The State of Social and Personality Science:

Rotten to the Core, Not so Bad, Getting Better, or Getting Worse?

Matt Motyl, Alexander P. Demos, Timothy S. Carsel, Brittany E. Hanson, Zachary J. Melton, Allison B. Mueller, JP Prims, Jiaqing Sun, Anthony N. Washburn, Kendal M. Wong, Caitlyn A. Yantis, and Linda J. Skitka

Paper forthcoming at the *Journal of Personality and Social Psychology* and may be cited as:

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J., Sun, J., Washburn, A. N., Wong, K., Yantis, C. A., & Skitka, L. J. (in press). 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*.

Author Note

Matt Motyl, Ph.D. Department of Psychology, University of Illinois at Chicago, Chicago, IL 60607. Contact: matt.motyl@gmail.com. All authors are affiliated with the University of Illinois at Chicago, Chicago, IL.

We thank Mickey Inzlicht for his blog post "Check Yourself Before You Wreck Yourself" for unwittingly inspiring this project. We thank Norbert Schwarz and Tom Pyszczynski for feedback on the survey used in Study 1. We thank Daniel Wisneski and Tomas Stahl for feedback on the project idea, and Mark Brandt and Kelly Hoffman for feedback on an early manuscript. We are grateful to the editor, Harry Reis, for his extensive constructive feedback, and to four self-identified reviewers, Alison Ledgerwood, R. Chris Fraley, Eli Finkel, and David Funder, who provided especially thoughtful and constructive feedback.

Abstract

The scientific quality of social and personality psychology has been debated at great length in recent years. Despite research on the prevalence of questionable research practices (QRPs) and the replicability of particular findings, the impact of the current discussion on research practices is unknown. The current studies examine whether and how practices have changed, if at all, over the last 10 years. In Study 1, we surveyed 1,166 social and personality psychologists about how the current debate has affected their perceptions of their own and the field's research practices. In Study 2, we coded the research practices and critical test statistics from social and personality psychology articles published in 2003-2004 and 2013-2014. Together, these studies suggest that (1) perceptions of the current state of the field are more pessimistic than optimistic; (2) the discussion has increased researchers' intentions to avoid QRPs and adopt proposed best practices, (3) the estimated replicability of research published in 2003-2004 may not be as bad as many feared, and (4) research published in 2013-2014 shows some improvement over research published in 2003-2004, a result that suggests the field is evolving in a positive direction.

Keywords: scientific quality, replicability, questionable research practices, QRPs, professional standards, methodology, meta-science

The State of Social and Personality Science:
Rotten to the Core, Not so Bad, Getting Better, or Getting Worse?

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way – in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil.

- Charles Dickens, *A Tale of Two Cities*

Science, like the two cities described by Dickens (1859), has faced a tumultuous few years. Numerous papers from many different disciplines argue that most published research findings are false (e.g., Ioannidis, 2005; for a recent review, see Begley & Ioannidis, 2015; Lehrer, 2010; Pashler & Harris, 2012). Following the publication of some particularly incredible and unbelievable findings (e.g., Bem, 2011; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkelman, & Pashler, 2009) and the discovery of outright fraud (e.g., Stapel, as summarized in Enserink, 2012), social and personality psychologists turned inward and began debating the truth value of the research published in our journals. This self-examination has generated dozens of impactful publications that have questioned the acceptability of once normative research practices and have replicated (or attempted to replicate) past findings. Although the content of this discussion is not new (e.g., Cohen, 1962; Greenwald, 1976; Hedges, 1984; Lane & Dunlap, 1978; Meehl, 1990), the most recent instantiation of it has garnered broader participation and catalyzed institutional changes at some of the field's top journals (Eich, 2014; Vazire, 2016).

Moreover, Twitter and Facebook discussions, media attention, and conference presentations during these years made these issues increasingly impossible to miss. In many ways, this discussion could be tantamount to a revolution, with increasing numbers striving toward a new "scientific utopia" (Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012; Spellman, 2015).

We know little, however, of the degree to which these ideas have permeated to and been accepted by those not at the front lines of the debate about both questionable and best research practices. To what extent is there consensus, for example, that research practices in the field are and/or were seriously flawed, and require major changes going forward? And, is there any empirical evidence that discussions about questionable or best research practices lead to changes in researchers' behavior? In other words, are social/personality psychologists evolving overtime into better scientists, maintaining the status quo, or perhaps even becoming worse?

The current paper aims to answer these questions, by examining social/personality psychologists' perceptions of the field and the acceptability/unacceptability of a range of proposed questionable and best practices. Additionally, this paper provides an initial inspection of whether there is evidence that scientific quality in social and personality psychology has changed in the midst of the current discussion on scientific practice. To do so, we conducted two studies. In the first study, we asked social and personality psychologists about how their research practices have changed over time and to estimate how replicable research in social and personality psychology is today compared to the past. The second study supplements these self-reports; we randomly sampled articles published in four well-respected journals in social and personality psychology from years before and after the current scientific quality discussion became mainstream. After selecting these articles, we manually coded methodological and statistical information from the sampled articles to calculate popular metrics designed to assess

research integrity and/or quality (e.g., *P*-curve, replicability index), allowing us to compare the prevalence of trace evidence of the use of questionable research practices (or QRPs), and potential replicability of studies published recently compared to those published 10 years ago as assessed by these metrics. With these data, we examined (a) the degree to which QRPs may in fact be rampant in the field's recent history, and (b) whether the scientific quality discussion is leading to improved scientific practice. Before turning to the particulars of these studies, we first summarize various perspectives that seem to have emerged in response to the status of our science discussion (SSD) in recent years.

Perspectives on the State of Social and Personality Science

Perspectives on the state of social and personality science vary along two main dimensions. First, researchers vary in the extent to which they view the literature as *rotten to the core*, where published findings are mostly false positives. Second, researchers vary in the extent to which they believe that quality of published findings *can get better*. Four main, non-mutually exclusive perspectives emerge and we delineate competing predictions from each perspective below.

Rotten to the Core

> "I'm in a dark place. I feel like the ground is moving from underneath me and I no longer know what is real and what is not." – Michael Inzlicht (2016, "Reckoning with the Past")
>
> "You might have noticed that the persons most likely to protest the importance of direct replications or who seem willing to accept a 36% replication rate as "not a crisis" are all chronologically advanced and eminent. And why wouldn't they want to keep the status quo? They built their careers on the one-off, counter-intuitive,

> amazeballs research model. You can't expect them to abandon it overnight can you? That said if you are young, you might want to look elsewhere for inspiration and guidance. At this juncture, defending the status quo is like arguing to stay on board the Titanic." – Brent Roberts (2015, "The New Rules of Research")

The *rotten to the core* perspective views science in general, and perhaps especially social and personality psychology, as especially troubled, containing many false positives, and facing great barriers to improvement. This perspective view the field as extraordinarily competitive with dwindling grant money available and relatively few jobs for a large number of applicants that creates intense pressure to have beautiful studies and perfect data demonstrating counterintuitive and novel phenomena. If these criteria are not met, then scholars cannot publish, are not competitive applicants for most academic jobs, and struggle to obtain tenure (Nosek et al., 2012). As in other organizational contexts, these competitive and individualist norms may promote cheating and unethical behavior (e.g., Kish-Gephart et al., 2010; Victor & Cullen, 1989). Therefore, it is unsurprising according to the *rotten to the core* perspective that many social and personality psychologists (as well as other scientists) torture their data into submission with the use of QRPs, statistical hacking, and post hoc justification (e.g., Bem, 2003; John et al., 2012; Kerr, 1996). The necessary consequence of these practices is impaired validity and reduced replicability of the purported effects in the published literature (Simmons et al., 2011).

The *rotten to the core* perspective is supported by some replication efforts that report that most findings selected for replication attempts from top psychology journals do not replicate. For example, the Open Science Collaboration (2015) was only able to successfully replicate 39% of 100 published effects. Similarly, Ebersole and colleagues (in press) conducted many simultaneous replications in many labs and found that only 30% of those effects replicated.

Survey research found that investigators admit to using QRPs at an alarmingly high rate (John et al., 2012). A recent meta-analysis concluded that there is very little evidence that ego depletion is a real phenomenon, despite hundreds of studies on the effect (Carter, Kofler, Forster, & McCullough, 2015), something that has led at least some to question whether any findings in the field can be trusted. As one prominent researcher put it: "At this point we have to start over and say, 'This is Year One'" (Inzlicht, as quoted in Engber, 2016).

Although some would argue that the field is essentially *rotten to the core*, it is less clear whether this pessimistic assessment also applies to proposed solutions to the problem. A pessimist could argue that because the academic reward system is so deeply entrenched and longstanding, with so many stakeholders invested in system maintenance, that reform may be nearly impossible. Social and personality psychology exists as only a small force within the larger organizational structures of academic publishing, university level productivity metrics (and associated rewards/punishments), promotion and tenure criteria, and job market pressures. In short, even if social and personality psychology attempts to make changes in research and dissemination practices, broader institutional structures may prove to be so strong that fundamental change is nearly impossible. If this is the dominant mindset in the field, we would expect to see (1) high self-reported rates of engaging in QRPs with mostly cynical justifications for doing so (e.g., that the use of these practices is necessary for academic survival), (2) little impact of the SSD on self-reported intentions to change research and dissemination practices, (3) little change in indices of replicability and other metrics of research quality from 2003-2004 to 2013-2014, and (4) low estimated replicability of research in social and personality psychology.

It Can Get Better

> "The essential causes of the replicability crisis are cultural and institutional, and transcend specific fields of research. The remedies are too." – David Funder ("What if Gilbert is right?," 2016)

> "I think psychology has a lot of potential, and I think we're improving it as a tool to answer really important questions, but I'm not sure we have a lot of answers yet." – Simine Vazire (as quoted in Resnick, 2016)

The *it can get better* perspective perceives that there are many false positives in the published literature, but is more optimistic that the research enterprise can improve and may be getting better over time. According to this view, now that problems with prior practices have been identified, widely discussed, and disseminated in conference presentations, journal articles, blogs, and other forms of social media, researchers and supporting institutions will begin to self-correct as new norms about best practices emerge. There is some basis for this kind of optimism. Research in organizational behavior, for example, finds that promotion of strong ethical cultures that clearly communicate the range of acceptable and unacceptable behavior through leader role-modeling, reward systems, and informal norms can reduce unethical behavior among its members (Kish-Gephart et al., 2010; Treviño, 1990). Social and personality psychology has a number of emerging leaders who are explicitly communicating which research practices are acceptable and which are not, for example, in setting new editorial standards for many of the field's journals (e.g., Giner-Sorolla, 2016; Funder, 2016; Vazire, 2016). Some of the central figures in promoting more open science practices are also being rewarded for their efforts, as Brian Nosek (a leader in open science and in the replication movement) was when he received the Society for Personality and Social Psychology's Distinguished Service to the Field Award in 2014. As more leaders in the field communicate what practices are desirable and scholars are

rewarded for using them, the informal norms will change and the replicability of the research produced should improve. The *it can get better* perspective may be best characterized by Barbara Spellman, former editor of *Perspectives on Psychological Science*, when she stated that "ultimately, after the ugliness is over … the science will end up being better" (as quoted in Resnick, 2016).

If the optimism inherent in the *it can get better* perspective is an accurate characterization of the field, then we would predict (1) relatively low self-reported use of QRPs, and justifications provided for using these practices will be independently coded as out of researchers' individual control (e.g., editors insist on them as a condition for publication), (2) high intentions to reduce these behaviors in light of the SSD, and (3) actual research practices and replicability indices should improve from 2003-2004 to 2013-2014.

It's Not So Bad

> "Science… is a method to quantify doubt about a hypothesis, and to find the contexts in which a phenomenon is likely. Failure to replicate is not a bug; it is a feature. It is what leads us along the path—the wonderful twisty path—of scientific discovery." - Lisa Feldman Barrett ("Psychology is not in crisis," 2015)
>
> "The claim of a replicability crisis is greatly exaggerated." – Wolfgang Stroebe & Fritz Strack, 2014
>
> "The reproducibility of psychological science is quite high." – Daniel Gilbert, Gary King, Stephen Pettigrew, & Timothy Wilson (2016, p. 1037)

In contrast to the *rotten to the core* and the *it gets better* perspectives on the SSD in social and personality psychology is the *it's not so bad* perspective. This view is skeptical about what it means for the field that some large scale replication efforts found that few studies in social and

personality psychology successfully replicated. For example, Feldman Barrett (2015) argued that a "failure" to replicate does not mean that the phenomenon in question is by definition non-existent. Presuming the replication study was well designed and implemented, she argues that a more likely explanation for a failure to replicate is hidden moderators. One very likely hidden moderator that could be operating in social and personality research is that of context. For example, the fundamental attribution error (i.e., when people fail to sufficiently take into account situational constraints on a target's behavior, and they attribute the behavior primarily to characteristics of the target instead) might replicate if the study were conducted in the United States or other Western cultural context, but very well might not replicate if the study were conducted in an Asian or Eastern cultural context. Consistent with this idea, Van Bavel and colleagues (2016) rated how much they thought each of the 100 studies in the Open Science Collaboration's (2015) massive replication effort would be contextually sensitive and found that contextual sensitivity predicted replication failure. In other words, effects that were deemed more contextually sensitive (e.g., "how diversity cues signal threat or safety to African Americans") were less likely to replicate than effects deemed less contextually sensitive (e.g., "extracting statistical regularities in sequences of visual stimuli;" cf. Inbar, 2016). From this point of view, failures to replicate are simply part of the usual progress of scientific discovery, as scientists subsequently seek to understand the conditions under which a given effect will emerge and when it will not.

Others argue that low estimates of replicability and high rates of self-reported use of QRPs in social and personality psychology are due to flawed research methods and/or analyses. For example, the survey method that revealed high levels of self-reported use of QRPs (John et al., 2012) has been critiqued because the questions were often ambiguous and because

participants were not given an opportunity to explain when and why they used a given practice (Fiedler & Schwarz, 2015). According to this view, there may be justifiable reasons to not report a measure (e.g., it did not factor as expected, or had low scale reliability) or a given study (e.g., a manipulation check revealed that the intended manipulation did not create the desired psychological effect). Consistent with this idea, a revised version of the John et al. (2012) survey that asked about more unambiguously questionable practice use revealed significantly lower levels of self-reported QRP use than originally reported (Fiedler & Schwarz, 2015). Additionally, Gilbert and colleagues (2016) argue that the Open Science Collaboration's (2015) massive replication effort contained three statistical errors, which erroneously led to the conclusion that replicability is low. When Gilbert and colleagues re-analyzed the data correcting for potential statistical errors, they concluded that the "data clearly provide no evidence for a 'replication crisis' in psychological science" (p. 1037).

Despite the high profile and large replication efforts that conclude that most findings in psychology journals do not replicate (e.g., Ebersole et al., in press; the Open Science Collaboration, 2015), other similarly large-scale replication efforts have had much higher levels of success. Klein and colleagues (2014) and Schweinsberg and colleagues (2016), for example, successfully replicated more than 86% of the studies they examined. Similarly, Mullinix, Leeper, Druckman, and Freese (2015) successfully replicated 80% of studies they examined and found a correlation of $r = .75$ between the original and replicated effect sizes[1]. Although there are some

[1] Mullinix et al. (2015) replicated experimental studies in political science. Experimental political science, however, is difficult to distinguish from experimental political psychology that is often published in the social psychological literature (e.g., priming effects). Although the correlation between effect sizes is quite large, it does not consider the

important distinctions between the sampling strategies different teams of replicators have used to select studies for replication that likely play a role in these widely variable estimates of replicability, these studies nonetheless point to the conclusion that the state of the field may not be as bad as earlier replication efforts may have suggested (see also Gilbert et al., 2016).

