Putting the Self in Self-Correction: Findings from the Loss-of-Confidence Project

Julia Rohrer,[1,2] Warren Tierney,[3] Eric L. Uhlmann,[3] Lisa M. DeBruine,[4] Tom Heyman,[5, 6] Benedict Jones,[7] Stefan C. Schmukle,[2] Raphael Silberzahn,[8] Rebecca M. Willén,[9] Rickard Carlsson,[10] Richard E. Lucas,[11] Julia Strand,[12] Simine Vazire,[13] Jessica K. Witt,[14] Thomas R. Zentall,[15] Christopher F. Chabris,[16] Tal Yarkoni[17]

1 International Max Planck Research School on the Life Course, Max Planck Institute for Human Development, Berlin; 2 Department of Psychology, University of Leipzig; 3 INSEAD, Singapore; 4 University of Glasgow; 5 Laboratory of Experimental Psychology, KU Leuven; 6 Institute of Psychology, Leiden University; 7 School of Psychological Sciences and Health, University of Strathclyde; 8 University of Sussex, Sussex Business School, Brighton, UK; 9 Institute for Globally Distributed Open Research and Education (IGDORE); 10 Linnaeus University; 11 Department of Psychology, Michigan State University; 12 Carleton College; 13 University of Melbourne; 14 Colorado State University; 15 University of Kentucky; 16 Geisinger Health Systems; 17 Department of Psychology, University of Texas at Austin

This is an unedited manuscript accepted for publication at Perspectives on Psychological Science. The manuscript will undergo further copyediting and typesetting.

Author Note

Correspondence concerning this article should be addressed to Julia M. Rohrer (julia.rohrer@uni-leipzig.de). Supplemental Materials for this manuscript are provided on the Open Science Framework, https://osf.io/bv48h/.

Contributions: TY and CFC initialized the project in 2016. JMR managed the project starting from 2017, launched the corresponding website, and took the lead in writing the manuscript. WT and ELU took the lead in designing the Loss-of-Confidence survey, with numerous authors providing feedback and edits on the survey content. LMD, TH, BJ, SCS, RS, ELU, RMW, and TY submitted Loss-of-Confidence statements during the first round of data collection. RC, REL, JS, SV, JKW, TRZ, and CFC submitted statements at a later point in time. All authors provided critical feedback and helped shape the manuscript. Authorship order was determined by the following rule: [lead author, JMR], [authors who led the survey, WT and ELU], [authors of Loss-of-Confidence statements received during first round of data collection in alphabetical author], [authors of Loss-of-Confidence statements received later in alphabetical order], [senior authors, CFC and TY].

Acknowledgments: We would like to thank Michael Inzlicht, Alison Ledgerwood, Kateri McRae, and Victoria Savalei, who all contributed to the initial draft of the project concept, and Nick Brown, who proofread an earlier version of the manuscript.

Funding: JKW is supported by the National Science Foundation (BCS-1632222), and WT and ELU's work was supported by an R&D grant from INSEAD. Part of this research was conducted while TH was a postdoctoral fellow of the Research Foundation-Flanders (FWO-Vlaanderen).

Abstract

Science is often perceived to be a self-correcting enterprise. In principle, the assessment of scientific claims is supposed to proceed in a cumulative fashion, with the reigning theories of the day progressively approximating truth more accurately over time. In practice, however, cumulative self-correction tends to proceed less efficiently than one might naively suppose. Far from evaluating new evidence dispassionately and infallibly, individual scientists often cling stubbornly to prior findings. Here we explore the dynamics of scientific self-correction at an individual rather than collective level. In thirteen written statements, researchers from diverse branches of psychology share why and how they have lost confidence in a published finding. We qualitatively characterize these disclosures and explore their implications. A cross-disciplinary survey suggests that such loss-of-confidence sentiments are surprisingly common among members of the broader scientific population, yet rarely become part of the public record. We argue that removing barriers to self-correction at the individual level is imperative if the scientific community as a whole is to achieve the ideal of efficient self-correction.

Keywords: self-correction, knowledge accumulation, metascience, scientific falsification, incentive structure, scientific errors

Putting the Self in Self-Correction: Findings from the Loss-of-Confidence Project

Science is often hailed as a self-correcting enterprise. In the popular perception, scientific knowledge is cumulative and progressively approximates truth more accurately over time (Sismondo, 2010). However, the degree to which science is genuinely self-correcting is a matter of considerable debate. The truth may (or may not) be revealed eventually, but errors can persist for decades; corrections sometimes reflect lucky accidents rather than systematic investigation, and can themselves be erroneous; and initial mistakes might give rise to subsequent errors before they get caught (Allchin, 2015). Furthermore, even in a self-correcting scientific system, it remains unclear how much of the knowledge base is credible at any given time (Ioannidis, 2012), since the pace of scientific self-correction may be far from optimal.

Usually, self-correction is construed as an outcome of the activities of the scientific community as a whole (i.e., collective self-correction): watchful reviewers and editors catch errors before studies get published; critical readers write commentaries when they spot flaws in somebody else's reasoning; replications by impartial groups of researchers allow the scientific community to update their beliefs about the likelihood that a scientific claim is true. Far less common are cases in which researchers publicly point out errors in their own studies, and question conclusions they have drawn before (i.e., individual self-correction). The perceived unlikeliness of such an event is facetiously captured in Max Planck's famous statement that new scientific truths become established not because their enemies see the light, but because those enemies eventually die (Planck, 1948). However, even if individual self-correction is not necessary for a scientific community as a whole to be self-correcting in the long run (Mayo-Wilson, Zollman, & Danks, 2011), we argue that it can increase the overall efficiency of the self-corrective process and thus contribute to a more accurate scientific record.

The Value of Individual Self-Correction

The authors of a study have privileged access to details about how the study was planned and conducted, how the data were (pre-)processed, and which analyses were performed. Thus, the authors remain in a special position to identify or confirm a variety of procedural, theoretical and methodological problems that are less visible to other researchers.[1] Even when the relevant information can in principle be accessed from the outside, correction by the original authors might still be associated with considerably lower costs. For an externally instigated correction to take place, skeptical "outsiders" who were not involved in the research effort might have to carefully reconstruct methodological details from a scant methods section (see, e.g., Chang, Li, et al., 2018; Hardwicke et al., 2018, for evidence that often, authors' assistance is required to reproduce analyses); write persuasive emails to get the original authors to share the underlying data (often to no avail; Wicherts, Bakker, & Molenaar, 2011); recalculate statistics, as reported values are not always accurate (e.g., Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016); or apply advanced statistical methods to assess evidence in the presence of distortions such as publication bias (Carter, Schönbrodt, Gervais, & Hilgard, 2019).

Eventually, external investigators might resort to an empirical replication study to clarify the matter. A replication study can be a very costly or even impossible endeavor. Certainly, it is inefficient when a simple self-corrective effort by the original authors might have sufficed. Widespread individual self-correction would obviously not eliminate the need for replication, but it would enable researchers to make better-informed choices about whether and how to replicate—with over 30 million scientific articles published since 1965

[1] Guidelines to promote openness (e.g., Nosek et al., 2015) might partly reduce this asymmetry and thus make it easier for third parties to spot flaws.

