (2004). The nature of social and personality psychology as reflected in JPSP, 1965-2000. *Journal of Personality and Social Psychology*, *86*(3), 435–452. https://doi.org/10.1037/0022-3514.86.3.435

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. (Version

4.2.0) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Red Team Market. (2020). *How it works*. https://redteammarket.com/how-it-works

Reis, H. T., & Stiller, J. (1992). Publication trends in JPSP: A three-decade review. *Personality and Social Psychology Bulletin*, *18*(4), 465–472. https://doi.org/10.1177/0146167292184011

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research*. (Version 2.2.5) [Computer software]. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych

Riddle, T. (2017). *Linguistic overfitting in empirical psychology*. PsyArXiv. https://psyarxiv.com/qasde/

Ring, K. (1967). Experimental social psychology: Some sober questions about some frivolous values. *Journal of Experimental Social Psychology*, *3*(2), 113–123. https://doi.org/10.1016/0022-1031(67)90016-9

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rowhani-Farid, A., Adrian, 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. http://dx.doi.org/10.1098/rsos.191818

Rudis, B. (2020). *hrbrthemes: Additional themes, theme components and utilities for 'ggplot2'*. (Version 0.8.0) [Computer software]. https://CRAN.R-

project.org/package=hrbrthemes.

Rudis, B., & Gandy, D. (2017). *waffle: Create waffle chart visualizations in R*. (Version 0.7.0) [Computer software]. https://CRAN.R-project.org/package=waffle.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

Schafmeister, F. (2021). The effect of replications on citation patterns: Evidence From a large-scale reproducibility project. *Psychological Science*, *32*(10), 1537–1548. https://doi.org/10.1177/09567976211005767

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

Schiavone, S. R., & Vazire, S. (2022). Reckoning with our crisis: An agenda for the field of social and personality psychology. *Perspectives on Psychological Science*, 17456916221101060. https://doi.org/10.1177/17456916221101060

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open data in global environmental research: The Belmont Forum's open data survey. *PLOS ONE*, *11*(1), e0146695. https://doi.org/10.1371/journal.pone.0146695

Shadish, W. R. (1989). The perception and evaluation of quality in science. In *Psychology of science: Contributions to metascience* (pp. 383–426). Cambridge University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia,

F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLOS ONE*, *8*(4), e56515. https://doi.org/10.1371/journal.pone.0056515

Sherman, R. C., Buddie, A. M., Dragan, K. L., End, C. M., & Finney, L. J. (1999). Twenty years of PSPB: Trends in content, design, and analysis. *Personality and Social Psychology Bulletin*, *25*(2), 177–187. https://doi.org/10.1177/0146167299025002004

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 Word Solution* (SSRN Scholarly Paper No. 2160588). https://doi.org/10.2139/ssrn.2160588

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science*, *13*(2), 255–259. https://doi.org/10.1177/1745691617698146

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file-drawer.

Journal of Experimental Psychology: General, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Singleton Thorn, F. (2020). The low statistical power of psychological research: Causes, consequences and potential remedies (Doctoral dissertation). http://hdl.handle.net/11343/251890

SIPS. (2022, January). *Report on SIPS Finances*. The Society for the Improvement of Psychological Science. https://improvingpsych.org/wp-content/uploads/2022/03/Report-on-SIPS-Finances-January-2022.pdf

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Small, H. (2004). Why authors think their papers are highly cited. *Scientometrics*, *60*(3), 305–316. https://doi.org/10.1023/B:SCIE.0000034376.55800.18

Small, H., Kushmerick, A., & Benson, D. (2008). Scientists' perceptions of the social and political implications of their research. *Scientometrics*, *74*(2), 207–221. https://doi.org/10.1007/s11192-008-0213-1

Smoke, K. L. (1935). The present status of social psychology in America. *Psychological Review*, *42*(6), 537–543. https://doi.org/10.1037/h0058585

Soderberg, C. K., Errington, T. M., & Nosek, B. A. (2020). Credibility of preprints: An interdisciplinary survey of researchers. *Royal Society Open Science*, *7*(10), 201520. https://doi.org/10.1098/rsos.201520

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997. https://doi.org/10.1038/s41562-021-01142-4

SPSP. (2022). *Frequently Asked Questions | Society for Personality and Social Psychology*.

https://spsp.org/events/annual-convention/faq

Stanley, M. L., Sinclair, A. H., & Seli, P. (2020). Intellectual humility and perceptions of political opponents. *Journal of Personality*, *88*(6), 1196–1216. https://doi.org/10.1111/jopy.12566