If the *it's not so bad* perspective provides the best account of the current SSD in social and personality psychology, we would expect to observe (1) low levels of self-reported QRP use, or that explanations for "QRPs" will either be rated by independent coders as mostly acceptable or as required by editors/reviewers as condition for acceptance (given current behavior is fine), (2) low self-reported intentions to change research practices as a consequences of the SSD, and (3) reasonably high estimates of replicability of studies and other indices of research quality in published findings not only in 2013-2014 (after the SSD became more widespread), but also in studies published in 2003-2004.

It's Getting Worse

"We have created a career niche for bad experimenters. This is an underappreciated fact about the current push for publishing failed replications. I submit that some experimenters are incompetent. In the past their careers would have stalled and failed. But today, a broadly incompetent experimenter can amass a series of impressive publications simply by failing to replicate other work and thereby publishing a series of papers that will achieve little beyond undermining our field's ability to claim that it has accomplished anything… Crudely put, shifting the dominant conceptual paradigm from Freudian psychoanalytic theory to Big Five research has

---

reliability of the effect sizes between the original and replication studies. A similarly strong, positive correlation was observed in OSC (2015).

reduced the chances of being wrong but palpably increased the fact of being boring." – Roy Baumeister (2016)

"Communicating an unfortunate descriptive norm ('almost everybody violates norms of good scientific practice anyway) undermines a desirable injunctive norm ('scientists must not violate rules of good scientific practice)." – Klaus Fiedler & Norbert Schwarz (2015)

The *it's getting worse* perspective argues that overall, past research outputs were mostly revealing truth (as opposed to false positives) because incompetent researchers were weeded out, and that the current push for improving research practices is making research weaker and less interesting because less competent researchers can focus on replication efforts. This perspective seems less common than the other three, but is an important possibility to consider. Past research suggests that descriptive norms can shape our behavior in positive or negative ways (e.g., Cialdini, Reno, & Kallgren, 1990). Therefore, the discussion of the prevalence of questionable research practices and potential fraud may communicate that these practices are normative, which may lead to an increased use of those practices.[2] Moreover, if the field requires large samples and rewards replication, it may lead to more findings that are less interesting conducted by researchers who are not sufficiently competent with their "intuitive flair" (Baumeister, 2016). This orientation may also discourage researchers from doing creative and exploratory research, and from publishing non-preregistered findings. And, these highly-publicized discussions of failures to replicate and questionable research practices might have unintended consequences of

[2] Indeed, one society that we contacted and asked to disseminate our survey declined to participate out of fear that our survey would give its membership cues that questionable research practices are normative, increasing the likelihood that its members would use those practices.

discouraging funding sources and universities from continuing their support of psychological research.

If the SSD is conveying social norms that QRPs are widespread and that non-replicable findings are publishable, then evidence consistent with the *it's getting worse* perspective would include (1) low levels of past usage of QRPs, (2) increased intentions to use QRPs in the future, and (3) declining estimates of replicability for more recent research compared to research from the past. In contrast, if the SSD is conveying that creative, exploratory research is risky and less publishable, then evidence consistent with the *it's getting worse* perspective might not appear in terms of QRP usage or estimated replicability. Rather, it would appear in decreased creativity and interestingness of research in recent years. This latter form of the *it's getting worse* perspective is beyond the scope of the current studies.

Study 1

The goal of Study 1 was to survey social and personality psychologists about their perceptions of the SSD, their current research practices, the perceived acceptability of various practices, and whether they intended to change their research practices as a consequence of the SSD. More specifically, we focused on the following questions: (a) perceptions of the SSD and whether it has been a good or bad thing for the field, (b) self-reported use of proposed questionable and best research practices, (c) perceptions of the acceptability/unacceptability of using proposed questionable and best practices, (d) open-ended explanations for why proposed questionable practices were sometimes perceived as acceptable, (e) and self-reported intentions to change research practices in light of the SSD.

Method

Sampling

To obtain as broad and representative of a sampling frame of social and personality psychologists as possible, we contacted the mailing lists of the *Society for Personality and Social Psychology* (6,172 members), *European Society for Social Psychology* (1,200 members), and the *Society of Australasian Social Psychologists* (166 members).[3] We requested the e-mail addresses of members or for the society to disseminate our invitation to participate to their members. Shortly after our invitation to participate was distributed via e-mail or through the society's mailing list, a recipient posted the survey link on Twitter. We therefore added a question about where participants learned about the survey (i.e., Twitter, Facebook, e-mail or other) so we could determine the degree to which this Twitter posting (and subsequent posts on Facebook) may have distorted our intended sampling of the largest relevant professional societies. Only 45 participants (< 4%) reported that they found the survey on social media, allowing us to make some rough estimates of response rates to the e-mail invitation. Of the 1,414 people who opened the survey, 1,166 responded to most of the survey questions (about 20% answered all but our demographic questions). Excluding participants who indicated that they found the survey through social media, we estimated that our response rate to the e-mail solicitation was between 15% (assuming 100% overlap of society memberships in our sampling frame) and 18% (assuming 0% overlap of society memberships in our sampling frame).

Participants

Of those who provided individuating background information, most identified primarily as social (79%) or personality psychologists (8%). The remainder of the sample consisted of

[3] We also contacted the Asian Association of Social Psychology, but they declined to disseminate the survey.

psychologists who have a primary specialization in something other than social or personality, most of whom reported being members of one of the societies in our sampling frame and were therefore retained. Participants were 49% male and 47% female (the remainder declined to answer the question or preferred not to identify). Twenty-six percent of our sample were graduate students, 11% non-tenure track Ph.D. holders (e.g., adjuncts, post-docs), 15% assistant professors, 12% associate professors, 16% full professors, and 20% declined to share their stage of career.[4] Fifty-seven percent of participants were affiliated with a public university, 25% with a private university, 1% did not have a university affiliation, and the remainder declined to provide this information.

Measures

Journal specific perceptions of replicability across time. First, we assessed perceptions of the replicability and quality of research in social and personality psychology across time. Specifically, we asked participants to estimate the percentage of results published in the *Journal of Personality and Social Psychology* (JPSP), *Personality and Social Psychology Bulletin* (PSPB), *Journal of Experimental Social Psychology* (JESP), and *Psychological Science* (PS) that would replicate in a direct replication study with 99% power, both 10 years ago and within the last year. Responses were provided on a 10-point scale with the point labels of 0-10%, 11-20% and so on up to 91-100%.

Broader perceptions of the SSD. In addition to journal-specific perceptions of replicability, we asked about the perceived replicability of results in our field more generally,

[4] Responses generally did not vary by career stage (these analyses are presented in our supplemental materials). When responses did vary by career-stage, career-stage explained less than 1% of the variance in the response.

specifically, "Do you think that research in social psychology is more replicable today than it was 10 years ago?" (*yes/no*), and, "How confident are you that the majority of findings in social psychology will replicate?" (*not at all, slightly, moderately,* and *very confident*).

We also included 3 items to assess perceptions of the SSD. More specifically, we asked how positive or negative the discussion has been for the field (i.e., "Do you think the 'status of our science' discussion has been more positive or negative for social psychology?" with the following 7 response options: *entirely negative, with no positives; mostly negative, with very few positives; slightly negative, with some positives; equally negative and positive; slightly positive, with some negatives; mostly positive, with very few negatives; entirely positive, with no negatives*). We also asked whether participants believed the discussion has improved research (i.e., "To what extent has the 'status of our science' discussion improved research in social and personality psychology?") and whether the discussion has changed the way they do research ("To what extent has the 'status of our science' discussion changed the way you do research?"). Both of these items had the following 5 response options of *not at all, slightly, moderately, much,* and *very much*.

Prevalence, acceptability, and intentions to change various practices. We next asked participants a number of questions about QRPs (e.g., not reporting all conditions of an experiment, reporting only studies that "worked," John et al.; for a full list, see Table 2). John and colleagues (2012) assessed prevalence of QRPs by asking participants whether they had personally engaged in specific practices, including falsifying data, not reporting all dependent measures, etc. (*yes/no*). Rather than using this approach, we asked how *frequently* participants engage in each practice, and provided them the opportunity to explain their answers (see Fiedler & Schwarz, 2015, for a critique of the "have you ever," approach, without opportunities for

explanation). Participants reported how frequently they engaged in a given practice on a 5-point scale (*never, rarely, sometimes, often,* and *always*). In addition to examining frequency, we also created a variable to indicate whether participants reported ever engaging in a given practice by recoding the frequency variable as a dichotomy, specifically, those who reported *never* versus those who reported having ever engaged in a practice (i.e., those who reported *rarely, sometimes, often* or *always*). Participants who reported ever engaging in a QRP or not always engaging in an acceptable research practice were also later presented with an open-ended textbox to explain their answer.

John et al. (2012) also asked participants whether various practices were defensible on a 3-point scale with the point labels *no, possibly,* and *yes* and treated this item as a continuous measure. Although it is common to treat certain ordinal measures as continuous, the point-labels usually reflect something about matter of degree (e.g., *not at all, moderately*), rather than categorical yes/no responses. We therefore opted to use a continuous measure of the acceptability of each practice, measured on a 7-point scale (*very unacceptable, moderately unacceptable, slightly unacceptable, uncertain, slightly acceptable, moderately acceptable,* and *very acceptable*). Moreover, participants who responded on the normatively questionable end of the scale (e.g., those who thought it was acceptable to selectively report studies that worked) were asked to elaborate using an open-ended text box with the prompt, "When is [research practice] acceptable?" Finally, to gauge whether researchers' behavior is likely to change as a function of the SSD conversation, we asked whether the likelihood of engaging in a given practice had changed following the SSD (on a 3-point scale with the point labels *decreased, stayed the same,* or *increased*). Because responses to each of our questions about specific practices did not correlate well across our various questions, we analyzed them separately.

Results

The materials and analysis scripts are all available on the Open Science Framework (see https://osf.io/xq3v5/).[5]

Journal Specific Perceptions of Replicability Across Time

One purpose of our survey was to assess scholars' perceptions of the perceived replicability of studies published in the four top tier journals that publish social and personality psychology research, as a function of our sampling periods (about 10 years ago versus the last year or so). Collapsing across all other considerations (e.g., time, journal), the average perceived replicability of studies was $M = 4.95$ ($SD = 1.81$), which translates to just short of 50% of studies. Participants perceived studies published within the last year as more replicable ($M = 5.27$, $SD = 1.89$) than studies published 10 years ago ($M = 4.63$, $SD = 1.90$), $F(1, 1099) = 187.07$, $p < .001$, $d = 0.34$. The perceived replicability of research also varied as a function of journal, $F(3, 3297) = 265.01$, $p < .001$. Tukey's Honestly Significant Difference (HSD) tests revealed that the order from most to least replicable journal was *JPSP*, *PSPB*, *JESP*, and *PS* (see Figure 1), and that the differences in perceptions of replicability over time were more pronounced for *JPSP* and *PS* than they were for *PSPB* and *JESP*.

Broader Perceptions of Replicability

Fifty percent of participants answered "yes" when asked if the field was more replicable now than it was 10 years ago. On average, participants were slightly confident ($M = 2.04$, $SD =$

[5] We hope to make the full data available, but currently our university's Institutional Review Board is prohibiting us from doing so (and are requesting that we not only withhold the data, but also destroy all data in 3-5 years as of this writing). We have filed a formal appeal and will upload the data to the OSF page, if the IRB grants us permission.

0.83, scale range 1-4) that the majority of findings in social psychology would replicate, and participants' perception that the SSD has been a good or bad thing was roughly the *neither good nor bad* response option *(M* = 4.35, *SD* = 1.54, scale range 1-7).

Participants thought that the SSD has moderately improved research in the field (*M* = 2.78, *SD* = 0.96, scale range 1-5), and that the discussion has moderately changed the way that they do research (*M* = 2.83, *SD* = 1.12, scale range 1-5).

Self-reported Use of QRPs and Best Practices

Lifetime use. Table 1 reports the percentage of participants who reported ever using a given practice in their research lifetime, as well as comparison percentages reported by John and colleagues (2012). Two practices—data falsification and stopping data collection early–were reported at similarly low rates in both samples. All other practices were reported at levels higher than observed by John and colleagues. The differences in lifetime prevalence rates suggest that participants may have opted to respond to the *yes/no* version of the question (John et al., 2012) by indicating what they "usually" do rather than something they have "ever done" (see also Fiedler & Schwarz, 2015).

Frequency. The average reported frequency of engaging in questionable practices was quite low in our sample (see Table 1). Participants reported that they *rarely* or *never* falsified data, claimed results were unaffected by demographics when they in fact were or the researcher did not know, stopped data collection early, rounded down *p*-values that were just over .05, excluded data after checking the impact of doing so, failed to report all conditions, decided to collect additional data after looking at results, or reframed unexpected findings as expected *a priori*. Participants reported that on average they *sometimes* failed to report all measures they collected or selectively reported studies that "worked." The average reported frequency of

engaging in proposed new best practices were more variable. Participants reported *never* or *rarely* pre-registering hypotheses, *sometimes* conducting a priori power analyses, and *often* reporting effect sizes (see Figure 2).

Acceptability/Unacceptability. As can be seen in Figure 3, the "questionable" behaviors we asked about were all seen as unacceptable to varying degrees. In order from least to most acceptable were falsifying data, falsely claiming results were unaffected by demographics, not reporting all conditions, stopping data collecting early, excluding some data after looking at its impact, rounding off *p*-values, not reporting all dependent variables, reporting that unexpected results were predicted, selectively reporting studies that worked, reporting effect sizes, deciding to collect additional data after looking at the results, conducting power analyses, making data publicly available, and pre-registering hypotheses.

When is it acceptable to use QRPs and not use proposed best research practices? Participants provided a range of explanations for engaging in research practices that have been called "questionable" (see Table 2 for examples). Each open-ended justification was coded by two members of our team for whether the explanation was one that most researchers would agree was acceptable versus being a clear example of a questionable practice. Specifically, coders were asked, "Do you think that most researchers today would think that this explanation for either using a QRP or not using a best practice is acceptable?" and were told to judge acceptability with the assumption that the behavior had been or would be disclosed in any publication. Response options were: *yes*, *no*, *unsure*, and *uncodeable*. Inter-rater agreement was quite high across behaviors (89 to 100 percent agreement). As can be seen in Table 1, our coders found researchers' justifications acceptable between 81 and 95% of the time for not reporting all dependent variables, collecting additional data after looking, excluding some data after looking

at the impact, rounding *p*-values, and stopping data collection early. Examples of behaviors that were generally coded as acceptable by our coders included dropping conditions or studies when manipulation checks failed, dropping items when they did not factor as expected, increasing *N* using sequential sampling procedures that correct the increase to Type I error, and rounding *p*-values to conform to APA Style, and excluding outliers using statistical conventions.

Other behaviors, however, were much more frequently rated as unacceptable research practices. Fifty-five percent of the justifications for selectively reporting studies, 26% of the justifications for reporting unexpected findings as expected, and 11% of the justifications for not reporting all conditions were judged to be unacceptable. Concerns about publication and/or mentions of direct pressure from reviewers and editors were frequently cited as explanations in each of these cases. More specifically, 83% of participants mentioned publication pressure or editorial/reviewer request as the reason they selectively reported only studies that worked. Thirty-nine percent of those who mentioned dropping conditions and 57% of those who reported unexpected findings as expected similarly mentioned publication pressure or being directed by reviewers and/or editors to do so.

We also discovered that self-reported confessions of data fabrication were almost always false confessions (11 out of 12). In all but one case (an uncodeable and seemingly snarky reference to Bem, 2011), participants in our sample who "admitted" to data falsification either misunderstood the question (e.g., they believed the question referred to Popperian *hypothesis* falsification, not data fabrication), or responded in such a way that made it clear that they mentally reversed the response options (i.e., their open-ended responses made it clear they never thought it was acceptable to falsify data, despite providing an "acceptable" response on the

close-ended measure). In short, self-reported data fabrication is extremely rare and much more often than not were examples of measurement error.