(Pan, Petersen, Pammolli, & Fortunato, 2018), limited research resources should not be expended mindlessly on attempts to replicate everything (see also Coles, Tiokhin, Scheel, Isager, & Lakens, 2018). In some cases, individual self-correction could render an empirical replication study unnecessary. In other cases, additionally disclosed information might render an empirical replication attempt even more interesting. And in any case, full information about the research process, including details that make the original authors doubt their claims, would help external investigators maximize the informativeness of their replication or follow-up study.

Lastly, in many areas of science, scientific correction has become a sensitive issue often discussed with highly charged language (Bohannon, 2014). Self-correction could help defuse some of this conflict. A research culture in which individual self-corrections are the default reaction to errors or misinterpretations could raise awareness that mistakes are a routine part of science and help separate researchers' identities from specific findings.

The Loss-of-Confidence Project

To what extent does our research culture resemble the self-correcting ideal, and how can we facilitate such behavior? To address these questions, and to gauge the potential impacts of individual self-corrections, we conducted the Loss-of-Confidence Project. The effort was born out of a discussion in the Facebook group PsychMAP following the online publication of Dana Carney's statement "My Position on Power Poses" (Carney, 2016). Carney revealed new methodological details regarding one of her previous publications and stated that she no longer believed in the originally reported effects. Inspired by her open disclosure, we conducted a project consisting of two parts: an open call for loss-of-confidence statements, and an anonymous online survey.

First, in our open call, we invited psychological researchers to submit statements describing findings that they had published and in which they had subsequently lost confidence.[2] The idea behind the initiative was to help normalize and de-stigmatize individual self-correction, while (hopefully) also rewarding authors for exposing themselves in this way with a publication. We invited authors in any area of psychology to contribute statements expressing a loss of confidence in previous findings, subject to the following requirements:

1. The study in question was an empirical report of a novel finding;
2. The submitting author has lost confidence in the primary/central result of the paper;
3. The loss of confidence occurred primarily as a result of theoretical or methodological problems with the study design or data analysis;
4. The submitting author takes responsibility for the errors in question.

The goal was to restrict submissions to cases where the stigma of disclosing a loss of confidence in previous findings would be particularly high; we therefore did not accept cases where an author had lost faith in a previous finding for reasons that did not involve their own mistakes (e.g., due to a series of failed replications by other researchers).

Second, to understand whether the statements received in the first part of the project are outliers, or reflect a broader phenomenon that goes largely unreported, we carried out an online survey and asked respondents about their experience with losses of confidence. Supplementary Table 1 provides the full list of questions asked. The link to the survey was posted on Facebook pages and mailing lists oriented towards scientists (Psych MAP,

[2] An archived version of the website can be found at https://web.archive.org/web/20171212055615/https://lossofconfidence.

Psychological Methods Discussion Group, International Social Cognition Network, JDM Society, SJDM mailing list), and further promoted on Twitter. Survey materials and anonymized data are made available on the project's OSF repository (https://osf.io/bv48h).

Results: Loss-of-Confidence Statements

The project was disseminated widely on social media (resulting in around 4,700 page views of the project website), and public commentary was overwhelmingly positive, highlighting how individual self-correction is aligned with perceived norms of scientific best practices. By the time we stopped the initial collection of submissions (December 2017 to July 2018), we had received Loss-of-Confidence statements pertaining to six different studies. After posting a preprint of an earlier version of this manuscript, we re-opened the collection of statements and received seven more submissions, some of them while finalizing the manuscript. Table 1 provides an overview of the statements we received.[3]

In the following, we list all statements in alphabetical order of the first author of the original study to which they pertain. Some of the statements have been abbreviated, the long versions are available on the Open Science Framework (https://osf.io/bv48h/).

[3] Readers are cautioned to infer nothing about original authors who did not join or sign a loss-of-confidence statement about their own paper. In some cases, these authors approved of the submission but did not get involved otherwise; in others, they had already left the field of research.

Table 1

Overview of the Loss-of-Confidence Statements

Authors	Title	Journal	JIF	Citations
Carlsson and Björklund (2010)	Implicit stereotype content: Mixed stereotypes can be measured with the implicit association test.	Social Psychology	1.36	74
Chabris and Hamilton (1992)	Hemispheric specialization for skilled perceptual organization by chessmasters.	Neuropsychologia	2.87	28
Fisher, Hain, DeBruine, and Jones (2015)	Women's preference for attractive makeup tracks changes in their salivary testosterone.	Psychological Science	4.90	9
Heyman, Van Rensbergen, Storms, Hutchison, and De Deyne (2015)	The influence of working memory load on semantic priming.	Journal of Experimental Psychology: Learning, Memory and Cognition	2.67	51
Lucas and Diener (2001)	Understanding extraverts' enjoyment of social situations: the importance of pleasantness.	Journal of Personality and Social Psychology	5.92	220
Schmukle, Liesenfeld, Back and Egloff (2007)	Second to fourth digit ratios and the implicit gender self-concept.	Personality and Individual Differences	2.00	20
Silberzahn and Uhlmann (2013)	It pays to be Herr Kaiser: Germans with noble-sounding surnames more often work as managers than as employees.	Psychological Science	4.90	28
Smith and Zentall (2016)	Suboptimal choice in pigeons: choice is primarily based on the value of the conditioned reinforcer rather than overall reinforcement rate.	Journal of Experimental Psychology: Animal Learning and Cognition	2.03	64
Strand, Brown, and Barbour (2018)	Talking points: A modulating circle reduces listening effort without improving speech recognition.	Psychonomic Bulletin & Review	3.70	9
Vazire (2010)	Who knows what about a person? The self-other knowledge asymmetry (SOKA) model.	Journal of Personality and Social Psychology	5.92	740
Willén and Strömwall (2012)	Offenders' lies and truths: an evaluation of the Supreme Court of Sweden's criteria for credibility assessment.	Psychology Crime & Law	1.46	19
Witt and Proffitt (2008)	Action-specific Influences on Distance Perception: A Role for Motor Simulation	Journal of Experimental Psychology: Human Perception and Performance	2.94	252
Yarkoni, Braver, Gray and Green (2005)	Prefrontal brain activity predicts temporally extended decision-making behavior.	Journal of the Experimental Analysis of Behavior	2.15	45

Note. JIF = Journal Impact Factor 2018 according to InCites Journal Citation Reports and citations according to Google Scholar, retrieved on April 27, 2019.

Statement on Carlsson and Björklund (2010) by Rickard Carlsson

In this study, we developed a new way to measure mixed (in terms of warmth and competence) stereotypes with the help of the implicit association test (IAT). In two studies, respondents took two IATs and results supported the predictions: Lawyers were implicitly stereotyped as competent (positive) and cold (negative) relative to preschool teachers. In retrospect, there are a number of issues with the reported findings. First, there was considerable flexibility in what counted as support for the theoretical predictions. In particular, the statistical analysis in Study 2 tests a different hypothesis than Study 1. This analysis was added after peer review round 2 and thus was definitely not predicted a priori. Later, when trying to replicate the reported analysis from Study 1 on the data from Study 2, I found that only one of the two effects reported in Study 1 could be successfully replicated. Second, when we tried to establish the convergent and discriminant validity of the IATs by correlating them with explicit measures, we committed the fallacy of taking a nonsignificant effect in an underpowered test as evidence for the null hypothesis, which, in this case, implied discriminant validity. Third, in Study 1, participants actually took a third IAT which measured general attitudes towards the groups. This IAT was not disclosed in the manuscript and was highly correlated with both the competence and the warmth IAT. Hence, it would have complicated our narrative and undermined the claim that we had developed a completely new measure. Fourth, data from an undisclosed behavioral measure was collected but never entered into dataset or analyzed because I made a judgement that it was invalid based on debriefing of the participants. In conclusion, in this 2010 article I claimed to have developed a way to measure mixed stereotypes of warmth and competence with the IAT. I am no longer confident in this finding.