Sterling, T. D. (1959). Publication decisions and their possible effects on Inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112. https://doi.org/10.2307/2684823

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Stürmer, S., Oeberst, A., Trötschel, R., & Decker, O. (2017). Early-career researchers' perceptions of the prevalence of questionable research practices, potential causes, and open science. *Social Psychology*, *48*(6), 365–371. https://doi.org/10.1027/1864-9335/a000324

Teixeira da Silva, J. A., & Bornemann-Cimenti, H. (2017). Why do some retracted papers continue to be cited? *Scientometrics*, *110*(1), 365–370. https://doi.org/10.1007/s11192-016-2178-9

Tennant, J. P., & Ross-Hellauer, T. (2020). The limitations to our understanding of peer review. *Research Integrity and Peer Review, 5*(6). https://doi.org/10.1186/s41073-020-00092-1

Tenopir, C. (2014). Trust in reading, citing and publishing. *Information Services & Use*, *34*(1–2), 39–48. https://doi.org/10.3233/ISU-140725

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

Thalmayer, A., Toscanelli, C., & Arnett, J. (2021). Neglected 95% revisited. *American Psychologist*, *76*, 116–129. https://doi.org/10.1037/amp0000622

Tijdink, J. K., Schipper, K., Bouter, L. M., Pont, P. M., Jonge, J. de, & Smulders, Y. M. (2016). How do scientists perceive the current publication culture? A qualitative focus group interview study among Dutch biomedical researchers. *BMJ Open*, *6*(2), e008681. https://doi.org/10.1136/bmjopen-2015-008681

Toribio-Flórez, D., Anneser, L., deOliveira-Lopes, F. N., Pallandt, M., Tunn, I., Windel, H., & on behalf of Max Planck PhDnet Open Science Group. (2021). Where do early career researchers stand on open science practices? A survey within the Max Planck Society. *Frontiers in Research Metrics and Analytics*, *5*, 586992. https://doi.org/10.3389/frma.2020.586992

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, *13*(2), 130–149. https://doi.org/10.1037/1082-989X.13.2.130

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1), 3. https://doi.org/10.1525/collabra.13

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of

low power in psychological science. *Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040. https://doi.org/10.1080/17470218.2014.885986

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

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

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, *351*, h6467. https://doi.org/10.1136/bmj.h6467

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. https://doi.org/10.1037/a0022790

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

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L. J., Yantis, C., Wong, K. M., Sun, J., Prims, J. P., Mueller, A. B., Melton, Z. J., & Carsel, T. S. (2018). Why do some

psychology researchers resist adopting proposed reforms to research practices? A description of researchers' rationales. *Advances in Methods and Practices in Psychological Science*, *1*(2), 166–173. https://doi.org/10.1177/2515245918757427

Wei T, Simko V (2021). *corrplot: Visualization of a correlation matrix*. (Version 0.92) [Computer software]. https://github.com/taiyun/corrplot

Wells, J. A., Thrush, C. R., Martinson, B. C., May, T. A., Stickler, M., Callahan, E. C., & Klomparens, K. L. (2014). Survey of organizational research climates in three research intensive, doctoral granting universities. *Journal of Empirical Research on Human Research Ethics*, *9*(5). https://doi.org/doi:10.1177/1556264614552798

West, S. G., Newsom, J. T., & Fenaughty, A. M. (1992). Publication trends in JPSP: Stability and change in topics, methods, and theories across two decades. *Personality and Social Psychology Bulletin*, *18*(4), 473–484. https://doi.org/10.1177/0146167292184012

Whitcomb, D., Battaly, H., Baehr, J., & Howard-Snyder, D. (2017). Intellectual humility: Owning our limitations. *Philosophy and Phenomenological Research*, *94*(3), 509-539. https://doi.org/10.1111/phpr.12228

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of

psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wilke, C. (2021). *ggridges: Ridgeline plots in 'ggplot2'*. (Version 0.5.3) [Computer software]. https://CRAN.R-project.org/package=ggridges

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zedelius, C. M., Gross, M. E., & Schooler, J. W. (2022). Inquisitive but not discerning: Deprivation curiosity is associated with excessive openness to inaccurate information. *Journal of Research in Personality*, *98*, 104227. https://doi.org/10.1016/j.jrp.2022.104227

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493–504. https://doi:10.1037/pspa0000056

Zmigrod, L., Zmigrod, S., Rentfrow, P. J., & Robbins, T. W. (2019). The psychological roots of intellectual humility: The role of intelligence and cognitive flexibility. *Personality and Individual Differences*, *141*, 200–208. https://doi.org/10.1016/j.paid.2019.01.016