The explanations for not using proposed best practices were much more variable and were not easily coded as acceptable or unacceptable. For this reason, we did not code reasons for not using best practices for acceptability. Justifications for not conducting power analyses included doing exploratory research, not having a basis for estimating the effect size, and planned or actual sample sizes were so large it was deemed unnecessary. People explained not pre-registering their research by arguing that their studies were largely exploratory, it is not required by journals or current standards of ethics (e.g., American Psychological Association; APA), and/or mentioned the extra burden associated with doing so. Participants also explained that publicly sharing data is not currently normative, that they share upon request, that they did not have Institutional Review Board (IRB) approval for doing so, and that they often had concerns about participant confidentiality and/or intellectual property.

Intentions to change. Many participants reported that their intentions to engage in various QRPs has decreased as a function of the SSD (see Figure 4 for more detail). More than 70% of participants indicated that they are now less likely to exclude data after looking at the impact of doing so, not report all dependent measures, not report all conditions, stop data collection early, selectively report only studies that work, falsely claim that results were unaffected by demographics, or falsify data. Participants were least likely to report an effect of the SSD on their decisions to collect additional data after looking, to use conventional rounding rules when reporting *p*-values close to .05, or reporting that unexpected findings were predicted. About half of the sample indicated that the SSD has increased their likelihood of pre-registering

hypotheses, making data publicly available, conducting power analyses, and reporting effect sizes.

Discussion

The results of Study 1 indicate that many social and personality psychologists are deeply pessimistic about whether the field is producing replicable science, even if they think there has been some increase (approximately a 10%) in the likelihood that studies conducted today will replicate relative to studies conducted 10 years ago (which, on average, were perceived as having only a 40% chance of replicating even with 99% power). Moreover, our sample does not seem to be particularly optimistic that the SSD is leading to wholesale improvements of this picture, given that the SSD is perceived to have led to only moderate rather than dramatic changes in research practices in the field.

Even though perceptions of the field as a whole were generally more pessimistic than optimistic (and therefore seemingly most consistent with the *rotten to the core* perspective), participants' reports of their own current and intended future research practices were more consistent with the *it gets better* perspective on the state of our science. At first glance, researchers' self-reported use of QRPs could be interpreted as problematic and hinting at significant rottenness. That said, independent coding of the circumstances in which our sample thought that these practices were acceptable were generally (but not always) encouraging. Our coders rated researchers' explanations for not reporting all measures, collecting additional data after looking, not reporting all conditions, excluding data, rounding of *p*-values, stopping data collection early, not reporting demographic differences, as acceptable on average roughly 90% of the time. Because only small percentages generally reported using a given QRP at all, and the majority of these had acceptable justifications for doing so, there seems reason for optimism that

most researchers' motivations are to do the best science they can with their resources. Researchers' explanations for specific other QRPs, that is, selectively reporting only studies that "work" and presenting unexpected findings as anticipated, indicate that our sample is often explicitly told to tell coherent and tidy stories in their paper submissions, and believe that doing so is necessary to successfully publish. Given the SSD, most researchers probably would not publicly endorse Bem's (2003) advice to tell a compelling and tidy story anymore, regardless of a priori hypotheses and how messy their studies may actually be. That said, our data indicates that researchers are still being told (implicitly, and sometimes explicitly) that Bem's advice on how to write for publication in social and personality psychology still holds. If editorial pressure is a major determinant of researchers' usage of QRPs, then the field should show improvement as editors at top journals become more accepting of messy, but more honest, results.

Despite the need to compete and publish in a world that does not yet uniformly reject all QRPs or reward proposed best practices, our respondents nonetheless reported intentions to change their behavior in ways that reduce use of the former and that increase the use of the latter. We think that these self-reported efforts to improve provide the strongest evidence that most of our respondents—despite their concerns about replicability of the field at large, and their cynicism about what it takes to publish—are trying to do sound science, a conclusion most consistent with the *it's getting better* perspective, modestly consistent with the *it's not so bad* perspective, and inconsistent with the *rotten to the core* and *it's getting worse* perspectives.

In summary, our survey revealed even though our sample seems to think the field overall might be pretty rotten (i.e., non-reproducible), they nonetheless personally report using justifiable research practices, and strong intentions to embrace higher standards of science going forward. Our survey also reveals the limits of how much researchers can do to improve the

science by themselves without greater institutional changes, such as the norms that still dominate publication decisions and practices.

But how much can we really trust these findings? One could argue, for example, that the response rate to Study 1 is too low to allow for inferences and generalization. Although low (roughly between 15-18%), we nonetheless argue that there are reasons to not dismiss our findings out of hand. Response rates to email survey solicitation vary widely based on the sample being targeted, but some studies suggest response rates to e-mail solicitations with no incentives or follow-up reminders to participate in research is generally about 10% (see Couper, 2000 for a review; see also Tourangeau, Conrad, & Couper, 2013). Other surveys administered to social/personality psychologists for studies published in recent years obtained a response rate of between 3.5 and 6.7% (see Inbar & Lammers, 2012). Our estimated response rate of between 15% and 18%, therefore, is considerably better than that average.

One can also take some reassurance from research that has tested side-by-side comparisons of identical procedures with the exception of using extra steps to garner higher response rates (e.g., whether the researchers used call-backs and other attempts to convert non-respondents into respondents). This research found greatly improved response rates at substantial cost, but only trivial differences in the demographic make-up of their samples, and no differences in substantive conclusions (Holbrook, Krosnick, & Pfent, 2007; Keeter, Miller, Kohut, Groves, & Presser, 2000; see also Curtin, Presser, & Singer, 2000, Merkle & Edelman, 2002, who arrived at similar conclusions).

Our ability to test the representativeness of our sample was limited because each society we sampled collects demographic information about its participants in slightly different ways and using different response options. That said, the number of males and graduate students in our

sample were within (respectively) 2 and 6 percentage points of what we expected given the demographics of the 2015 membership of the Society for Personality and Social Psychology, the largest group we sampled (numbers that did not change when we considered only respondents who indicated that they were members of SPSP). Given we did not find many career stage differences in responses the slight under-representation of graduate students should pose little threat to our interpretation of our results or their likely generalizability to the population of social/personality psychologists.

Finally, we understand the limitations of self-reports (e.g., Wilson & Dunn, 2004) and pressures toward socially desirable responding (e.g., Richman, Kiesler, Weisband, & Drasgow, 1999). For this reason, the results of Study 1 should be interpreted as suggestive rather than definitive with respect to whether the field is best characterized or social/personality psychologists perceive it as *rotten to the core*, *getting better*, *getting worse*, or *not so bad* in the first place. To complement our reliance on self-report in Study 1, we therefore turned to a very different method to examine the status of our science in Study 2.

Study 2

The goal of Study 2 was to examine the statistical support for the key hypothesis test in various social and personality psychology journal articles, and to examine likely replicability of these results using a variety of new metrics designed to estimate replicability. Toward this end, we manually coded and compared research published in the past (2003-2004) and more recently published research (2013-2014).

Our decision to manually code the statistics from selected articles departs from what has become popular practice when statistically estimating replicability. Other researchers have taken two general approaches to examine the scientific soundness of the literature. The first approach

is to use computer programs that indiscriminately collect all test statistics from a paper and drop them into a database for subsequent analyses (e.g., Schimmack, 2015). The strength of this approach is that it allows for a quick look at a lot of research over many years. A limitation of this approach, however, is that it includes many statistics that are not critical to the theory being tested in a paper, which could distort the picture of the literature that it paints. For example, if there are many significance tests in a paper and few of them are significant (as is common in research on personality, or on individual differences in attitudes/social cognition), then that research may be seen as less replicable because the proportion of significant findings is so low relative to the number of significance tests conducted and the study's sample size, even if most of those hypothesis tests are irrelevant to the theory being tested. Alternately, papers that include many significance tests that are not relevant to the critical hypothesis (again, as is common in research on personality and individual differences) may appear more replicable than other areas of research because the proportion of significant findings is more normally distributed which might suggest that fewer findings were hacked or hidden in a file drawer. Therefore, findings from this first method must be viewed cautiously and supplemented with more in-depth methods.

The second approach takes the test statistics from a series of studies within a single paper and then examines them closely for departures of what would be expected under usual assumptions of probability distributions (e.g., Simonsohn, Nelson, & Simmons, 2014). This approach is useful in identifying whether the key statistics within a set of studies suggest evidential value, and is an improvement over the first method because it focuses on statistics critical to the theory being tested. The main limitation of this second approach is that it only looks at test statistics and pays little attention to the research practices used to obtain those test statistics (e.g., only looking at *p*-values without considering the complexity of research design

that yielded those *p*-values).[6] Therefore, in Study 2, we manually examined a random sample of articles published in four major journals and coded them for the general statistics they include, the statistics that are critical to the hypotheses being tested, and for the research methods reported in those articles. This approach allows us to examine whether social and personality psychologists' claims of rarely using QRPs and claims of being more likely to use (some) better research practices map onto their actual behavior (or metrics aimed to estimate QRPs), as well as stronger tests of the predictions made by the four perspectives on the status of our science.

Estimating Research Integrity

We used a multipronged approach to estimate the research integrity of the studies we coded. First, we coded for evidence that researchers are using various best practices (e.g., reports of exact rather than rounded *p*-values, evidence of increased transparency by including supplementary materials, and reporting effect sizes). Second, we calculated a variety of indices of "replicability". We chose to use several methods that have received considerable attention in our journals and in blogs popular among social and personality psychologists. These methods include the Test for Insufficient Variance (TIVA), *p*-curve, and *z*-curve. Another class of indices represent different ways of estimating statistical power.

It is important to note that none of the indices actually measure or predict replicability directly. Moreover, there is no agreed upon method for defining a successful replication (see Asendorpf et al., 2013; Open Science Collaboration, 2015). In the case of using statistical

---

[6] In a recent blog post, Simonsohn (2015, "Falsely reassuring: Analyses of all p-values") demonstrated that *p*-curves suggest greater evidential value for *p*-values collected using the automatic approach (approach 1 described above), unless those *p*-values were collected from an analysis that included a covariate.

significance and/or post-hoc estimations of power (derived from statistical significance) and effect size may also not be informative as to future replication success (Hoenig & Heisey, 2001; Sohn, 1998). We will first describe the "replicability indices," and will then turn to indices of power. For lack of a better term, we will continue to call these metrics estimates of replicability.

Estimating Replicability

TIVA. The Test for Insufficient Variance (TIVA; Schimmack, 2014a) is one test designed to estimate how much variability there is around the critical statistics in a set of studies that use null hypothesis significance tests. Theoretically, due to measurement and sampling error, there should be considerable variation of the test statistic across studies. However, due to the importance of having a *p*-value at or below .05, researchers may engage in questionable practices to get their *p*-value below that magical cut-off value. If researchers use these questionable practices, there will be insufficient variance around the test statistic that corresponds to *p*-values around .05 (e.g., *Z* scores around 1.96 have *p*-values close to .05). To generate this statistic, we first used Rosenthal's (1978) method to convert all test statistics into *Z*-scores. Then, we computed the variance of the full set of *Z*-scores and multiplied that by the degrees of freedom (i.e., *N* of *Z*-scores – 1). If the TIVA statistic is small (i.e., less than 1), then the research is more likely to have resulted from QRPs and be less replicable. If the TIVA statistic is large, then the research should be more replicable. We must note, however, that this is an unpublished metric and additional simulation work is required to validate this index.

***P*-curve and *Z*-curve.** *P*-curve is another test that assesses the likelihood of QRPs to obtain *p*-values just below .05 (in other words, *p*-hacking, Simonsohn, Nelson, & Simmons, 2014). *P*-curves are examined in a couple of main ways, i.e., visual and analytical. The visual approach is simply plotting the distribution of *p*-values from near 0 to .05. If the distribution is

skewed such that there are more *p*-values closer to .05 than to .01, it is suggestive that the researchers used QRPs to artificially reduce their *p*-value and increase their likelihood of committing a Type I error. If the distribution is skewed such that there are more *p*-values closer to .01 than .05, it suggests that the findings in that analysis contain evidentiary value. Two analytical approaches to estimating evidentiary value and *p*-hacking have been suggested: One that estimates how many *p*-values reported in a paper are between .04 and .05 (Simonsohn, Nelson & Simmons, 2014, what we call the "original *P*-curve"), and a more recently updated approach that estimates the number of *p*-values in a paper that are between .025 and .05 (Simonsohn, Simmons & Nelson, 2015, or what we call the "ambitious *p*-curve"). We present the *P*-curve analysis using each of these definitions. Evidentiary value of research is considered higher when the number of *p*-values within these ranges are small rather than large.

One limitation of this *P*-curve approach is that it only considers *p*-values less than .05. In response, some (e.g., Schimmack, 2015) have proposed examining the *Z*-curve, which looks at all *Z*-scores from 0 to infinity. This latter approach should reinforce the findings of the *P*-curve, by showing that the distribution of *Z*-scores is skewed in one direction or another. Moreover, the *Z*-curve may clearly demonstrate a publication bias if the distribution of scores is leptokurtic around $Z = 1.96\text{-}2.06$ (the range of Z-scores corresponding to $ps < .04\text{-}.05$).

Estimating Statistical Power

In addition to metrics of the likelihood of questionable research methods, we also examined indices of replicability and more traditional estimates of possible replicability, namely statistical power. A priori power is essentially the log-log linear relationship between sample size and true effect size at a given alpha, provides an estimate of the likelihood that a study will achieve a significant effect (Cohen, 1988; 1992). Post hoc statistical power can be estimated after

a study has been conducted by using the sample size collected, observed effect size, and setting an alpha. The challenge has been to estimate post hoc power as a proxy for a priori power based on the reporting practices in journals.

There is one issue regarding calculating post hoc power after a study has been conducted, namely that any significant effect (*p*-value <= .05) will yield a post hoc power at or above .50 (Hoenig & Heisey, 2001). Post hoc power calculated from a biased literature will produce a strongly negatively skewed distribution with most of the values above .50. Therefore, care should be taken with respect to interpreting post hoc power calculated after a study has been conducted; according to Hoenig and Heisey (2001), "Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature" (p. 1).

Post hoc observed power. Taking all of the above into account, post hoc power can be imperfectly estimated from a study using the reported test value (e.g., *t, F, r,* χ^2) and the degrees of freedom to extrapolate an effect size (Cohen, 1988), or one can ignore the degrees of freedom and convert the reported probability values (or reported test values) into *Z*-scores (see Hoenig & Heisey, 2001). These methods are biased and will inflate estimated post hoc power in small sample studies and do not correct for violations of assumptions, such as heterogeneity of variance, which also inflate post hoc power estimates. It is also important to note that this *Z*-score method for approximating post hoc power will generally provide lower estimates of power than Cohen's *observed power* based on post-hoc effect size approximation methods, especially in multi-factor designs. Importantly, both of these post hoc power estimates are biased to suggest

higher power than actually obtained when the literature being examined contains publication bias in favor of significant effects.

***N*-Pact.** Sample sizes can be used to infer robustness of a particular study, because larger samples usually are better able to accurately detect the medium to small effect sizes that are seen in social experiments (Fraley & Vazire, 2014). Further, larger sample sizes are likely to correspond to higher estimates of power, if we assume most social and personality psychology research studies report small to medium effect sizes. Based on this logic, Fraley and Vazire (2014) proposed the *N*-Pact factor, which is the median sample size of a set of studies being examined and one simple estimation of replicability. The reader should be cautioned, however, that *N*-Pact as a proxy index of *a priori* power has some limitations. Specifically, when effect sizes are heterogeneous, such as in a large collections of studies, the sample size only acts as a guess at the likely replicability, because it is only one half of the power estimation. Moreover, *N*-Pact treats all research designs the same, which would lead to lower power estimates for within-subjects studies relative to between-subjects designs with the same sample and effect sizes. Further, *N*-Pact does not consider other important issues that also affect statistical power (e.g., measurement error, assumption violations). *N*-Pact is therefore a convenient, albeit quick and dirty way to look at likely power, given the number of assumptions it makes about the underlying studies.