Statement on Chabris and Hamilton (1992) by Christopher F. Chabris

This paper reported a divided-visual-field (DVF) experiment showing that the skilled pattern recognition that chess masters perform when seeing a chess game situation was performed faster and more accurately when the stimuli were presented briefly in the left visual field, and thus first reached the right hemisphere of the brain, than when the stimuli were presented in the right field. The sample was large for a study of highly skilled performers (16 chess masters), but we analyzed the data in many different ways and reported the result that was most favorable. Most critically, we tried different rules for removing outlier trials, and picked one that was uncommon but led to results consistent with our hypothesis. Nowadays I would analyze this type of data using more justifiable rules, and pre-register the rules I was planning to use (among other things), to avoid this problem. For these reasons I no longer think that the results provide sufficient support for the claims that the right hemisphere is more important than the left for chess expertise and for skilled visual pattern recognition. These claims may be true, but not because of our experiment.

Two other relevant things happened with this paper. First, we submitted a manuscript describing two related experiments. We were asked to remove the original Experiment 1 because the *p* value for the critical hypothesis test was below .10 but not below .05. We complied with this request. We were also asked by one reviewer to run approximately ten additional analyses of the data. We did not comply with this—instead, we wrote to the editor and explained that doing so many different analyses of the same data set would invalidate the *p* values. The editor agreed. This is evidence that the dangers of multiple testing were not exactly unknown as far back as the early 1990s. The sacrificed Experiment 1 became a chapter of my Ph.D. thesis. I tried to replicate it several years later, but I could not recruit enough chess master participants. Having also lost some faith in the DVF methodology, I put that data in the "file drawer" for good.

Statement on Fisher et al. (2015) by Ben Jones and Lisa M. DeBruine

The paper reported that women's preferences for wearing makeup that was rated by other people as being particularly attractive were stronger in test sessions where salivary testosterone was high than in test sessions where salivary testosterone was relatively low. Not long after publication, we were contacted by a colleague who had planned to use the open data and analysis code from our paper for a workshop on mixed effect models. They expressed some concerns about how our main analysis had been set up. Their main concern was that our model did not include random slopes for key within-subject variables (makeup attractiveness and testosterone). Having looked into this issue over a couple of days, we agreed that not including random slopes typically increases false positive rates and that, in the case of our study, the key effect for our interpretation was no longer significant. To minimise misleading other researchers, we contacted the journal immediately and asked to retract the paper. While this was clearly an unfortunate situation, it highlights the importance of open data and analysis code for allowing mistakes to be quickly recognised and the scientific record corrected accordingly.

Statement on Heyman, Van Rensbergen, Storms, Hutchison, and De Deyne (2015) by Tom Heyman

The goal of the study was to assess whether the processes that presumably underlie semantic priming effects, are automatic in the sense that they are capacity-free. For instance, one of the most well-known mechanisms is spreading activation, which entails that the prime (e.g., cat) pre-activates related concepts (e.g., dog), thus resulting in a head start. In order to disentangle prospective processes, those initiated upon presentation of the prime like spreading activation, from retrospective processes, those initiated upon presentation of the target, three different types of stimuli were selected. Based on previously gathered word association data, we used symmetrically associated word pairs (e.g., cat–dog; both prime and

target elicit one another) as well as asymmetrically associated pairs in the forward direction (e.g., panda–bear; the prime elicits the target, but not vice versa) and in the backward direction (e.g., bear–panda; the target elicits the prime, but not vice versa). However, I now believe that this manipulation was not successful in teasing apart prospective and retrospective processes. Critically, the three types of stimuli do not solely differ in terms of their presumed prime–target association. That is, I overlooked a number of confounding variables, for one because a priori matching attempts did not take regression effects into account (see supplementary statement for more details). Unfortunately, this undercuts the validity of the study's central claim.

Statement on Lucas and Diener (2001) by Richard E. Lucas

The paper reported three studies that examined the types of situations that extraverts enjoy. Our goal was to assess whether—as intuition and some models of personality might suggest—extraverts are defined by their enjoyment of social situations, or whether extraverts are actually more responsive to the pleasantness of situations regardless of whether these are social. We concluded that extraversion correlated more strongly with ratings of pleasant situations than unpleasant situations, but not more strongly with social situations than nonsocial situations once pleasantness was taken into account. There are two primary reasons why I have lost confidence in this result. First, the sample sizes are simply too small for the effect sizes one should expect (Schönbrodt & Perugini, 2013). I do not remember how our sample size decisions were made, and the sample sizes vary substantially across studies even though the design was essentially the same. This is especially important given that one important effect from the third and largest study would not have been significant with the sample sizes used in Studies 1 and 2. We did report an internal meta-analysis; but I have become convinced that these procedures cannot correct for other problematic research practices (Vosgerau, Simonsohn, Nelson, & Simmons, 2019). Second, many participants

were excluded from our final analyses. Two participants were excluded because they were outliers who strongly affected the results. We were transparent about this and reported analyses with and without these outliers. However, the results with the outliers included do not support our hypothesis. We also excluded a second group because their results seemed to indicate that they had misinterpreted the instructions. I still find our explanation compelling, and it may indeed be correct. However, I believe that the appropriate step would be to rerun the study with new procedures that could prevent this misunderstanding. Because we would never have been motivated to look for signs that participants misunderstood the instructions if the results had turned out the way we wanted in the first place, this is an additional researcher degree of freedom that can lead to unreplicable results.

Statement on Schmukle, Liesenfeld, Back, and Egloff (2007) by Stefan C. Schmukle

The original main finding was that the implicit gender self-concept measured with the Implicit Association Test (IAT) significantly correlated with 2D:4D ratios for men ($r = .36$, $p = .02$), but not for women. We used two different versions of a gender IAT in this study (one with pictures and one with words as gender-specific stimuli; $r = .46$), and we had two different 2D:4D measures (the first measure was based on directly measuring the finger lengths using a caliper, the second was based on measuring the scans of the hands; $r = .83$). The correlation between IAT and 2D:4D was, however, only significant for the combination of picture IAT and 2D:4D scan measure, but insignificant for other combinations of IAT and 2D:4D measures. When I was writing the manuscript, I thought that the pattern of results made sense, because a) the literature suggested that for an IAT pictures were better suited as stimuli than words, and because b) I assumed that the scan measures should lead to better results for psychometric reasons (as measurements were averaged across two raters). Accordingly, I only reported the results for the combination of picture IAT and 2D:4D scan measure in the article (for all results see the long version of the Loss-of-Confidence

statement). In the meantime, I have lost confidence in this finding and I now think that the positive association between the gender IAT and 2D:4D is very likely a false-positive result, because I should have corrected the *p* value for multiple testing.

Statement on Silberzahn and Uhlmann (2013) by Raphael Silberzahn and Eric Uhlmann

In 2013 we published a paper providing evidence that the meaning of a person's name might affect her career outcomes. In a large archival dataset with over 200,000 observations, we found that German professionals with noble-sounding last names such as Kaiser ("emperor"), König ("king"), and Fürst ("prince") were more often found as managers compared to German people with common, ordinary last names such as Koch ("cook") or Bauer ("farmer"). We applied what we believed to be a solid statistical approach, using generalized estimating equations first and during the review process applied hierarchical linear modelling and controlled for various potential third variables, including linear controls for name frequency. A post-publication re-analysis by Uri Simonsohn using an expanded version of our dataset identified a curvilinear name-frequency confound in the data, whereas we had used only linear controls. Applying the improved matched-names analysis to the larger dataset conclusively overturned the original paper's conclusions. Germans with noble and non-noble names are equally well represented in managerial positions. We subsequently co-authored a collaborative commentary (Silberzahn et al., 2014) reporting the new results. This experience inspired us to pursue our line of work on crowdsourcing data analysis, in which the same dataset is distributed to many different analysts to test the same hypothesis and the effect size estimates are compared (Silberzahn et al., 2018; Silberzahn & Uhlmann, 2015).

Statement on Smith and Zentall (2016) by Thomas R. Zentall

We have found, paradoxically, that pigeons are indifferent between a signaled 50% reinforcement alternative (leading half of the time to a stimulus that signals 100% reinforcement and otherwise to a stimulus that signals 0% reinforcement) over a guaranteed 100% reinforcement alternative. We concluded that the value of the signal for reinforcement (100% in both cases) determines choice and curiously, the signal for the absence of reinforcement has no negative value. More recently, however, using a similar design but involving extended training, we found that there was actually a significant preference for the 50% signaled reinforcement alternative over the 100% reinforcement alternative (Case & Zentall, 2018). This finding required that we acknowledge that there is an additional mechanism involved, namely the contrast between what was expected and what was obtained (positive contrast). In the case of the 50% reinforcement alternative 50% reinforcement was expected but on half of the trials a signal indicated that 100% reinforcement would be obtained ("elation," analogous to the emotion felt by a gambler who hits the jackpot). Choice of the 100% reinforcement alternative comes with an expectation of 100% reinforcement and as 100% reinforcement is obtained there no positive contrast and no elation. The recognition of our error in not acknowledging the positive contrast effect has led to a better understanding of the motivation that gamblers have, to gamble in the face of repeated losses and occasional wins.

Statement on Strand, Brown, and Barbour (2018) by Julia Strand

The paper reported that when participants listened to spoken words in noise, the cognitive resources necessary to understand the speech (referred to as “listening effort”) were reduced when the speech was accompanied by dynamic visual stimulus—a circle that modulated with the amplitude of the speech. When attempting to replicate and extend that work, I discovered an error in the original stimulus presentation program that was responsible

for the observed effect. The listening effort task we used was response time based, so the critical comparison was participant response times in conditions with and without the visual stimulus. There was an unintentional delay set in the timer of the condition without the visual stimulus, leading to artificially slowed response times in that condition. We contacted the journal and they invited us to submit a replacement article. Given that the timing delay affected every observation for one condition in a systematic way, it was straightforward to reanalyze the data and present the results as they would have been without the error. The original paper was not retracted, but now links to the new paper (Strand, Brown, & Barbour, 2020) that presents the corrected results.

Statement on Vazire (2010) by Simine Vazire

In this article, I suggested a model in which self-reports are more accurate than peer reports for traits that are low in observability and low in evaluativeness, whereas peer reports are more accurate than self-reports for traits that are high in observability and high in evaluativeness. The main issue was that I ran many more analyses than I reported, and I cherry-picked which results to report. This is basically p-hacking but since most of my results were not statistically significant, I did not quite successfully p-hack by the strict definition. Still, I cherry-picked the results that made the contrast between self- and peer-accuracy the most striking, and that fit with the story about evaluativeness and observability. That story was post hoc and chosen after I had seen the pattern of results.

Statement on Willén and Strömwall (2012) by Rebecca M. Willén

This study evaluated the criteria used by Swedish courts for assessing credibility of plaintiffs' accounts. The main reasons for my loss of confidence in the results reported are listed below.

1. The main coder (myself) was not blind to the veracity of the statements. In addition, the main coder had also conducted the interviews, which means that she might have been influenced by the memory of non-verbal cues which were not supposed to have influenced the codings. The second coder was blind, and did indeed come to different conclusions in his codings. These differences may have been a consequence of the conditions and non-verbal cues being known to the main coder, and this possibility remained undisclosed in the article.

2. All four hypotheses described as confirmatory in the introduction of the paper were in fact not formalised until after the data had been collected. It could be argued that the first three hypotheses were "obvious" and thereby implicitly already decided upon. The fourth hypothesis, however, was far from obvious and it was the result of exploratory analyses made by myself.

3. No gender differences were predicted and gender was never planned to be analysed at all. The gender findings are thus the result of exploratory analyses. This fact is however never made very explicit; instead are these (unexpected) results highlighted even in the abstract.

That said, I do think there is reason to believe that one particular main finding is worth trying to replicate: "False and truthful confessions by 30 offenders were analysed, and few significant effects were obtained." That is, true and false statements by criminally experienced offenders might be more difficult to distinguish than true and false statements provided by the typical participants in deception and interrogation research (i.e., undergraduates without criminal experience).

Statement on Witt and Proffitt (2008) by Jessica K. Witt

The paper reported that squeezing a rubber ball interferes with the processes necessary for the perceiver's ability to reach to a target to affect perceived distance to the target (Experiment 3a). Participants judged the distance to targets that were beyond the reach of the arm, then picked up a conductor's baton and reached to them. One group of participants applied a constant, firm pressure on a rubber ball while making their distance judgments, whereas another group did not. The analysis reported in the paper used an incorrectly specified model. Specifically, we calculated the mean estimated distance for each participant at each distance for a total of 10 estimates per participant, then analyzed these means as if they were independent observations. This inflated the degrees of freedom, which resulted in lower *p* values. When the data are analyzed correctly, the critical effect of ball squeeze on estimated distance is not significant, $F(1, 14) = 2.31$, $p = .151$, $\eta_p^2 = .14$. The mean difference between the two conditions was 3.5 cm, and the standard deviation for the mean estimate across all participants was 4.8 cm. To achieve 80% power to find an effect of $d = .73$, we would have needed 33 participants per condition. Instead, we only had 8 participants per condition. Thus, we do not have enough data to make claims about whether squeezing the ball had an effect. This lack of sufficient data was confirmed by calculating the Bayes factor for a two-sample t-test comparing mean estimated distance across the two squeeze conditions. The Bayes factor was 0.90, which means the data support neither the null or the alternative hypothesis. Incorrect model specification and subsequent discovery of lack of sufficient power also applies to Experiments 1, 2, and 3b. Experiment 4 is believed to have been analyzed correctly based on the reported degrees of freedom, but those data have been lost and therefore cannot be confirmed.

Statement on Yarkoni, Braver, Gray, and Green (2005) by Tal Yarkoni

This study used a dynamic decision-making task to investigate the neural correlates of temporally-extended decision-making. The central claim was that activation in areas of right lateral prefrontal cortex (PFC) strongly and selectively predicted choice behavior in two different conditions; peak between-subject brain-behavior correlations were around r = .75. I now think most of the conclusions drawn in this article were absurd on their face. My understanding of statistics has improved a bit since writing the paper, and it's now abundantly clear to me that (a) I p-hacked to a considerable degree (e.g., the choice of cluster thresholds was essentially arbitrary), and (b) because of the "winner's curse", statistically significant effect sizes from underpowered studies cannot be taken at face value (see Yarkoni, 2009). Beyond these methodological problems, I also now think the kinds of theoretical explanations I proposed in the paper were ludicrous in their simplicity and naivete—so the results would have told us essentially nothing even if they were statistically sound (see Meehl, 1967, 1990).