R-Index. The Replicability Index (R-Index; Schimmack, 2014b) is an attempt to correct the estimate of power, and subsequent estimated likely replicability, given publication bias in research. The R-Index reduces the "inflated" publication post-hoc estimations of power by adjusting the degree of *incredibility* (i.e., the number of significant effects they have relative to their post-hoc power; Schimmack, 2012). This statistic requires three pieces of information.

First, it requires the median of post-hoc power from a series of studies. Second, it requires determining the percentage of significance tests at $p < .05$. Third, it requires estimating an inflation rate by subtracting the median post-hoc power from the percentage of significant significance tests. Then, the R-Index can be computed by taking the median of estimated post-hoc power and subtracting the inflation rate from it. For example, if you had 5 studies with post-hoc power ranging (.25, .40, .50, .75, .90) and corresponding theoretical dichotomized significant tests with 1 being significant (0, 0, 1, 1, 1; or, ns, ns, $p < .05$, $p < .05$, $p < .05$), your R-Index would be median power (.5) minus the inflation rate (.60 - .50) = .40. Larger R-Index scores should indicate greater likely replicability and smaller R-Index scores should indicate reduced likely replicability. However, caution should be taken in interpreting the R-index, because it simply represents an amalgamated level of "replicability" across a group of studies that are not necessary related. For example, take another 5 studies with powers (.25, .25, .25, 1.00, 1.00) and respective theoretical dichotomized significance (0, 0, 0, 1, 1; or, ns, ns, ns, $p < .05$, $p < .05$). The calculated R-index of .10 in this situation suggests extremely low replicability for the set of studies, but clearly the pattern of dichotomized significance is bound to the power estimate. In sum, this index assumes that studies with power below .5 are basically the result of Type I error, which is not necessarily a safe assumption to make because even low power studies can occasionally find a true effect. Like TIVA, this is an unpublished metric and the validity of the metric has not yet been established; nonetheless, we are including it because of the attention it receives on blogs and discussion boards where social and personality psychologists discuss replicability and research practices (e.g., "Sometimes I'm Wrong," "Psychological Methods," and "PsychMAP").

If the *rotten to the core* perspective is true, we would expect to see poor metrics of replicability and low statistical power across both time periods, with no change from before to after the SSD. If the *it's not so bad* perspective is true, we would expect to see acceptable metrics of replicability and statistical power across both time periods, with little change from before to after the SSD. If the *it gets better* perspective is true, these metrics should indicate that studies should show improvement in replicability metrics and statistical power over time as the SSD changed norms of best practice. If the *it's getting worse* perspective is true, these metrics should indicate that studies should show declines in replicability metrics and statistical power over time.

Method

This project was pre-registered on the Open Science Framework prior to data collection and analysis. All *a priori* hypotheses, coding forms, data, materials, and data management R-scripts are available on the Open Science Framework (https://osf.io/he8mu/).[7]

Article Selection

To enhance generalizability of the findings, we chose to sample articles from four important journals within social and personality psychology—*JPSP*, *PSPB*, *JESP*, and *PS*. We identified that the replicability discussion began gathering more attention around 2005 and was widely discussed at conferences, in peer-reviewed journal articles, and the popular press by 2012 (Ioannidis, 2005; Nosek et al., 2012). Although subjective and somewhat arbitrary, the fact that a special issue of an important journal—*Perspectives on Psychological Science*—was dedicated to discussing the problems with our science and published in 2012 makes this a reasonable year to

[7] Analysis scripts are also available on the OSF page, but were written after the preregistration of the project and while the project was on-going as the first-author learned to use R.

identify as a watershed moment for this discussion when a number of papers started appearing in our journals that embraced improved research practices. Therefore, we chose to include articles published in 2013 and 2014. To examine whether the improving science discussion had an effect on our science, we needed to include articles from a time point before this discussion started rising to the mainstream. To that end, we chose to include articles in those same journals but from 10 years prior—2003 and 2004. After identifying these four journals and four years, we downloaded all 2,228 articles published in them and used a random number generator to assign each of them a number from 1 to 2,228. We then sampled 30% of these articles by selecting the articles numbered 1 through 705. The final sample consisted of 161 articles from *JPSP*, 71 articles from *PS*[8], 147 articles from *JESP*, and 164 articles from *PSPB*. These 543 articles contained 1505 individual studies. A comprehensive list of the selected articles is available on the Open Science Framework (https://osf.io/9mtyi/).

Article Coding

We coded the statistics that the authors reported that pertained to their critical hypothesis test (for a full list of all variables coded, see https://osf.io/9mtyi/). Often, the authors of the studies being coded would identify their critical hypothesis test with verbal markers (e.g., by

[8] PS publishes research from areas beyond social and personality psychology, but the distinctions between areas is blurry and somewhat arbitrary. Therefore, we did two things. First, we analyzed all articles in PS (n = 233). Second, we coded the specialization of the lead author of the paper as social/personality or something else, and only examined papers with lead authors who specialize in social/personality psychology (n = 71). Given the focus of the current manuscript, we exclude papers by non-social/personality authors, but include the analyses including them in the supplemental materials. The pattern of the data does not change whether these articles are included or excluded.

declaring, "as predicted…," or "critical to our hypothesis…") prior to reporting those statistics. If these statements were not present, we read the abstract and hypothesis paragraph preceding the study, and tried to connect the hypothesis present in those with one of the statistical analyses in the results section. If the connection was not obvious, or if there were multiple critical hypothesis tests (e.g., if the authors predicted that all Big 5 personality traits would predict an outcome measure), we coded the first statistical test reported in the results section that was not a manipulation check. In coding the statistics, we recorded the type of test, number of predictor and outcome variables, degrees of freedom, the number of covariates included in the model, the actual test statistic, whether the authors reported an exact *p*-value, the *p*-value they reported, and what effect size they reported. After coding the critical hypothesis testing statistics, we also coded the number of significance tests with *p*-values less than .05, the total number of significance tests conducted that were reported in the article, and the number of footnotes pertaining to analyses. Lastly, we rated the subjective difficulty of coding each study on a 7-point scale (1 = *very difficult* to 7 = *very easy*).

All articles were coded by the authors who either have a doctorate in social psychology or are in a doctoral program in social/personality psychology. To ensure consistency in coding, we reviewed articles together, created a list of what to do with the most complicated articles, and coded articles in a group setting. Upon completion, all raters re-coded 10 studies and we found that raters correctly identified the same critical statistics 80% of the time, suggesting acceptable interrater reliability.

Computation

Using the main statistic reported and degrees of freedom provided for each critical hypothesis test, we were able to estimate effect size and calculate *observed* power (Cohen,

1988). We calculated effect sizes for *t* and *F* values, as well as correlations and chi-squares (See Appendix for the formulas; Cohen, 2008). For multiple regression, because we coded only the critical predictors, we treated those as *t*-tests. Other reported statistics, such as hierarchical linear modeling, mixed models, or other non-parametric statistics were not considered because the calculations of estimated effect-size and power were too complex or impossible to compute given the information coded (and often reported in the manuscripts). All *observed* power calculations assumed between-subject designs because this is the more conservative estimation[9]. In total, 66.78% of the studies reported statistics that have a standard procedure by which they could be converted into effect sizes. Finally, these parametric statistics were converted to a common effect size measure, R^2, for statistical analysis. Because effect size and sample size are the main two components of calculating observed power, we provide analysis of each. *Post-hoc* power can also be computed by converting the test statistic into a *Z*-score and then determining the probability of obtaining that *Z*-score if there is a true effect in the population (see Hoenig & Heisey, 2001). Estimated post-hoc power was used to calculate R-index and to generate *Z*-curves for subsequent analysis.

Statistical analysis

[9] *Observed power* is a post hoc power calculation that is very similar to *post hoc power*, but is slightly improved because it allows consideration of research design, instead of simply converting the observed test statistic into a *z*-score and determining the probability of obtaining that *z*-score if there is a true effect in the population. We use the type of power specified in the formula for calculating each replicability metric. Given the high correlation between the two types of power ($r = .94$, $p < .001$), we only report observed power in our summary tables. See supplement for post hoc power.

Because the assumptions of many parametric tests (e.g., normal distributions, homoscedasticity) are not met for many of the variables we analyze in this study, we turn to modern bootstrapping methods to provide a standardized and unified way to examine all the metrics and indices provided (Efron & Tibshirani, 1994). Bootstrapping is a resampling with replacement method that allows us to avoid making strong distributional assumptions and can be used to provide prediction error for traditional parametric style null hypothesis significance tests (NHST) that produced *p*-values or can be used non-parametrically without *p*-values to provide confidence intervals (CI) to compare between point-estimates and indices. Given the peculiarity of the distributions we encountered, we applied the CI method to make inferences because it is the more conservative approach. Although bootstrapping can also be used to provide an estimate of the point-estimator itself, we also reported the arithmetic values of the point-estimators with bootstrapped confidence intervals to provide a more comprehensive picture of the data.

Confidence intervals. We use the ordinary non-parametric bias-corrected and accelerated (BCa) bootstrap ("Boot" package in R: Davison & Hinkley, 1997), which creates two-tailed 95% non-symmetrical confidence intervals around any point estimator (e.g., mean, median) and any index (e.g., R-index). BCa confidence intervals are computationally intense methods that first calculate the 95% percentile of a resampled distribution of the point estimator and correct the systematic difference between the resampled distribution and the population distribution (bias) and the degree of skewness (acceleration). This method preserves the natural asymmetry in a distribution, which can be understood as the distance between each CI end point and the point-estimator. If the BCa CIs do not overlap, we can infer the two values are statistically different.

Distributions. The distributions of power, estimated effect size, and sample size, were displayed using kernel density estimation (KDE) following the procedures in Venables and Ripley (2002; "GGplot" package in R: Wickham 2009). KDE non-parametrically calculates and displays the probability density of a distribution. To simplify the visual comparison across the two time periods (i.e., 2003-2004 and 2013-2014), we normalized our estimations so that the maximum value for each year period is 1. In other words, KDE distributions can be understood as normalized and smoothed histograms. This method allows us to calculate where the peaks (modes) of distributions are located. We applied this method in conjunction with bootstrapping.

We directly compared the shape of the distributions using an entropy-based method (Maasoumi & Racine, 2002). Entropy is a non-parametric metric of dispersion (like the standard deviation, but it does not compare each value to a mean) that can be applied to discrete or continuous data. Entropy-based methods are commonly used in time-series analyses where distributions have unusual dispersion patterns, such as in EEG data (e.g., Bezerianos, Tong & Thakor, 2003) or human movement data (Stergiou & Decker, 2011). We calculated the entropy for each KDE density and compared between them under the null hypothesis that the densities are equal ("np" package in R: Hayfield & Racine, 2008) to test whether the distribution of values changed from 2003-2004 to 2013-2014. This method relies on a similar type of bootstrap method that we used for BCa CI, but instead calculates the standard error of the bootstrapped samples to give a probability value associated with NHST.

Results

How are research practices changing over time?

Summary statistics and BCa 95% CI are reported in Table 3.

Significance reporting practices. The proportion of studies that reported exact *p*-values

more than doubled between 2003-2004 (19.29%) and 2013-2014 (54.51%). Given that past research has demonstrated that articles commonly include misreported *p*-values that are smaller than they should be considering the test statistic and degrees of freedom (e.g., Bakker & Wicherts, 2011), we also examined the prevalence of misreported *p*-values in the current set of studies. If we simply compare the exact *p*-values that were reported to the *p*-values we computed without making any adjustments for rounding or for accordance with APA style (i.e., no more than 2 decimal places), we found that about 34% of all studies round *p*-values down. However, if we try to account for conforming to APA standards (i.e., 2 decimal places) and look at *p*-values rounded down by more than .004, we see that fewer than 10% of exact *p*-values reported were rounded down excessively. In 2003-2004, 10.53% of studies included *p*-values that were rounded down excessively; in 2013-2014, 5.01% of studies included *p*-values that were rounded down excessively. Although numerically different, this decrease is not statistically different.

Reporting effect sizes. The proportion of studies that reported a measure of standardized effect size more than doubled between 2003-2004 (19.22%) and 2013-2014 (49.65%).

Additional information. The number of analysis-related footnotes has not changed significantly over time, but there was an increase in the proportion of studies that referred to additional analyses available in supplemental materials with 1.36% in 2003-2004 to 8.59% of articles in 2013-2014. This suggests that scholars are increasingly disclosing additional analyses and materials.

Has Replicability Changed Over Time (According to Replicability Metrics)?

Test for insufficient variance. Small Test for Insufficient Variance (TIVA) statistics (i.e., smaller than 1) suggest questionable practices of manipulating *p*-values to be just at .05, which would imply non-replicable results as they were likely manufactured. In this

sample, TIVA yielded statistics of 1610.36 [$CI_{95\%}$ = 1309.05, 1933.91] in 2003-2004 and 2402.14 [$CI_{95\%}$ = 2080.41, 2864.50] in 2013-2014, both of which were statistically significantly greater than 0 (testing against the Chi-square distribution). This result suggests that there was sufficient variance in studies we coded across each time period examined, and that test statistics were not particularly constrained around a singular value. Moreover, the variance around test statistics has increased nearly 34% from 2003-2004 through 2013-2014, and the confidence intervals around the TIVA statistics do not overlap. Thus, although there is increased variance around statistics in more recent years than there was a decade ago, research published a decade ago had sufficient variance and, according to the TIVA criteria, it is less manipulated and therefore more replicable. TIVA suggests that the field is *not so bad* and that it is *getting better*.

***P*-Curve.** According to the binning strategy of detecting "ambitious *p*-hacking," *P*-Curves of research that contains evidentiary value should have *p*-values that are right skewed both for the *p*-values below .025 (half-curve) and also for *p*-values less than .05 (full curve; Simonsohn, Simmons & Nelson, 2015). The original version of the *P*-curve has a slightly more conservative definition of *p*-hacking, where "just" significant hacked results are defined as *p*-values just below .05 but above .04 (Simonsohn, Simmons & Nelson, 2014). *P*-curves have been tested using both binomial tests (i.e., original: more *p*-values below .04 than between .04 to .05, and ambitious: more *p*-values below .025 than between .025 to .05). The ambitious *P*-curve has also alternatively been operationalized by using a continuous method taking Stouffer's mean of the *p*-values for only χ^2, *F*, *t*, and *r* tests.

For the original *P*-curve we will only use the binomial test and for the ambitious *P*-curve we used the most recent published procedure using the scripts from the online application which use both the binomial test and the z-score method (http://www.p-curve.com/app4/ [version

4.04]). Further, we only examined the *p*-values that we were able to calculate from the test statistic reported in a given article. *P*-curves are normally calculated within papers, but there is some discussion as to whether it can be applied across papers when there are heterogeneous effects and we report those results in supplemental information (see van Aert, Wicherts, van Assen, Bakker, Flore, Francis, & Hartgerink, 2016). We show the distribution of the *p*-values in bins by both the mean and median of each paper (Figure 5). When using the *z*-score method for the ambitious *P*-curve, we sometimes only had a single *p*-value to include in the analyses because we only coded one statistic for each study pertaining to the critical hypothesis test in that study. Therefore, in single-study papers, we make evidentiary value judgments based on a single *p*-value (consistent with Simonsohn et al., 2015, and their web application). For the original P-curve, we used a median, and not a mean, to average within paper before tabulating the number of papers that had evidentiary value.

For each paper, we examined each critical hypothesis test reported and asked if the paper had evidentiary value (0 for no and 1 for yes) based on the either the original *P*-curve (median $p < .04$) and for ambitious *P*-curve based on *p*-half and *p*-full curves. We then tabulated the percentage of papers that had evidentiary value in each year and used our bootstrapping method to compare between the two periods.