Discussion of the Loss-of-Confidence Statements

The studies for which we received statements spanned a wide range of psychological domains (stereotypes, working memory, auditory perception, visual cognition, face perception, personality and well-being, biologically-driven individual differences, social cognition, decision-making in non-human animals, deception detection) and employed a diverse range of methods (cognitive tasks, implicit and explicit individual differences measures, archival data analyses, semi-structured interviews, functional MRI), demonstrating the broad relevance of our project. Overall, the respective original articles had been cited 1,559 times as of April 27, 2020 according to Google Scholar, but the number of citations varied widely, from 9 to 740. The reasons given for the submitters' loss of confidence also

varied widely, with some statements providing multiple reasons. Broadly speaking, however, we can group the explanations into three general categories:

1. *Methodological Error.* Five of the statements reported methodological errors in the broadest sense. In three instances, submitters (Jones & DeBruine; Silberzahn & Uhlmann; Witt) lost confidence in their findings upon realizing that their key results stemmed from misspecified statistical models. In those three cases, the submitters discovered, post-publication, that a more appropriate model specification resulted in the key effect becoming statistically non-significant. In another instance, Carlsson reported that, upon reconsideration, two studies included in his article actually tested different hypotheses–a reanalysis testing the same hypotheses in Study 2 actually failed to fully support the findings from Study 1. Lastly, Strand lost confidence when she found out that a programming error invalidated her findings.

2. *Invalid inference.* Four of the statements reported invalid inferences in the broadest sense. In two cases (Heyman and Yarkoni), the submitters attributed their loss of confidence to problems of validity—that is, to a discrepancy between what the reported results actually showed (a statistically significant effect of some manipulation or measure) and what the paper claimed to show (a general relationship between two latent constructs). In a similar vein, Zentall lost confidence in a conclusion when a follow-up experiment revealed that an extension of the experimental procedures suggested that the original mechanism was not sufficient to account for the phenomenon. While the latter Loss-of-Confidence statement might be closest to normative assumptions about how science advances–new empirical insights lead to a revision of past conclusions–it also raises interesting questions: At what point should researchers lose confidence in a methodological decision made in one

study based on the results of other studies that are, in principle, also fallible?

3. *P-hacking.* Seven of the statements (Carlsson, Chabris, Lucas, Yarkoni, Schmukle, Vazire, and Willén) reported some form of *p*-hacking—i.e., failing to properly account for researcher degrees of freedom when conducting or reporting the analyses. We hasten to emphasize that our usage of "*p*-hacking" here does not imply any willful attempt to mislead. Indeed, some of the submitters noted that the problems in question stemmed from their poor (at the time) understanding of relevant statistical considerations. The statement by Lucas also highlights how subtle researcher degrees of freedom can affect analyses: Even though the justification for a specific exclusion criterion still seems compelling, the researcher would not have been motivated to double-check data points if the desired results had emerged in the initial analysis.

Results and Discussion of the Anonymous Online Survey

Overall, 316 scientists completed the survey. Most (93%) reported being affiliated with a university or a research institute, and all career stages from graduate students to tenured professors were represented. We did not limit the survey to particular fields of research but asked respondents to indicate their department (if applicable); 43% did not report a department, 37% worked at a psychology department, and the remaining respondents were distributed over a broad range of fields (e.g., business, economics, medicine). Almost all respondents reported working either in Europe (44%) or the US (47%). Figure 1 provides an overview of the survey results.

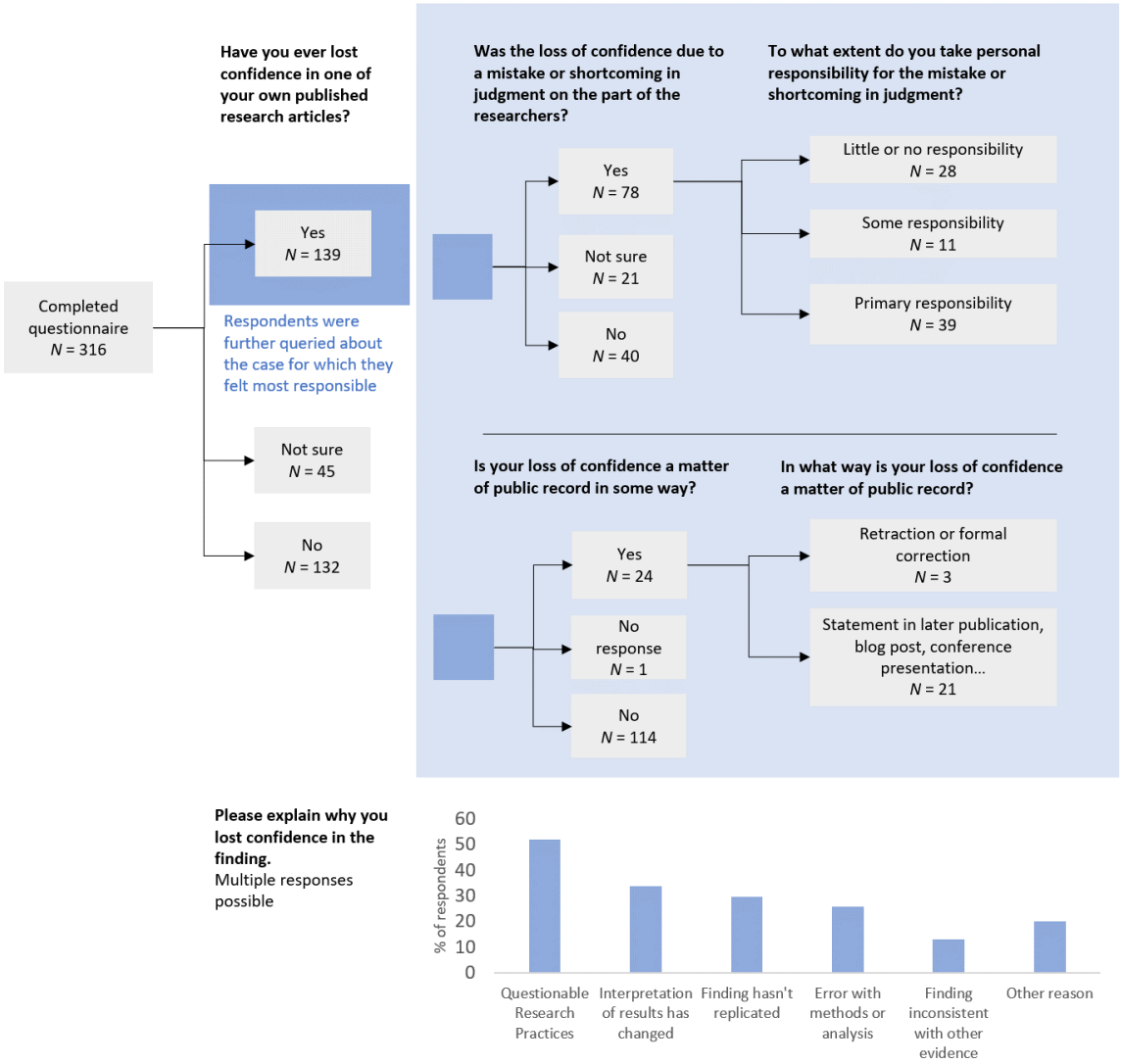

Figure 1. An overview of the findings from the Loss-of-Confidence survey.