For the original *P*-curve we found that in 2003-2004, 94.03% [$CI_{95\%}$ = 88.59, 97.09] and in 2013-2014, 95.38% [$CI_{95\%}$ = 91.64, 97.53] of the papers had evidentiary value and there was no difference between the two time periods (as evidenced by the overlapping bootstrapped confidence intervals). For the ambitious *P*-curve we found that in 2003-2004, 50.00% [$CI_{95\%}$ = 39.07, 58.97] and in 2013-2014, 49.64% [$CI_{95\%}$ = 41.48, 57.75] of the papers had evidentiary value and there was no difference between the two time periods. It is important to note that

papers with more studies were significantly more likely to report evidentiary value, r_{pb}= .306, [$CI_{95\%}$ =.153, .441]. This finding may be due to the way the method was implemented for critical hypotheses as opposed to all *p*-values in a paper. Thus, when examining the ambitious *P*-curve, approximately half of papers contain evidentiary value, whereas when using the original *P*-curve, we found that approximately 95% percentage of the papers had evidentiary value. Importantly, caution should be taken in interpreting the ambitious *P*-curve, because it includes a number of single-study papers where evidentiary value judgments are determined based on whether that *p*-value is below .025 or not. The original and ambitious *P*-Curves by paper both suggest no change suggests the field not *getting better* or *getting worse*. Estimates based on the original *P*-curve suggest that the field is *not so bad*, whereas estimates using the ambitious *P*-curve suggest that the field is half *rotten.*

***Z*-curve.** An alternative to the *P*-curve is the *Z*-curve; the *Z*-curve allows examination of potential publication bias, by looking at whether there is a precipitous drop at 1.96 (i.e., the *z*-score corresponding to the *p*-value of .05; Schimmack, 2015). If there is a sharp rise of distribution at around 1.96, it suggests publication bias. If there are many more *z*-scores of 1.96 than 1.95, that would be evidence of *p*-hacking. For computational reasons, however, we limited the *z*-scores to be between 0 and 8.2, which is an equivalent to a *p*-value of 2.2 x 10^{-16}. Further, a *p*-value of 1 x 10^{-13} or a *z*-score of 7.44 is equivalent to a power of 1. We plotted all *z*-scores using a KDE density plot that we were able to create given the statistics reported (*n* = 1061). As depicted in the top panel of Figure 6, the curves for each period are fairly similar and both positively skewed, but with peaks around 2.01. The bootstrapped CIs for the distribution peak in 2003-2004 [$CI_{95\%}$ = 2.03, 2.75] and 2013-2014 [$CI_{95\%}$ = 2.02, 2.47] did not contain 1.96 (*z*-score for *p* = .05). There is clear evidence of publication bias (i.e., a sharp rise of the distribution near

1.96), and less clear evidence for *p*-hacking, as there are many *z*-scores that are greater than 1.96 (or, 2.06, if classifying *p*-values between .04 and .05 as likely the result of *p*-hacking). The observation that the peak of the *z*-curve and the confidence interval around it does not overlap 1.96, suggests that the literature is not replete with "just" significant effects. The peaks in the *Z*-curve distribution did not differ across time periods, which suggests no change in publication bias over time, and using the entropy density equality test, we found that the two density plots for each period did not differ significantly, $S_p = .0132$, $p = .18$. The same conclusion of no difference across time periods can also be seen in the three measures of central tendency (mean, median, peak/mode) found in the bottom panel of Figure 6, that all have overlapping CIs. When examining the *Z*-curves, we see clear evidence of publication bias and low rates of *p*-hacking, and no difference between time periods. The *Z*-curve analysis suggests *it's not so bad* and the field is not *getting better* or *getting worse*.

Sample size and *N*-Pact. Sample size ranged from extremely small ($N = 9$) to extremely large ($N = 1{,}129{,}334$). To generate stable estimates, we transformed *N* into Log_{10} which affects the estimation of the mean, but will leave the median and peak unchanged. We plotted all sample sizes using a KDE density plot ($n = 1483$). As depicted in the top panel of Figure 7, the transformed distribution of *N* was fairly normal, but the distribution for 2013-2014 looked more symmetrical than 2003-2004. Using the entropy density equality test, we found that the two density plots for each time period differed significantly, $S_p = .0163$, $p < .0001$. Both the mean and median sample size (N-Pact) increased from 2003-2004 to 2013-2014, but the peaks remained fairly stable. This increased sample size suggests studies may have higher statistical power in 2013-2014 than in 2003-2004, assuming effect sizes of interest remained constant.

Following Fraley and Vazire (2014), we estimated the percentage of studies conducted that should have power of .80, assuming the average effect size in social/personality psychology of $r = .21$ or $d = .43$ and assuming all designs are between-subjects *t*-tests (Richard, Bond, & Stokes-Zoota, 2003). Under these assumptions, studies with sample sizes at or greater then $n = 172$ would be sufficient sample size to claim evidentiary value at *a priori* power = .80. Under these admittedly conservative assumptions, the percentage of studies with a sufficient sample to obtain a power of .80 significantly increased from 15.20% [$CI_{95\%}$ = 11.70, 19.00] in 2003-2004 to 24.27% [$CI_{95\%}$ = 20.92, 27.18] in 2013-2014.[10] Additionally, these changes are not the same at each of the four journals examined.

If we instead account for differences in design and number of conditions (instead of assuming all mixed and within designs are between-subjects designs), the picture suggests more evidentiary value than when we assume two condition between-subject designs. Moreover, evidentiary value increases from 34.50% [$CI_{95\%}$ = 29.76, 39.82] in 2003-2004 to 46.87% [$CI_{95\%}$ = 43.03, 50.49] in 2013-2014. Thus, this method suggests weak evidentiary value at both time points, but does demonstrate improvement over time, that is, estimated a prior power based on sample size suggests that the field is *rotten to the core* (but less rotten when taking into account design and number of conditions), and is *getting better*.

Power. If the effects of the studies in our sample were all the products of *p*-hacking and publication bias, we would see power at .50 and not above. If power estimations are all above .50 but not centered there, it suggests a strong degree of publication bias. Power estimates near 1.0

[10] This analysis only includes studies that used χ^2, *F*, *t*, and *r* tests.

suggest studies that have extremely small *p*-values (.00001) or some combination of large effect size and degrees of freedom. Cohen's recommendation for (pre-experimental) power is .80, which corresponds to a *p*-value = .005 for post-hoc observed power, which we used as a criterion of adequate power (with caution in mind). Figure 8 shows the KDE distributions of *observed* power by year (see supplementary information for more rudimentary form of post-hoc power used in some replicability metrics). The distribution was highly negatively skewed with peaks at power of 1, suggesting publication bias. Because the values were not centered at .50, however, there is no evidence of extensive and frequent *p*-hacking. The entropy density equality test did not differ significantly across time periods for observed power, S_p = .0173, p = .150. In other words, the post hoc observed estimate of power has remained stable from 2003-2004 to 2013-2014.

Although post hoc observed power estimates are extremely upwardly biased and should be interpreted with great caution, our median values were very near Cohen's .80 threshold for both time periods, a conclusion more consistent with an interpretation of *it's not so bad* than *it's rotten to the core*, because power is not at 50% which would indicate that most results are borderline significant. For post hoc power estimates to be around .80, the observed p-values would need to be .005, which is much lower than the .05 or .025 cut-offs for evidentiary value (Simonsohn et al., 2011, 2015). Therefore, we conclude that the post hoc power estimate is suggestive that there is not rampant p-hacking in the literature, but we cannot rule out how publication bias might contribute to this estimate. The stability of power across time periods

despite the increase in sample size, however, is curious and suggests that effect sizes decreased across time periods.[11]

R-Index. The R-Index is interpreted as the percent likelihood of an effect replicating, where 0 corresponds with 0% likelihood of replicating an effect and 1 corresponds with 100% likelihood of replicating an effect. Therefore, if the field's targeted power level is .80, an R-Index around .80 would indicate acceptable replicability. The R-index decreased numerically, but not statistically over time, from .62 [$CI_{95\%}$ = .54, .68] in 2003-2004 to .52 [$CI_{95\%}$ = .47, .56] in 2013-2014. This metric suggests that the field is not *getting better* and that it may consistently be *rotten to the core*.

As summarized in Table 4, 4 of 7 indices of evidentiary value indicated that the status of the science in 2003-2004 was acceptable and had evidentiary value even before the SSD. In other words, there is some suggestive trace evidence that the field was not wholly *rotten to the core* in 2003-2004. Since the metrics in Table 4 are on different scales, we converted them all to be presented in power units[12] and displayed them in Figure 9. There is also some indication that *it's*

---

[11] To confirm this, we converted available test statistics from our sample into estimated effect sizes and found that the median estimated effect size shrunk significantly from R^2 = .11 [$CI_{95\%}$ = .09, .13] to R^2 = .08 [$CI_{95\%}$ = .07, .08]. The entropy density equality test suggests that the distributions of estimated effect sizes differed significantly across these time periods, S_p = .035, $p < .0001$. With increasing sample sizes and constant power, these smaller effect sizes being observed may be more accurate estimates of the true effect.

[12] Observed power and *R*-index are already in power units. *P*-curve and *a priori* power we simply converted to proportions as we assume .8 value is the target value. *N*-pact value was used in a power estimate assuming the average effect size in social/personality psychology of d = .43 for a between subjects *t*-test. *Z*-curve was converted into a *p*-value and translated into post-hoc power. For TIVA, we first converted it into an effect size and then made

getting better given that we observed significant uptake in various best practices (reporting exact *p*-values, effect sizes, and providing supplemental materials) over the next 10 years, and improvements in some indices of evidentiary value over the same time period (TIVA, *N*-Pact, estimated a priori power).

Discussion

One goal of Study 2 was to determine how well social and personality psychologists' actual research behaviors—and not just self-reported behaviors—fit the four perspectives on the SSD. Of course, we cannot directly infer whether QRPs were used from the published literature, but we can examine whether there is trace or indirect evidence of QRPs using statistical estimates of, for example, sufficient variance, to explore just how endemic possible problems of evidentiary value might be. Although there are still some hints of possible rottenness, the results of Study 2 indicate that the majority of the indices suggest that research published in 2003-2004 may not be the dystopian landscape many had come to fear was the most accurate characterization of the field prior to the SSD.

The results of Study 2 also revealed that reporting practices in the field have changed in the direction of greater transparency over time. However, only three of the metrics of likely replicability (*N*-Pact, estimated a priori power and TIVA) showed improvement from 2003-2004 to 2013-2014, whereas the other methods showed no reliable change (R-Index, *P*-curve, observed post-hoc power). Importantly, none of these metrics indicated that replicability was *getting worse*. The SSD may have led to new normatively accepted best practices, which is

---

the same assumptions as with N-pact. The details can be found in supplemental information which includes the R script for these custom functions.

yielding more transparency but not necessarily more replicable science, yet (at least as inferred from some replicability metrics). Despite the replicability metric-dependent conclusion, transparency itself is a self-evident good, because it allows reviewers and readers to better evaluate the quality of the research, and will assist scholars attempting to do high-fidelity replications of the original research. Taken together, these results are most consistent with the *it's getting better* perspective but do not contradict the *it's not so bad* perspective.

Consistent with the *it's getting better* perspective, Study 2 provides evidence that some research practices have changed over the past decade. In 2003-2004, very few studies reported exact *p*-values, effect sizes, confidence intervals, or included supplemental information. In 2013-2014, about half of studies reported exact *p*-values and effect sizes, and there was an eightfold increase in studies that included supplemental information. Additionally, 6 research practices improved over time, 10 remained the same, and 0 got worse. These changes suggest some improvement in reporting practices. Study 2 therefore provides some evidence that social and personality psychologists have recently implemented at least some of the recommendations to improve replicability (e.g., Asendorpf et al., 2013).

Post hoc observed power levels have not changed since 2003-2004. This seems like a curious finding given that sample sizes have increased overtime. Yet, we also see a relative drop in effect size. There are three possible explanations for the drop in effect size over time. The first explanation is that social and personality psychologists may be examining different phenomena with smaller true effect sizes, requiring larger samples. The second explanation is that social and personality psychologists are conducting studies on more heterogeneous samples and/or are collecting data in noisier environments (e.g., via the internet vs. in lab) that together could increase measurement error or reduce the impact of experimental manipulations. The third

explanation is that effect sizes are often inflated in small samples because the standard deviation is often underestimated (Cohen, 1988). Although sample sizes have increased on the whole, the estimate of effect size has decreased, which could mean we are now seeing better representations of true effect sizes.

One unusual discrepancy in the assessment of evidentiary value comes from the two methods of calculating the *P*-Curve. The change seems to involve a growing fear that researchers are changing their strategy to not only *p*-hack, but to do it "ambitiously." The original *P*-Curve assumed that researchers would see a marginal result and hack at their data to make it into a significant result. This has since evolved into the ambitious *P*-Curve, which instead assumes that researchers are working hard to hack their marginal or even non-significant results into not being "just" significant, right at $p = .05$, but "more significant" with $p < .025$. Moving the bar back of what qualifies as significant, after an alpha has been set, is in complete violation of the spirit of significance testing. Whether or not significance testing is a good or a bad thing is beyond our scope (see Cohen, 1994 for an eloquent discussion on the flaws and misinterpretations of null hypothesis significance testing; see also Meehl, 1997), but it is important to remember that a result is either significant or it is not. A result cannot be "just" or "more" significant, and making these distinctions is clearly leading to a slippery slope (see Lehmann, 1993, for an overview of the history of setting fixed *p*-values for significance testing). At which point do we trust that researchers are not "ambitiously" manipulating their *p*-values: $p < .01$, $p < .001$, $p < .0001$? Rather than assume malicious ambition on the part of researchers, it might make more sense to assume ignorance—something that is better addressed with education.

To date, no studies have directly compared or demonstrated the predictive validity of these indirect metrics of replicability and research quality on the same set of studies. This study

provides the first test of convergent validity of these replicability metrics. There are two dimensions on which we examine convergent validity: (a) whether there is evidentiary value, and (b) whether evidentiary value is increasing. First, the Original P-Curve, Z-Curve, and TIVA converge with each other in suggesting evidentiary value. In contrast, the N-Pact, R-Index, and Ambitious P-Curve converge with each other in suggesting lower evidentiary value. Second, N-Pact, a prior power, and TIVA converge with each other in suggesting increased evidentiary value over time. In contrast, Ambitious P-Curve, R-Index, Z-Curve, and observed post-hoc power, converge with each other in suggesting no change over time. Future simulation studies are needed to further examine the convergent and discriminant validity of these measures. Additionally, future research should apply these metrics to predict replicability of studies included in Many Labs and other replication efforts in the future.

Overall, Study 2 provided some evidence that published research in social and personality psychology may not be as *rotten to the core* as many feared, or *getting worse* as some fear, and that it seems to be *getting better*. When we do see evidence of change, it is changing in the direction of recommended research practices described in the literature on how to improve reporting our science (e.g., Asendorpf et al., 2013). The story becomes more complicated when examining the metrics of replicability and scientific quality, but in general, these metrics suggest that the replicability of social and personality science as a whole may not be as bad as some fear and there are some hints that replicability may be improving (e.g., Ioannidis, 2005; Study 1 of this paper).

Study 2 avoided the self-presentational concerns of the self-report data from Study 1, but is not without limitations. Specifically, we could only observe practices reported in final published articles and not the practices reported in prior drafts of articles, or the methods used to

conduct the research. Therefore, the observed changes may be due to changes in reporting over time rather than changes in the actual conduct of the science being reported. Additionally, it is impossible to say with certainty that it is necessarily the SSD causing the observed changes in research practices and quality. The SSD began and became more widespread after 2003-2004 (e.g., Bem, 2011; Enserink, 2012; Ioannidis, 2005; Lehrer, 2010; Nosek et al., 2012; Simmons, Nelson, & Simonsohn, 2011; Vul et al., 2009; Wicherts, Borsboom, Kats, & Molenaar, 2011) and we observed changes in research practices and replicability as estimated by new metrics in 2013-2014. Indeed, the temporal precedence of the SSD is a necessary, but not sufficient, condition to establish causality. The discussion may be causing these changes, but it may also be that the arc of science bends toward better practice and greater replicability over time.