Almost half of the respondents (44%) reported losing confidence in at least one of their findings. Another 14% were not sure whether they had lost confidence according to our definition for a variety of reasons–for example, some reported that their confidence in one of their own research articles was low to begin with; some had lost confidence in their theoretical explanation, but not in the general effect—or conversely, in the effect but not in the theory; others doubted whether their results would generalize to other contexts. Respondents who reported losing confidence were then asked to elaborate on the case for

which they felt most responsible.[4] Of the respondents who stated that they had experienced a loss of confidence, more than half (56%) said that it was due to a mistake or shortcoming in judgment on the part of the researchers, and roughly one in four (28%) took primary responsibility for the error.

Strikingly, the primary reason indicated for a loss of confidence was self-admitted questionable research practices (such as *p*-hacking and selective reporting; 52%). However, a broad variety of other reasons were also reported. The loss of confidence was a matter of public record in fewer than a fifth of the reported cases (17%), and if it was a matter of public record, the outlets primarily chosen (statement in later publication, conference presentation, social media posting) were not directly linked to the original research article. Respondents whose loss of confidence was not public reported multiple reasons for the lack of disclosure. Many felt insufficiently sure about the loss of confidence to proceed (68%). Some stated the belief that public disclosure was unnecessary, as the finding had not attracted much attention (46%); expressed concerns about hurting the feelings of co-authors (33%); or cited the lack of an appropriate venue (25%); uncertainty about how to best communicate the matter (25%); and worries about how the loss of confidence would be perceived (24%).

On the whole, these survey results suggest a nuanced view of losses of confidence. Researchers may start to question their own findings for a broad variety of reasons, and different factors may then keep them from publicly disclosing this information. Collectively,

[4] Respondents who were not sure whether they had experienced a loss of confidence could also answer the follow-up questions. However, many decided not to answer, and for those who answered, responses are hard to interpret given the broad variety of scenarios they were referring to. Thus, we decided to restrict the following analyses to respondents with an unambiguous loss of confidence.

the responses suggest that a sizeable proportion of active researchers have lost confidence in at least one of their findings—often due to a recognized error of their own commission.

It is important to note that our respondents do not constitute a representative sample of researchers. Further, estimating article-level rather than researcher-level loss of confidence requires assumptions and extrapolations.[5] Thus, caution should be exercised when interpreting the specific numerical estimates reported here. Nevertheless, one can attempt a very conservative extrapolation: over one million academic papers are currently published each year (Jinha, 2010). Supposing that at least a third of these are empirical research reports, and that even just one percent of these reports are affected, that still leaves us with thousands of articles published each year that will eventually lose the confidence of at least some of their authors—often due to known errors, yet typically without any public disclosure.

General Discussion

The Loss-of-Confidence Project raises a number of questions about how we should interpret individual self-corrections.

First, on a substantive level, how should we think about published empirical studies in cases where the authors have explicitly expressed a loss of confidence in the results? One intuitive view is that authors have no privileged authority over "their" findings, and thus such statements should have no material impact on a reader's evaluation. On the other hand, even if authors lack any privileged authority over findings they initially reported, they clearly often have privileged access to relevant information. This is particularly salient for the p-hacking

[5] In the survey, we also asked researchers to indicate in how many of their articles they had lost confidence. An analysis of these numbers suggested that respondents had collectively lost confidence in more than 10% of their publications in total; or more than 7% counting only those articles in which they had lost confidence due to an error for which they took primary responsibility. Of course, these are extrapolations based on retrospective self-reports, and we cannot assume respondents are able to give perfect estimates of the relevant quantities. For this reason, a number of our key analyses focus on the respondents' description of the one case for which they felt most responsible.

disclosures reported in the Loss-of-Confidence statements. Absent explicit statements of this kind, readers would most likely not be able to definitively identify the stated problems in the original report. In such cases, we think it is appropriate for readers to update their evaluations of the reported results to accommodate the new information.

Even in cases where a disclosure contributes no new methodological information, one might argue that the mere act of self-correction should be accorded a certain weight. Authors have presumably given greater thought to, and are more aware of, their own study's potential problems and implications than a casual reader. The original authors may also be particularly biased to evaluate their own studies favorably—so if they have nonetheless lost confidence, this might heuristically suggest that the evidence against the original finding is particularly compelling.

Second, on a meta-level, how should we think about the reception our project received? On the one hand, one could argue that the response was about as positive as could reasonably be expected. Given the unconventional nature of the project and the potentially high perceived cost of public self-correction, the project organizers (JMR, CFC, TY) were initially unsure whether the project would receive any submissions. From this perspective, even the thirteen submissions we ultimately received could be considered a clear success and a testament to the current introspective and self-critical climate in psychology.

On the other hand, the survey responses we received suggest that the kinds of errors disclosed in the statements are not rare. Approximately 12% of the 316 survey respondents reported losing confidence in at least one of their articles for reasons that matched our stringent submission criteria (i.e., due to mistakes that the respondent took personal responsibility for), and nearly half acknowledged a loss of confidence more generally.

This suggests that potentially hundreds, if not thousands, of researchers could have submitted loss-of-confidence statements, but did not do so. There are many plausible reasons for this, including not having heard of the project. However, we think that at least partially, the small number of submitted statements points to a gap between researchers' ideals and their actual behavior—that is, public self-correction is desirable in the abstract, but difficult in practice.

Fostering a Culture of Self-Correction

As we have seen, researchers report a variety of reasons for both their losses of confidence, and their hesitation to publicly disclose a change in thinking. However, we would like to suggest that there is a broader underlying factor: in the current research environment, self-correction, or even just critical reconsideration of one's past work, is often disincentivized professionally. The opportunity costs of a self-correction are high; time spent on correcting past mistakes and missteps is time that cannot be spent on new research efforts, and the resulting self-correction is less likely to be judged a genuine scientific contribution. Moreover, researchers may worry about self-correction potentially backfiring. Corrections that focus on specific elements from an earlier study might be perceived as undermining the value of the study as a whole, including parts that are in fact unaffected by the error. Researchers might also fear that a self-correction that exposes flaws in their work will damage their reputation, and perhaps even undermine the credibility of their research record as a whole.

To tackle these obstacles to self-correction, changes to our research culture are necessary. Scientists make errors (and this statement is certainly not limited to psychological researchers, see, e.g., Eisenman, Meier, & Norris, 2014; García-Berthou & Alcaraz, 2004; Salter et al., 2014; Westra et al., 2011), and rectifying these errors is a genuine scientific contribution—whether it is done by a third party or the original authors. Scientific societies

could consider whether they want to more formally acknowledge efforts by authors to correct their own work. Confronted with researchers who publicly admit to errors, other researchers should keep in mind that willingness to admit error is not a reliable indicator of propensity to commit errors—after all, errors are frequent throughout the scientific record. On the contrary, given the potential (or perceived) costs of individual self-corrections, public admission of error could be taken as a credible signal that the issuer values the correctness of the scientific record. However, ultimately, given the ubiquity of mistakes, we believe that individual self-corrections should become a routine part of science, rather than an extraordinary occurrence.