Study 2 also assumes that researchers were embracing best research practices by sometime earlier than 2013, because there is a publication lag that could lead to some research that was conducted in a different scientific climate that tolerated questionable research practices appearing at some point after the SSD. To us, it seems that the SSD was widespread by 2009 and peer-reviewed research adhering to the best practices being advocated in this discussion were appearing in print by 2011 (e.g., LeBel & Paunonen, 2011; LeBel & Peters, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). Yet, even if the uptake of this discussion occurred later, it would make finding evidence of change less likely. The fact that we observed change suggests that the field is evolving and the SSD may be contributing to that evolution—and that we might observe even stronger evidence of improvement in the years to come with greater time for researchers to further adapt to new norms in the field (e.g., larger samples).

While Study 2 provides a snapshot of social/personality psychological research, it does not address potential differences in practices used by psychologists within the separate subfields

of social and personality psychology. Given that social psychologists often run smaller experiments looking at how contextual factors moderate effects of interest and personality psychologists often run larger correlational studies looking at the structure of personality and how it relates to behavior, there may be differences sample size, statistical power, the prevalence of questionable research practices, and replicability. Indeed, some analyses demonstrate that personality research tends to rely on larger samples (e.g., Fraley & Vazire, 2014). Yet, it remains unclear whether these larger samples are needed due to analysis requirements and do not actually reflect greater statistical power when accounting for analytical and methodological details (e.g., larger samples are needed when doing structural equation modeling than t-tests, ceteris paribus). The current work focused on one general psychology journal (*PS*), two general social/personality journals (*JPSP*, *PSPB*), one experimental social journal (*JESP*), and no purely personality journals (e.g., *Journal of Personality*, *Journal of Research in Personality*). Therefore, Study 2 may apply slightly more to social psychology than to personality psychology, but that it is unclear whether there are major differences in estimated replicability or the prevalence of questionable research practices between these two related subfields.

Future research should expand on the current method to see how practices are changing and whether practices vary by subfield. Given technological advances, it is increasingly possible to study the research process. For example, the Open Science Framework allows researchers to upload revised versions of their hypotheses, analysis scripts, and conclusions. As more social and personality psychologists adopt the Open Science Framework, it may be possible to examine whether the original hypotheses are the ones that appear in the eventual publication. If not, that would be clear evidence of hypothesizing after the results are known (Kerr, 1998). This approach

would be especially informative because it would allow for an in-depth audit of the actual research process rather than its final output.

General Discussion

Together, Studies 1 and 2 provide initial evidence that social and personality psychologists are changing the way that they report research. The clearest changes are greater consideration of statistical power, collecting larger samples, reporting exact *p*-values and measures of effect sizes, and appending supplemental information regarding methodological and analytical details. Although these changes are encouraging, social and personality psychological research is not yet a scientific utopia. For example, researchers described considerable publication pressure to selectively report only studies that work, and the literature continues to demonstrate a publication bias in favor of statistically significant results. Notwithstanding, social and personality psychologists are embracing research practices that should result in more replicable research moving forward.

The current studies also raise some important points pertaining to past methods used to evaluate the status of our science. In the impactful initial examination of the prevalence of different research practices, the majority of social and personality psychologists chose "yes" as opposed to "no" when asked if they had ever engaged in any of those practices (John et al., 2012). That survey used "yes" responses as evidence that people's practices were necessarily questionable without allowing them to explain their response. In Study 1, we provided an open-ended textbox where participants could elaborate on why they indicated using a QRP. Sometimes, our participants responded in ways that made it clear that they do not use the practice or that they misunderstood the question (e.g., they interpreted our question about "falsifying data" as about "falsifying hypotheses" instead), which suggests that the self-reported

prevalence of QRP use is probably lower than reported here and elsewhere due to measurement error. In many other cases, an examination of the reasons why researchers engaged in a "questionable" research practice was judged by independent coders as normatively acceptable (e.g., not conducting an a priori power analysis when there is no clear effect size estimate to use in the calculation, as is the case for more complex designs and analyses; dropping a condition or study when manipulation checks indicated that the manipulation did not work as intended; dropping items to improve scale reliability or because they do not factor with other items). Where researchers could improve when engaging in questionable but nonetheless acceptable research behaviors, however, is in transparently reporting when they engage in them.

Additionally, until now, metrics of replicability or scientific quality created by social and personality psychologists have generally only been applied to small sets of suspicious studies (e.g., Bem, 2011) and have done well in generating some statistic that suggests the findings were false positives and contain little or no evidentiary value (e.g., Simonsohn et al., 2014; for a review, see Schimmack, 2014b). As others have attempted to apply these methods more broadly to see whether evidential value varies by geographic regions, journals, sub-disciplines, university, or even time, these methods do less well. This difference may be due to the way the *p*-values and test statistics are collected. In some cases, *p*-values are collected by automatically searching abstracts using a search engine (e.g., de Winter & Dodou, 2015). This method is peculiar because norms about including statistics in the abstract change over time. De Winter and Dodou (2015) tried to account for this by also searching for words that seemed to imply significance (e.g., "significant difference" vs. "no significant difference"). Again, the past decade ushered in very short reports with tight word limits, which would make using those phrases less likely (Ledgerwood & Sherman, 2012). Others have tried collecting *p*-values and related test

statistics automatically by using computer programs that scrape these values from papers available online (e.g., Schimmack, 2015). Given the enormous number of significance tests included in most studies (hundreds in some studies included in Study 2), however, the vast majority of *p*-values included in these analyses are likely to be irrelevant to the critical hypothesis being tested. Therefore, it remains unclear what conclusions can be drawn from using scraping methods to obtain *p*-values, because the ways that test statistics are culled can affect the conclusions drawn.

The cautiously optimistic conclusions of the current studies cohere with the results of some replication efforts that suggest many effects do replicate. For example, the first Many Labs effort consistently replicated 10 out of 13 effects and found additional evidence that another effect sometimes replicated (Klein et al., 2014). Similarly, another Many Labs-type group replicated 8 out of 10 effects in 25 different research labs who ran their replication studies (Schweinsberg et al., 2016). Therefore, some replication efforts tell a similar story as the data presented in this manuscript – the field may not be *rotten to the core* and at least some social/personality psychology effects replicate. The Open Science Collaboration's (2015) Reproducibility Project, however, only successfully replicated 39% of the 100 studies it set out to replicate and the Many Labs 3 (Ebersole et al., in press; cf. Gilbert et al., 2016) effort replicated just 3 out of 10 effects. How can we account for these discrepancies in the estimated replicability of social and personality science?

One possible answer is that replication rates may be linked to how studies are selected for replication. Many Labs 1, for example, chose studies that the authors predicted were highly likely to replicate, whereas Many Labs 3 selected studies that could be conducted quickly. The Open Science Collaboration chose the last study appearing in various social and experimental

journals in 2008. Schweinsberg and colleagues (2016) chose effects that they had observed in the principal investigator's past unpublished research. The field has only just begun to give serious attention to replication. Among other things, the field does not have a clear definition of replication success versus failure. Different definitions of replicability can lead to very different conclusions. For example, Patil, Peng, and Leek (2016) defined a replication success as when the 95% prediction interval for the effect size estimate of the replication study, computed around the results of the original study, includes the actual point estimate from the replication. Using this definition of replicability, they concluded that 75%, not 39%, of the studies in the Reproducibility Project replicated. It is also likely to take some time to amass a large enough corpus of evidence to allow for strong inferences about the replicability of the field at large. Alternatively, given some of our replicability indices were consistent with a conclusion of *rottenness*, and may be better signals than the majority of indices that did suggest evidentiary value.

That said, we believe the totality of our findings—the survey data, actual research practices, and the majority of the indices of evidentiary value--converge on an optimistic conclusion. The field may not yet be perfect, but it does not appear to be (or becoming) rotten. While our findings provide additional information in the field's continuing quest to assess the status of our science, our efforts should be considered as only one of a growing number of contributions to understanding the evidentiary value of the field, and not the final word.

Furthermore, the current endeavor focuses on replicability and research practices that may affect future replicability of publishing findings. We acknowledge that this is only one constituent element of research quality, and that replicability should not be conflated with research quality. Quality research should replicate, but replicable research is not necessarily valid

(e.g., if the measures lack construct validity). Beyond replicability, there are numerous other features of high quality research. For example, high quality research should be cumulative, building bridges between past findings and paving a path towards innovative future studies (Finkel, Eastwick, & Reis, in press). Therefore, the current studies should be viewed as one element of a broader discussion on how to promote better science.

Generalizability and Broader Implications

Publication bias and research replicability have been the subject of much discussion in numerous fields besides social and personality psychology in recent years. Ioannidis (2005) proclamation that most research findings are false, for example, focused primarily on biomedical research. Scientists in other disciplines have begun to examine how replicable their findings are as well (e.g., Mullinix et al., 2015). To our knowledge, however, ours is the first effort to estimate the replicability of an existing corpus of knowledge in a field by comparing the conclusions of various statistical measures of replicability on a large sample of existing studies. Our results sound a cautiously optimistic note that at least some (according to one group of indices) or even a good deal (according to other indices) of our existing knowledge in social/personality is likely to be replicable. That said, research practices and incentives vary dramatically across disciplines. Few social/personality psychology studies have millions of dollars of investment and potential profit at stake, which may sometimes be the case (for example) in clinical trials of certain drugs or biomedical interventions. We therefore advise caution in generalizing anything about our findings to other disciplines, who should conduct their own investigations into the replicability of their findings in the context of their disciplines.

The broader implications of our work are therefore limited to conclusions about the soundness of social/personality psychology, and even within this specific context, caution is

warranted. We believe the work reported in this paper represents one small advance in addressing the soundness of our science. This study should be viewed as only a brick in the wall of evidence that is needed: Only converging evidence across studies using a variety approaches will provide a definitive answer to the question of the soundness of our science.

Conclusion

Is and/or was social and personality psychology wholly *rotten to the core*? No. Are the fruits being produced by social and personality psychologists getting better as the field embraces better norms? It seems so. In the current studies, we find evidence that social and personality psychologists are changing their research practices in ways that cohere with the ideals of the scientific utopia. At this early stage, it remains unclear whether replicability is improving, but if our research practices continue to improve, replicability should follow suit. The horrible revelation of outright fraud and the difficult observation that a number of social and personality psychological phenomena do not replicate may be seen as the worst of times, the age of foolishness and incredulity, and the winter of despair for a dystopian science. Our evidence suggests that the field was not wholly *rotten to the core* a decade ago (prior to the discussion on the state of our science, and the current emphasis on best practices) and that our field shows some improvements. Greater awareness of the problems with questionable practices and the benefits of at least some of the proposed best practices suggest that there are reasons to judge the product of these discussions as the best of times that will bring us that much closer to an increasingly utopian science.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fielder, K., Fiedler, S., Funder, D. C., Kleigl, 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,* 108–119. http://dx.doi.org/10.1002/per.1919

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678. http://dx.doi.org/10.3758/s13428-011-0089-5

Barrett, L. F. (2015, September 1). Psychology is not in crisis. Retrieved from http://www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html?_r=0

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. http://dx.doi.org/10.1016/j.jesp.2016.02.003

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research, 116*, 116-126. http://dx.doi.org/10.1161/CIRCRESAHA.114.303819

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The complete academic: A career guide (pp. 171–201)*. Washington, DC: American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425. http://dx.doi.org/10.1037/a0021524

Bezerianos, A., Tong, S., & Thakor, N. (2003). Time-dependent entropy estimation of EEG rhythm changes following brain ischemia. *Annals of Biomedical Engineering, 31*, 221-232. http://dx.doi.org/10.1114/1.1541013

Carter, E. C., Kofler, L.M., Forster, D.E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General, 144,* 796 – 815. http://dx.doi.org/10.1037/xge0000083

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015-1026. http://dx.doi.org/10.1037/0022-3514.58.6.1015

Cohen, B. (2008). *Explaining psychological statistics*. Hoboken, NJ: Wiley & Sons, Inc.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65,* 145–153. http://dx.doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition).* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64,* 464-494. http://dx.doi.org/10.1086/318641

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrapping methods and their application*. New York, NY: Cambridge University Press.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly, 64*, 413-428. http://dx.doi.org/10.1086/318638

de Winter, J. C., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, *3*, e733. http://dx.doi.org/10.7717/peerj.733

Dickens, C. (1859). *A tale of two cities*. London: Champman & Hall

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (in press). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Boca Raton, FL: CRC press.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3-6. http://dx.doi.org/10.1177/0956797613512465

Engber, D. (2016, March, 6). Everything is crumbling. An influential psychological theory, borne out in hundreds of experiments, may have just been debunked. How can so many scientists have been so wrong? Retrieved from http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.html

Enserink, M. (2012). Final report: Stapel affair points to bigger problems in social psychology. *Science Insider*. Retrieved December 26, 2012 from http://news.sciencemag.org/people-events/2012/11/final-report-stapel-affair-points-bigger-problems-social-psychology.

Fiedler, K. & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological & Personality Science, 7*, 45-52. http://dx.doi.org/10.1177/1948550615612150

Finkel, E., Eastwick, P. W., & Reis, H. T. (in press). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*.

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*, e109019. http://dx.doi.org/10.1371/journal.pone.0109019

Funder, D. C. (2016, March, 6). What if Gilbert was right? Retrieved from https://funderstorms.wordpress.com/2016/03/06/what-if-gilbert-is-right/

Gilbert, D. T., King, G. Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*, 1037. http://dx.doi.org/10.1126/science.aad7243

Giner-Sorolla, R. (2016). Approaching a fair deal for significance and other concerns. *Journal of Experimental Social Psychology*, *65*, 1-6. http://dx.doi.org/10.1016/j.jesp.2016.01.010

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, *83*, 314-320. http://dx.doi.org/10.1037/0033-2909.83.2.314

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software, 27*, 1-32. http://dx.doi.org/10.18637/jss.v027.i05

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61 – 85. http://dx.doi.org/10.3102/10769986009001061

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis *The American Statistician, 55*, 19-24. http://dx.doi.org/10.1198/000313001300339897

Holbrook, A., Krosnick, J. A., & Pfent, A. (2007). The causes and consequences of response rates in surveys by the news media and government contractor survey research firms. *Advances in telephone survey methodology*, 499-528.

Inbar, Y. (2016). Association between contextual dependence and replicability in psychology may be spurious. *Proceedings of the National Academy of Sciences*, 201608676.

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

Inzlicht, M. (2016). Reckoning with the past. Retrieved from http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124. http://dx.doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. http://dx.doi.org/10.1177/0956797611430953

Keeter, S., Miller, C., Kohut, A., Groves, R. M., & Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, *64*, 125-148.

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

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about the sources of unethical decisions at work. *Journal of Applied Psychology, 95,* 1 – 31. http://dx.doi.org/10.1037/a0017103

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Cemalcilar, Z. (2014). Data from investigating variation in replicability: A "Many Labs" Replication Project. *Journal of Open Psychology Data*, *2*.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107 – 112.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*, 570-583. http://dx.doi.org/10.1177/0146167211400619

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

Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science, 7*, 60 – 66. http://dx.doi.org/10.1177/1745691611427304

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

Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? The New Yorker. Retrieved from http://jackmccallum .com/The%20Truth%20Wears%20Off--Scientific%20 Method%20Error.pdf

Maasoumi, E., & Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics, 107,* 291-312. http://dx.doi.org/10.1016/S0304-4076(01)00125-7

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

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

Merkle, D., & Edelman, M. (2002). Non-response in exit polls: A comprehensive analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A., Little (Eds.) *Survey Non-response* (pp. 243-258). New York: Wiley.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2,* 109-138.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23,* 217-243. http://dx.doi.org/10.1080/1047840X.2012.692215

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

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

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531-536. http://dx.doi.org/10.1177/1745691612463401

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*, 539-544. http://dx.doi.org/10.1177/1745691616646366

Resnick, B. (2016, March 25). What psychology's crisis means for the future of science. Retrieved from http://www.vox.com/2016/3/14/11219446/psychology-replication-crisis

Richard, F. D., Bond, C.F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*, 754-775.

Robert, B. W. (2015, September 17). The new rules of research. Retrieved from https://pigee.wordpress.com/2015/09/17/the-new-rules-of-research/

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551-566. http://dx.doi.org/10.1037/a0029487

Schimmack, U. (2014a). *The test of insufficient variance (TIVA): A new tool for the detection of questionable research practices.* Retrieved from https://replicationindex.wordpress.com/2014/12/30/the-test-ofinsufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices

Schimmack, U. (2014b). *Quantifying statistical research integrity: The replicability index.* Retrieved from http://www.r-index.org/uploads/3/5/6/7/3567479/introduction_to_the_r-index__14-12-01.pdf

Schimmack, U. (2015). Distinguishing questionable research practices from publication bias. Retrieved on December 16, 2015 from https://replicationindex.wordpress.com/2015/12/08/distinguishing-questionable-research-practices-from-publication-bias/

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., *Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C. Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei,

C., Wetter, E., Wilson, C., Darroux, S-C., & Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. http://dx.doi.org/10.1016/j.jesp.2015.10.001

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

Simonsohn, U. (2015). *Falsely reassuring: Analyses of ALL p-values*. Retrieved from http://datacolada.org/41.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-Curve: A key to the file drawer. *Journal of Experimental Psychology: General*, *143*, 534-547. http://dx.doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*, 1146-1152. http://dx.doi.org/10.1037/xge0000104

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8*, 291-311. http://dx.doi.org/10.1177/0959354398083001

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science*, *10*, 886-899. http://dx.doi.org/10.1177/1745691615609918

Stergiou, N., & Decker, L. M. (2011). Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Human Movement Science*, *30*, 869-888. http://dx.doi.org/10.1016/j.humov.2011.06.002

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59-71. http://dx.doi.org/10.1177/1745691613514450

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199747047.001.0001

Treviño, L. K. (1990). A cultural perspective on changing and developing organizational ethics. In R. Woodman & W. Passmore (Eds.), *Research in organizational change and development* (Vol. 4, pp. 195 – 230). Greenwich, CT: JAI Press.

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p-values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11,* 713–729. http://dx.doi.org/10.1177/1745691616650874

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*, 6454-6459. http://dx.doi.org/10.1073/pnas.1521897113

Vazire, S. (2016). Editorial. *Social Psychological & Personality Science, 7,* 3-7.

Venables, W. N., & Ripley, B. D. (2002). Random and mixed effects. In *Modern applied statistics with S* (pp. 271-300). New York, NY: Springer.

Victor, B., & Cullen, J. B. (1989). The organizational bases of ethical work climates. *Administrative Science Quarterly, 33,* 101 – 124. http://dx.doi.org/10.2307/2392857

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274-290.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. (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.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726-728. http://dx.doi.org/10.1037/0003-066X.61.7.726

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Houston, TX: Springer Science & Business Media.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology*, *55*, 493-518.

Figure 1. Perceived likelihood that studies published in various journals will replicate as a function of time with 95% Confidence Intervals.

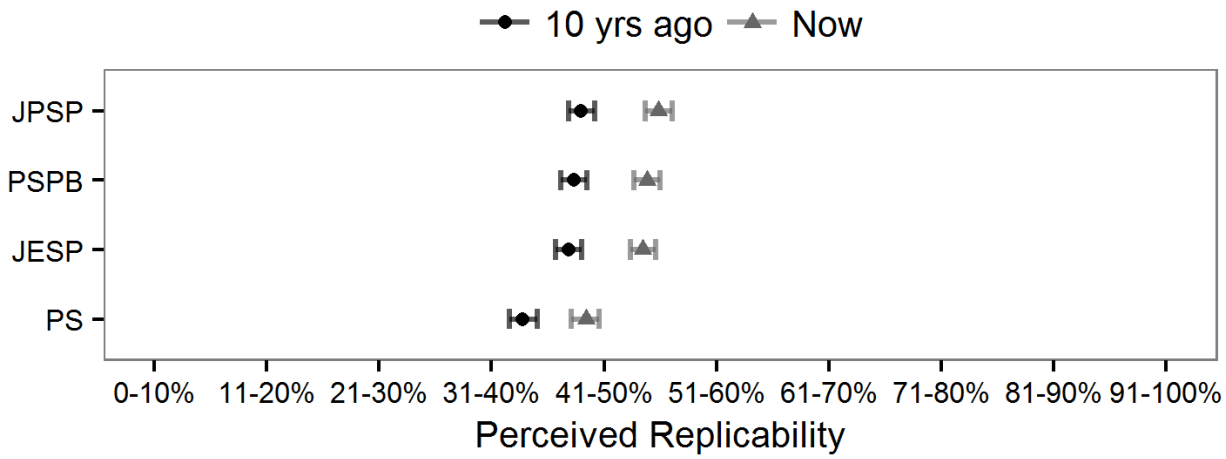

Figure 2. Self-reported frequency of using each research practice with 95% Confidence Intervals.

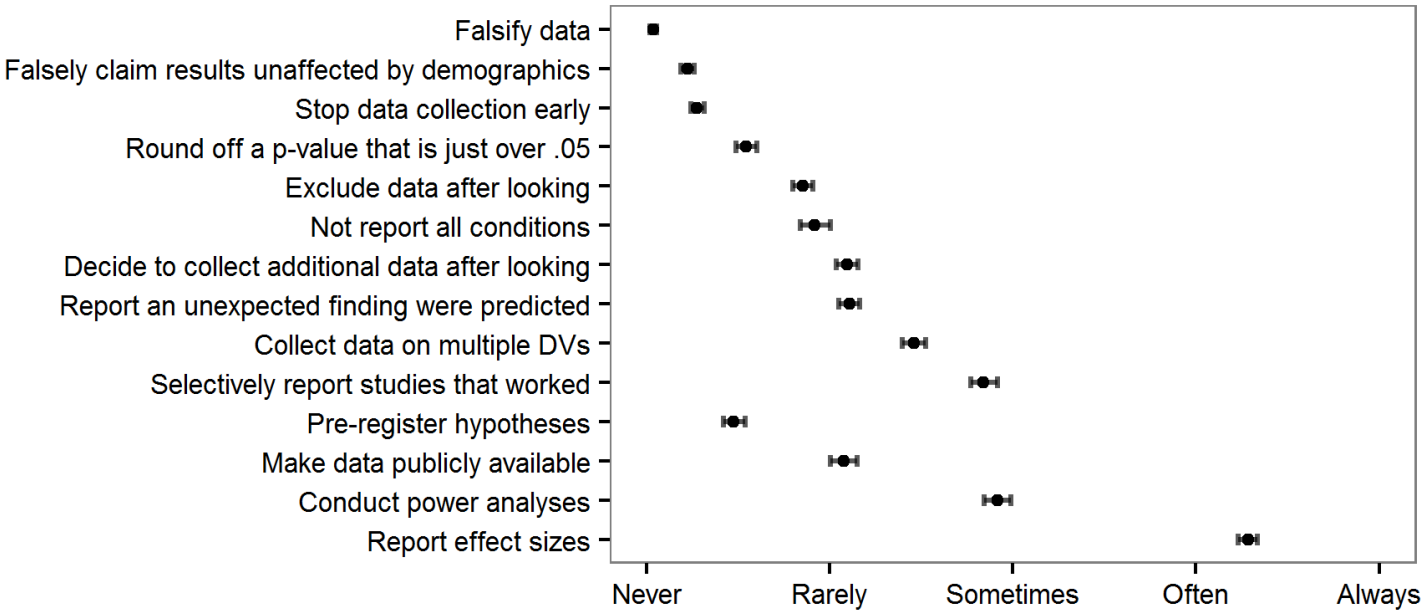

Figure 3. Acceptability and unacceptability of various research practices.

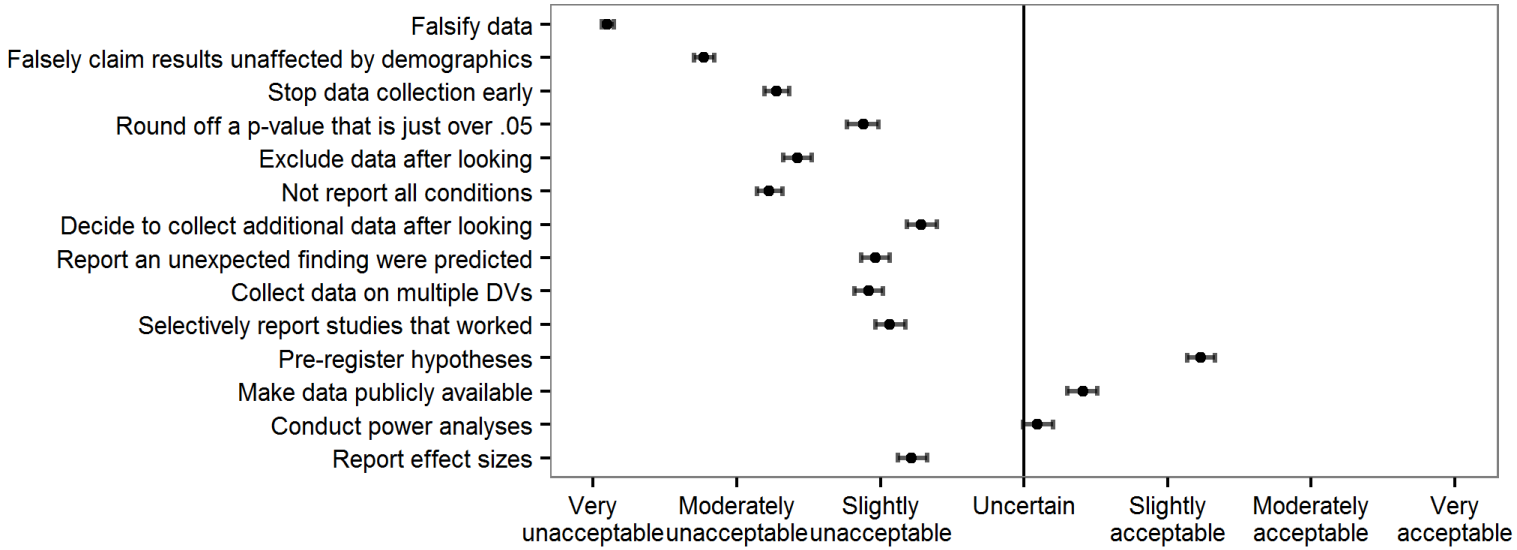

Figure 4. Reported likelihood of changing behavior as a function of the "status of our science" discussion.

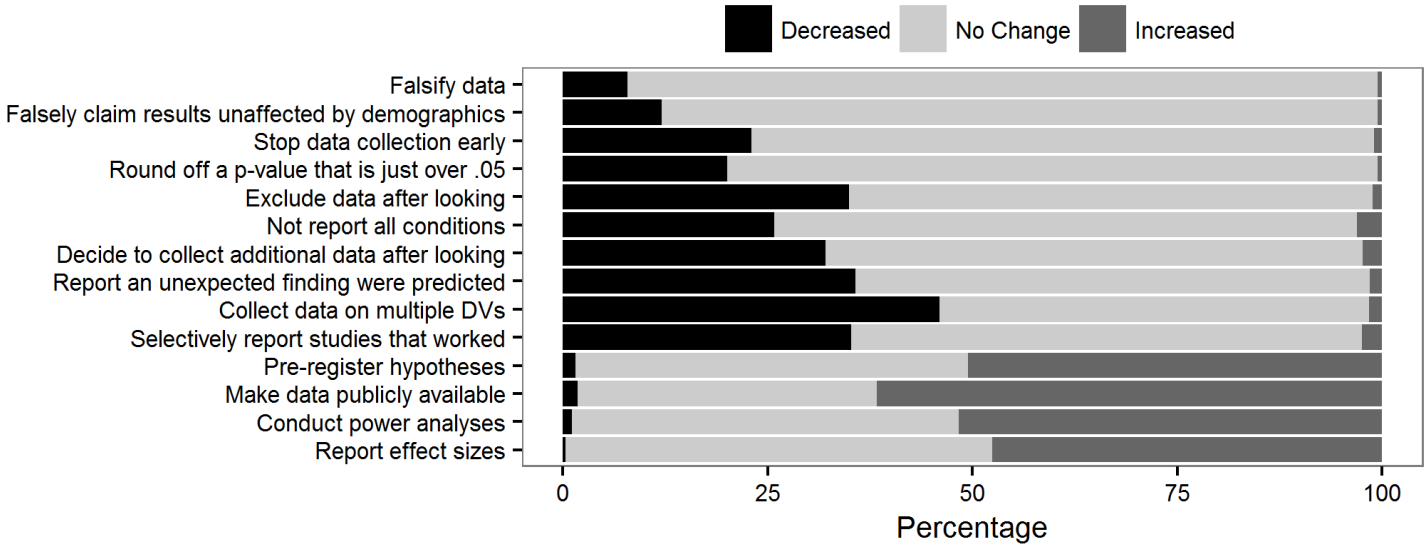

Figure 5. The distribution of *p*-values ranging from 0 to .05 calculated by Stouffer's mean within paper (solid line) and by median within paper (dotted line).

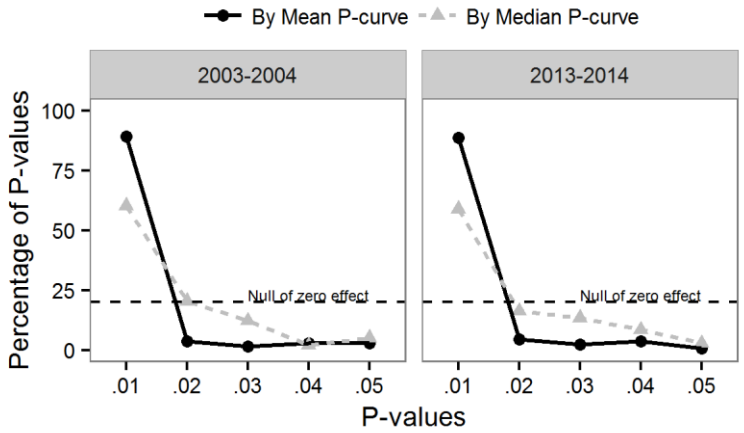

Figure 6. Top panel is the KDE density plot for the z-curves by time period and the bottom panel is a forest plot with 95% BCa CI for measures of central tendency.

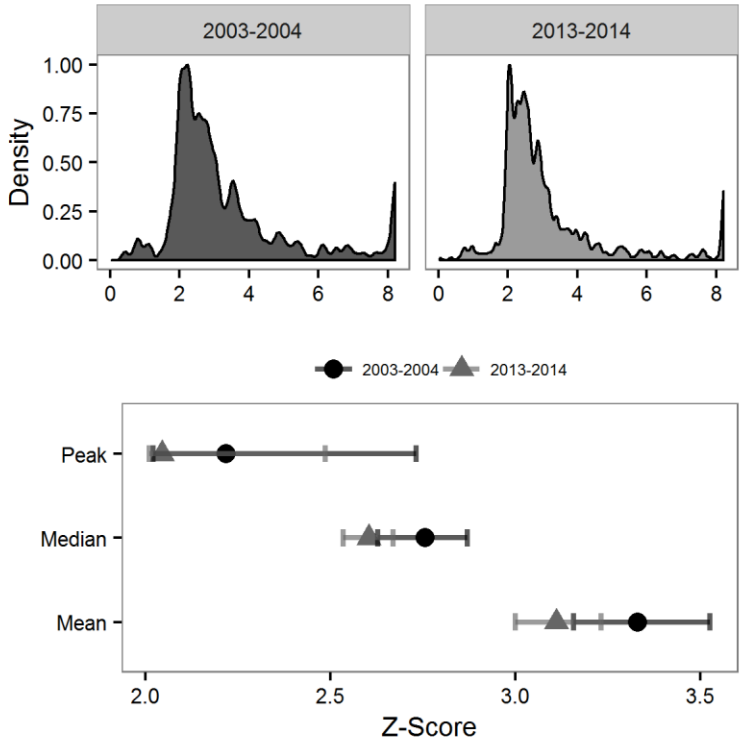

Figure 7. Top panel is the KDE density plot for the sample size (in Log_{10}) by time period and the bottom panel is a forest plot with 95% BCa CI for measures of central tendency.