Different Media for Self-Correction

Unfortunately, good intentions are not enough. Even when researchers are committed to public self-correction, it is often far from obvious how to proceed. Sometimes, self-correction is hindered by the inertia of journals and publishers. For example, a recent study suggested that many medical journals published correction letters only after a significant delay, if at all (Goldacre et al., 2019), and authors who tried to retract or correct their own articles after publication have encountered delays and reluctance from journals (e.g., Grens, 2015). Even without such obstacles, there is presently no standardized protocol describing what steps should be taken when a loss of confidence has occurred.

Among the participants of the loss-of-confidence project, Fisher, Hahn, DeBruine, and Jones (2015) decided to retract their article after they became aware of their misspecified model. But researchers may often be reluctant to initiate a retraction, given that retractions occur most commonly as a result of scientific misconduct (Fang et al., 2012) and are therefore often associated in the public imagination with cases of deliberate fraud. To prevent this unwelcome conflation and encourage more frequent disclosure of errors, journals could introduce a new label for retractions initiated by the original authors (e.g., “Authorial Expression of Concern” or “voluntary withdrawal”; see Alberts et al., 2015). Furthermore, an

option for authorial amendments beyond simple corrections (up to and including formal versioning of published articles) could be helpful.

Thus, it is not at all clear that widespread adoption of retractions would be an effective, fair, or appropriate approach. Willén (2018) argued that retraction of articles in which questionable practices were employed could deter researchers from being honest about their past actions. Furthermore, retracting papers because of questionable research practices (QRPs) known to be widespread (e.g., John et al., 2012) could have the unintended side effect that some researchers might naively conclude that a lack of a retraction implies a lack of QRPs. Hence, Willén (2018) suggested that all articles should be supplemented by transparent retroactive disclosure statements. In this manner, the historical research record remains intact, as information would be added rather than removed.

Preprint servers (such as PsyArXiv.com) and other online repositories already enable authors to easily disclose additional information to supplement their published articles or express their doubts. However, such information also needs to be discoverable. Established databases such as PubMed could add links to any relevant additional information provided by the authors. Curate Science (curatescience.org), a new online platform dedicated to increasing the transparency of science, is currently implementing retroactive statements that could allow researchers to disclose additional information (e.g., additional outcome measures or experimental manipulations not reported in the original article) in a straightforward, structured manner.

Another, more radical step would be to move scientific publication entirely online and make articles dynamic rather than static, such that they can be updated based on new evidence (with the previous version being archived) without any need for retraction (Nosek &

Bar-Anan, 2012). For example, the Living Reviews journal series in physics by Springer Nature allows authors to update review articles to incorporate new developments.

The right course of action once one has decided to self-correct will necessarily depend on the specifics of the situation, such as the reason for the loss of confidence; publication norms that can vary between research fields and evolve over time; and the position that the finding takes within the wider literature. For example, a simple but consequential computational error may warrant a full retraction, whereas a more complex confound may warrant a more extensive commentary. In research fields in which the published record is perceived as more definitive, a retraction may be more appropriate than in research fields in which published findings have a more tentative status. And an error in a manuscript that plays a rather minor role in the context of the wider literature may be sufficiently addressed in a corrigendum, whereas an error in a highly cited study may require a more visible medium for the self-correction to reach all relevant actors.

That said, we think that both the scientific community and the broader public would profit if additional details about the study, or the author's re-assessment of it, were always made public, and always closely linked to the original article—ideally in databases and search results as well as the publisher's website and archival copies. A cautionary tale illustrates the need for such a system: In January 2018, a major German national weekly newspaper published an article (Kara, 2018a) which uncritically cited the findings of Silberzahn and Uhlmann (2013). Once the journalist had been alerted that these findings had been corrected in Silberzahn, Simonsohn, and Uhlmann (2014), she wrote a correction to her newspaper article which was published within less than a month of the previous article (Kara, 2018b), demonstrating swift journalistic self-correction and making a strong point that any post-publication update to a scientific article should be made clearly visible to all readers of the original article.

Outlook

All of these measures *could* help to transform the cultural norms of the scientific community, bringing it closer to the ideal of self-correction. Naturally, it is hard to predict which ones will prove particularly fruitful, and changing the norms of any community is a non-trivial endeavor. However, it might be encouraging to recall that over the last few years, scientific practices in psychology have already changed dramatically (Nelson, Simmons, & Simonsohn, 2018). Hence, a shift towards a culture of self-correction may not be completely unrealistic, and psychology with its increasing focus on openness may even serve as a role model for other fields of research to transform their practices.

Finally, it is quite possible that fears about negative reputational consequences are exaggerated. It is unclear whether, and to what extent, self-retractions actually damage researchers' reputations (Bishop, 2018). Recent acts of self-correction such as those by Carney (2016), which inspired our efforts in this project, Silberzahn and Uhlmann (Silberzahn et al., 2014), Inzlicht (2016), Willén (2018), and Gervais (2017) have received positive reactions from within the psychological community. They remind us that science can advance at a faster pace than one funeral at a time.

References

Alberts, B., Cicerone, R. J., Fienberg, S. E., Kamb, A., McNutt, M., Nerem, R. M., ...Jamieson, K. H. (2015). Self-correction in science at work. *Science*, *348*(6242), 1420–1422.

Allchin, D. (2015). Correcting the "self-correcting" mythos of science. *Filosofia e História da Biologia*, *10*(1), 19–35.

Bishop, D. V. M. (2018). Fallibility in science: responding to errors in the work of oneself and others. *Advances in Methods and Practices in Psychological Science*, *1*(3), 432–438.

Bohannon, J. (2014). Replication effort provokes praise—and 'bullying' charges. *Science*, *344*(6186), 788–789.

Carlsson, R. & Björklund, F. (2010). Implicit stereotype content: mixed stereotypes can be measured with the implicit association test. *Soc. Psychol. 41*(4), 213–222.

Carney, D. (2016). My position on power poses. https://faculty.haas.berkeley.edu/dana_carney/pdf_my\%20position\%20on\%20power\%20poses.pdf. Accessed: 2018-11-27.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science, 2*, 115–144. doi:10.1177/2515245919847196

Case, J. P. & Zentall, T. R. (2018). Suboptimal choice in pigeons: does the predictive value of the conditioned reinforcer alone determine choice? *Behav. Processes*, *157*, 320–326.

Chabris, C. F. & Hamilton, S. E. (1992). Hemispheric specialization for skilled perceptual organization by chessmasters. *Neuropsychologia*, *30*(1), 47–57.

Chang, A. C., Li, P. et al. (2018). Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Often Not". *Critical Finance Review*, *7*.

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behav. Brain Sci. 41*, e124.

Eisenman, I., Meier, W. N., & Norris, J. R. (2014). A spurious jump in the satellite record: Has Antarctic sea ice expansion been overestimated? *The Cryosphere*, *8*(4), 1289–1296.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America, 109*(42), 17028–17033.

Fisher, C. I., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2015). Women's preference for attractive makeup tracks changes in their salivary testosterone. *Psychol. Sci. 26*, 1958–1964.

García-Berthou, E. & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Med. Res. Methodol. 4*, 13.

Gervais, W. (2017). Post publication peer review. http://willgervais.com/blog/2017/3/2/post-publication-peer-review. Accessed: 2018-11-27.

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., ... Mahtani, K. R. (2019). COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118.