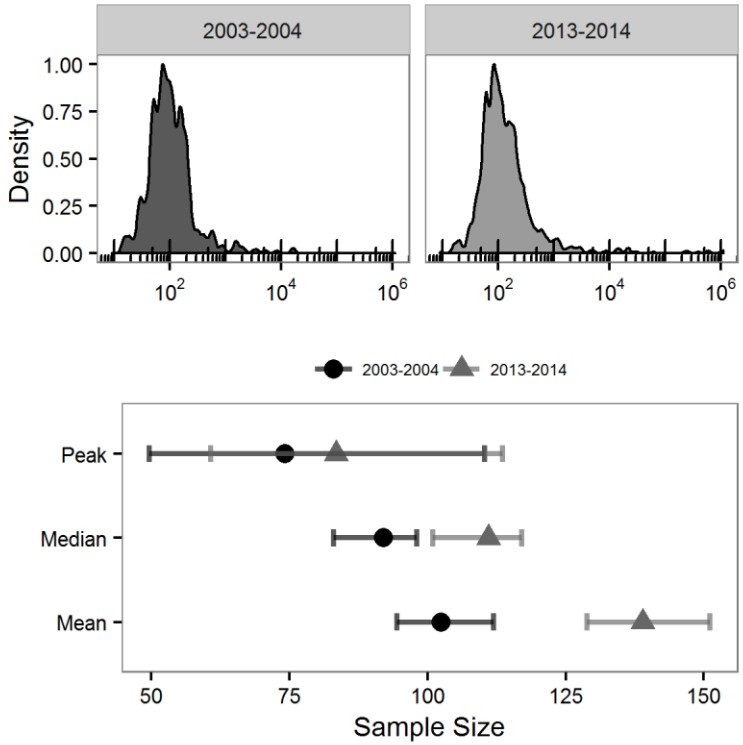

Figure 8. Top panel is the KDE density plot for observed power from reported test statistics (*t*, *F*, *r*) by time period and the bottom panel is a forest plot with 95% BCa CI for measures of central tendency.

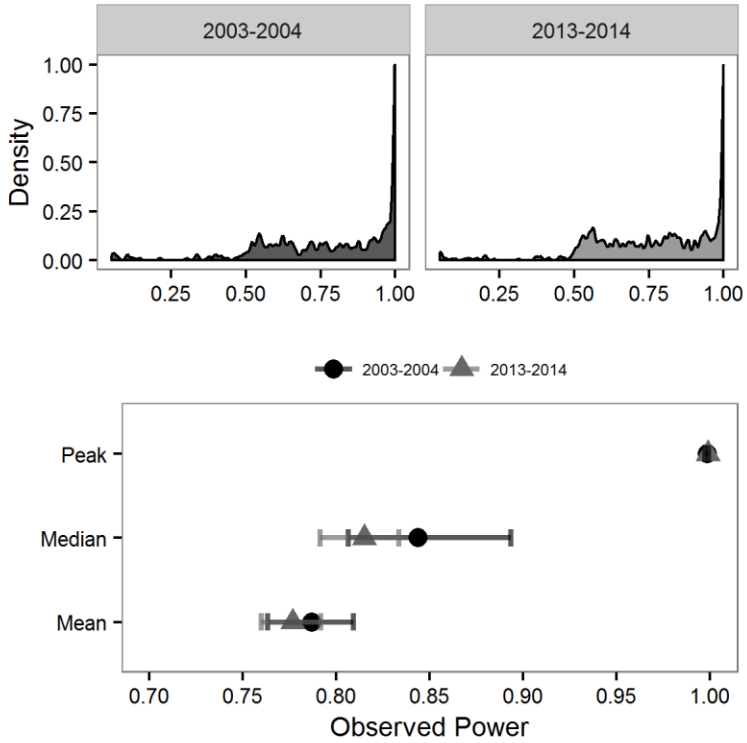

Figure 9. Summary of estimates of scientific quality rescaled to a common metric, where .80 is a conventionally accepted threshold for statistical power, except for the P-Curves which may be interpreted as the percent of studies containing evidentiary value (see Footnote 12). Error bars represent 95% BCa confidence intervals.

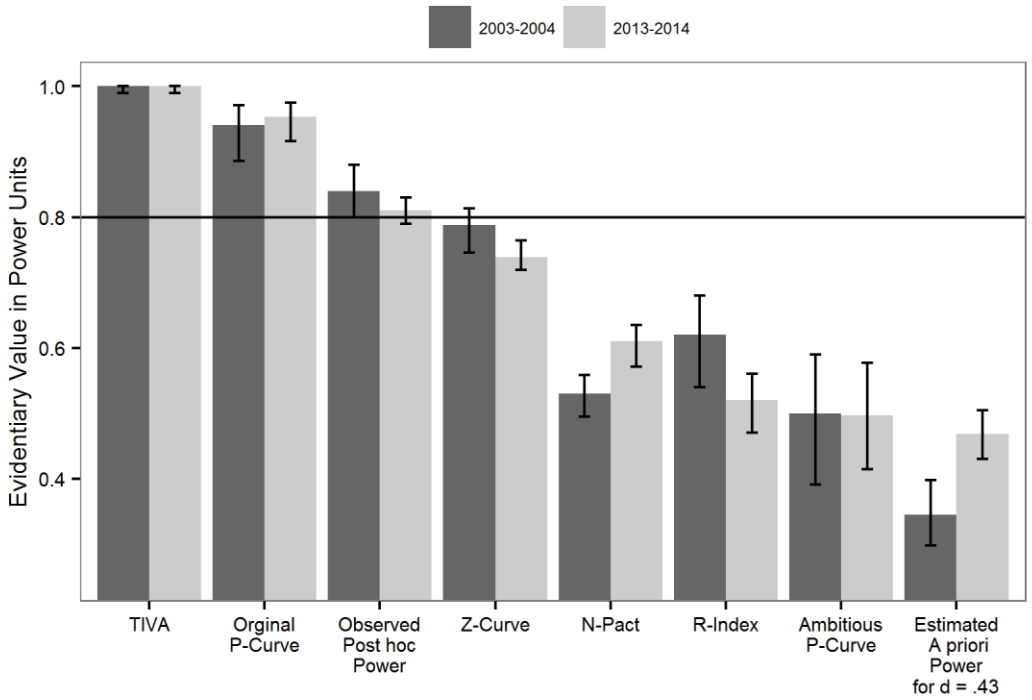

Table 1

Self-reported frequency of using "QRPs" and acceptability/unacceptability ratings of participants' justifications for using them

Practice	**John et al. (2012) % Yes**	**% Ever**	**Average frequency (SD)**	**% Acceptable**	**% QRP**
Selectively report studies that worked	46	84	2.84 (1.18)	41	55
Not report all measures	63	78	2.46 (1.06)	91	3
Report that unexpected findings were expected	27	58	2.11 (0.99)	72	26
Decide to collect additional data after looking	56	66	2.10 (0.96)	88	3
Not report all conditions	28	45	1.92 (1.32)	89	11
Exclude some data after looking at impact	38	58	1.85 (0.87)	95	1
Rounded down *p*-values > .05	22	33	1.54 (0.90)	89	6

Stop data collection early	16	18	1.27 (0.59)	81	8
Claim results were unaffected by demographics when they were	3	16	1.22 (0.57)	90	4
Falsify data	1	2[#]	1.04 (0.28)	0	0
Report effect sizes	-	99	4.29 (0.86)		
Conduct power analyses	-	87	2.92 (1.17)		
Make data publicly available	-	56	2.08 (1.19)		
Pre-register hypotheses	-	27	1.48 (0.92)		

Note: All differences in "lifetime ever" percentages were statistically different between our sample and John et al.'s sample at $p < .01$ except stopping data collection early and falsifying data. Given the two surveys used very different scales of measurement (John et al. used a *yes/no* measure, whereas we assessed frequency on a *not at all, rarely, sometimes, often, always* scale), these differences should be interpreted with considerable caution.

% Acceptable = cases where both coders rated the justification for using the behavior as acceptable, % QRP = cases where both coders rated the justification for using the behavior as unacceptable. Other responses were either disputed or uncodeable. Justifications were only provided when participants indicated that the behavior was acceptable.

[#] Twelve participants reported having ever falsified data. Open-ended explanations for their behavior, however, revealed that all but one of them clearly misunderstood the question or accidentally responded on the wrong end of the response scale. The one respondent who did not clearly misunderstand the question or evidence of a measurement responded with snark (an ambiguous reference to Bem, 2011).

Table 2

Example Justifications for Engaging in QRPs or Not Engaging in New Best Practices

Justifications for Using "Questionable Research Practices"	**Example Justifications**
Selectively report studies that worked	•When studies did not work for uninteresting reasons (e.g., experimenter error, failed manipulation checks) •Editors/reviewers requested that the study be dropped •Non-inclusion is acknowledged in the write-up •No one publishes failed studies
Not report all measures	•Variables were exploratory •Poor reliability/measurement problems •When reporting findings from large scale studies/surveys •Reviewers/editors ask you to drop them
Report that unexpected findings were expected	•When presented as an alternative and theoretically grounded hypothesis •A more careful literature review would have yielded the hypothesis •It was followed up with additional studies and it replicated

	•Editors'/reviewers' strong suggestion
Decide to collect additional data after looking	•Always acceptable •Observed power is lower than anticipated •Acceptable if one adjusts *p*-value to account for having peeked •After reaching stopping rule, want greater confidence results are real •Results are in the expected direction but are not significant
Not report all conditions	•Manipulation checks fail •The omitted conditions do not qualify the reported results/had no effect •An intended control condition is not perceived as neutral •Reporting a subset of a larger study/data set •Conditions included for exploratory purposes and not relevant to the main research question
Exclude some data after looking at impact	•Report results with and without exclusions •Participants fail instructional or other manipulation checks, reveal suspicion, etc.

	•Outliers/influence metrics •Non-native first language
Round off p-value	•APA style requirements •Conforms to norms on rounding •.05 is arbitrary •*p*-values are incidental, uninteresting
Stop data collection early	•The effect size or anticipated power is larger than anticipated •A stopping rule besides achieving a given sample size (e.g., end of the semester) •It becomes infeasible to persist (e.g., exhaust resources, graduation deadlines)
Claim results were unaffected by demographic variables when they were	*Nearly all responses involved explaining why respondents didn't test demographic differences (e.g., not enough power); none reported false claims of no differences when there were in fact differences*

Falsify data	*Nearly all responses in this category indicated that participants misunderstood the question, or reversed the scale anchors*
Justifications for Not Using Proposed Best Research Practices	
Report effect sizes	•Exploratory/pilot studies
Conduct power analyses	•There is no way to estimate effect size *a priori* •Exploratory/pilot studies
Make data publicly available	•Not required/not normative •Makes data available upon request •No IRB approval/confidentiality issues/legal issues/sensitive information •Intellectual property/plans to publish additional papers from data set •Data file too large/complex
Pre-register hypotheses	•Not required/not normative •Exploratory research/pilot studies

•Secondary analysis of existing data

•Fear of being scooped

Table 3

Summary statistics and best practices for all articles from JESP, JPSP, PSPB, and PS where the first author specialized in social/personality over two time periods with 95% BCa CIs

	2003-2004			2013-2014			
	Arithmetic Statistic	95% BCa CI LCI	UCI	Arithmetic Statistic	95% BCa CI LCI	UCI	Change?
Significance Testing							
ian # of Significance Tests	**15**	12	16	**10**	9	10	**Better**
n # of Significance Tests	**28.16**	23.85	35.54	**17.24**	15.74	19.12	**Better**
Significant Significance Hypothesis Tests	60.34	58.33	62.42	61.54	59.93	63.63	Same
Significant Significance Critical Hypothesis	89.17	85.43	92.13	92.01	89.69	93.82	Same
Significance Reporting Practices							
act p-values reported	**19.29**	16.10	23.03	**54.51**	51.29	57.89	**Better**
values inappropriately rounded down (> 0)	32.89	23.19	44.63	34.34	29.79	39.18	Same
values inappropriately rounded down (>-.004)	10.53	4.69	19.16	5.01	3.15	7.50	Same
ypothesis tests reporting an effect size	**19.22**	16.09	22.87	**49.65**	46.31	52.58	**Better**

Participant Exclusion							
entage of Hypotheses Excluding Participants	25.60	22.01	29.28	28.75	25.85	31.53	Same
entage of Participants Excluded	2.52	1.99	3.36	3.28	2.74	3.98	Same
Additional Reporting of Details							
n # of Analysis-related Footnotes	0.76	0.68	0.86	0.70	0.62	0.81	Same
Studies Reporting Additional Analyses in SI	**1.36**	0.00	3.18	**8.59**	5.52	11.96	**Better**
Complexity of Designs							
n # of Studies per Paper	**2.39**	2.20	2.59	**3.15**	2.94	3.61	·
n # of Predictors per Hypothesis	**8.42**	7.95	8.91	**7.54**	7.24	7.88	·
n # of Conditions per Hypothesis	8.81	8.34	9.29	8.76	8.43	9.09	·
n # of Covariates per Hypothesis	3.37	3.06	3.75	3.26	3.06	3.51	·

Note: **Bold values indicate where the 95% BCa CIs do not overlap. We interpreted reporting exact p-values, not rounding down, reporting effect sizes, including additional analyses in supplemental materials, and reducing the number of predictors, as better practices. For the Complexity of Designs measures, we report comparisons but cannot infer whether this reflects anything about quality or replicability.*

Table 4

Statistics regarding replicability for all articles from JESP, JPSP, PSPB, and PS where the first author was specialized in social/personality over two time periods with 95% BCa CIs

	2003-2004				**2013-2014**				
		95% BCa CI				95% BCa CI			
	Statistic	LCI	UCI	EV*	Statistic	LCI	UCI	EV*	Change?
Published Methods									
inal P-Curve	94.03%	88.59	97.06	Yes	95.38%	91.64	97.53	Yes	Same
itious P-Curve	50.00%	39.07	58.97	Med	49.64%	41.48	57.75	Med	Same
ıct (median sample size)	**92**	84	98	-	**111**	101	117	-	**Better**
iori Power (% .8 power at *d* = .43)	**34.50%**	29.76	39.82	Low	**46.87%**	43.03	50.49	Low	**Better**
erved (Post-hoc) Power (median)	0.84	0.80	0.88	Yes	0.81	0.79	0.83	Yes	Same
Unpublished Methods									
A	**1610.36**	1309.05	1933.91	Yes	**2402.14**	2030.41	2864.5	Yes	**Better**
ırve (median)	2.76	2.62	2.85	Yes	2.60	2.54	2.68	Yes	Same

dex	0.62	0.54	0.68	Low	0.52	0.47	0.56	Low	Same

Note: *Bold values indicate where the 95% BCa CIs do not overlap.*

**Summary judgments for whether there was evidentiary value were made using the interpretations provided by the authors of those metrics. Summary judgments regarding what change there was over time are based on whether the time points differed and how the authors of those metrics suggest interpreting those metrics.*

Appendix

For *t*-values conversion to effect size see equations 1 and 2.

$$R^2 = \frac{t^2}{t^2 + df} \quad (1)$$

$$cohen's\ d = \frac{2r}{1-R^2} \quad (2)$$

For, *F*-values conversion to effect size see equations 3 and 4.

$$\eta^2 = \frac{df_{num}F}{df_{num}F + df_{denom}} \quad (3)$$

$$f^2 = \frac{\eta^2}{1-\eta^2} \quad (4)$$

For, χ^2-values conversion to effect size see equation 5.

$$\omega = \sqrt{\left(\frac{\chi^2}{N}\right)} \quad (5)$$