Grens, K. (2015). Self correction. What to do when you realize your publication is fatally flawed. https://www.the-scientist.com/careers/self-correction-34431. Accessed: 2019-2-21.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... Frank, M. C. (2018, August). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal cognition. *R Soc Open Sci*, *5*(8), 180448.

Heyman, T., Van Rensbergen, B., Storms, G., Hutchison, K. A., & De Deyne, S. (2015). The influence of working memory load on semantic priming. *J. Exp. Psychol. Learn. Mem. Cogn. 41*(3), 911–920.

Inzlicht, M. (2916, February 29). Reckoning with the past [Blog post]. Retrieved from http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspect. Psychol. Sci. 7*(6), 645–654.

Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learn. Publ. 23*(3), 258–263.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci. 23*(5), 524–532.

Kara, S. (2018a). Bewusstsein: Wir wissen nicht, was wir tun. https://www.zeit.de/2018/05/unterbewusstsein-psychologie-forschung-manipulation. Accessed: 2018-11-27.

Kara, S. (2018b). Forschung: Wer Karriere machen will, hat es mit einem noblen Namen leichter. “Koch” sollte er nicht heißen, besser “König”*. https://www.zeit.de/2018/08/forschung-fehler-korrektur-wissenschaft. Accessed: 2018-11-27.

Lucas, R. E. & Diener, E. (2001). Understanding extraverts’ enjoyment of social situations: the importance of pleasantness. *J. Pers. Soc. Psychol. 81*(2), 343–356.

Mayo-Wilson, C., Zollman, K. J. S., & Danks, D. (2011). The independence thesis: When individual and social epistemology diverge. *Philos. Sci. 78*(4), 653–677.

Meehl, P. E. (1967). Theory-Testing in psychology and physics: a methodological paradox. *Philos. Sci. 34*(2), 103–115.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychol. Rep. 66*(1), 195–244.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annu. Rev. Psychol. 69*, 511–534.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ...Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A. & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychol. Inq. 23*(3), 217–243.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods*, *48*(4), 1205–1226.

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *J. Informetr. 12*(3), 656–678.

Planck, M. (1948). *Wissenschaftliche Selbstbiographie. Mit einem Bildnis von der von Max von Laue gehaltenen Traueransprache*. Leipzig: Johann Ambrosius Barth Verlag.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., ... Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol. 12*, 87.

Schmukle, S. C., Liesenfeld, S., Back, M. D., & Egloff, B. (2007). Second to fourth digit ratios and the implicit gender self-concept. *Pers. Individ. Dif. 43*(5), 1267–1277.

Schönbrodt, F. D. & Perugini, M. (2013). At what sample size do correlations stabilize? *J. Res. Pers. 47*(5), 609–612.

Silberzahn, R., Simonsohn, U., & Uhlmann, E. L. (2014). Matched-names analysis reveals no evidence of name-meaning effects: A collaborative commentary on Silberzahn and Uhlmann (2013). *Psychol. Sci. 25*(7), 1504–1505.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.

Silberzahn, R. & Uhlmann, E. L. (2013). It pays to be Herr Kaiser: Germans with noble-sounding surnames more often work as managers than as employees. *Psychol. Sci. 24*(12), 2437–2444.

Silberzahn, R. & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189–191.

Sismondo, S. (2010). *An Introduction to Science and Technology Studies*. Chichester: Willey-Blackwell, 1-11.

Smith, A. P. & Zentall, T. R. (2016). Suboptimal choice in pigeons: choice is primarily based on the value of the conditioned reinforcer rather than overall reinforcement rate. *J Exp Psychol Anim Learn Cogn*, *42*(2), 212–220.

Strand, J. F., Brown, V. A., & Barbour, D. L. (2018). Talking points: A modulating circle reduces listening effort without improving speech recognition. *Psychon. Bull. Rev. 26*(1), 291–297.

Strand, J. F., Brown, V. A., & Barbour, D. L. (2020). Talking points: A modulating circle increases listening effort without improving speech recognition in young adults. *Psychon. Bull. Rev*

Vazire, S. (2010). Who knows what about a person? the self–other knowledge asymmetry (SOKA) model. *J. Pers. Soc. Psychol. 98*(2), 281.

Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology: General, 148*(9), 1628.

Westra, H.-J., Jansen, R. C., Fehrmann, R. S. N., te Meerman, G. J., van Heel, D., Wijmenga, C., & Franke, L. (2011). MixupMapper: Correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics*, *27*(15), 2104–2111.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One*, *6*(11), e26828.

Willén, R. M. (2018). Retroactive disclosure statements make the past more useful. https://medium.com/@IGDORE/retroactive-disclosure-statements-make-the-pastmore-useful-c1b2e73f4bae. Accessed: 2018-11-27. Medium.

Willén, R. M. & Strömwall, L. A. (2012). Offenders' lies and truths: an evaluation of the Supreme Court of Sweden's criteria for credibility assessment. *Psychol. Crime Law*, *18*(8), 745–758.

Witt, J. K. & Proffitt, D. R. (2008). Action-specific influences on distance perception: a role for motor simulation. *J. Exp. Psychol. Hum. Percept. Perform. 34*(6), 1479–1492.

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspect. Psychol. Sci. 4*(3), 294–298.

Yarkoni, T., Braver, T. S., Gray, J. R., & Green, L. (2005). Prefrontal brain activity predicts temporally extended decision-making behavior. *J. Exp. Anal. Behav. 84*(3), 537–5

Table S1

Questions Included in the Loss-of-Confidence Survey

Question	Response options
Have you ever lost confidence in one of your own published research articles? In other words, have you become convinced that a primary effect or conclusion you initially reported does not accurately describe the world as you currently understand it?	Yes, I have lost confidence in one or more of my own findings in this way No, I have not lost confidence in one or more of my own findings in this way Not sure [text box provided for elaboration]
How many total research articles have you published, regardless of your position on the authorship list? Please feel free to give an approximate number.	[Range of options from "0" to "500 or more"]
What total number of your published research articles have you largely or completely lost confidence in for any reason? This could be 0 papers, 1 paper, 2 papers, and so on. If you do not think that you can give a precise number of articles, please provide an estimated number.	[Range of options from "0" to "500 or more"]
What total number of your published research articles have you largely or completely lost confidence in, specifically due to a mistake or shortcoming in judgment for which you take primary responsibility? This could be 0 papers, 1 paper, 2 papers, and so on. If you do not think that you can give a precise number, please provide an estimated number.	[Range of options from "0" to "500 or more"]
For the remaining questions, please focus on <u>one</u> research article in which you have lost confidence. If you have lost confidence in multiple papers, please answer based on the <u>single</u> paper for which you feel the most personally responsible for the loss of confidence.	
Please explain why you lost confidence in the finding.	The finding hasn't replicated The finding is inconsistent with other reported evidence My interpretation/understanding of the original results has changed I think that the finding is unlikely to replicate because of Questionable Research Practices (e.g., p-hacking, HARKing or hypothesizing-after-the-results-are-known, selective reporting) [text box provided for elaboration] There was an error with the methods or analysis procedures in the original study [text box provided for elaboration] Other reason [text box provided for elaboration]
Was the loss of confidence due to a mistake or shortcoming in judgment on the part of the researchers?	Yes No Not sure [text box provided for elaboration]
If you answered yes to Question 6 above, to what extent do you take personal responsibility for the mistake or shortcoming in judgment?	Little or no responsibility Primary or sole responsibility Some responsibility

	Not sure [text box provided for elaboration]
Is your loss of confidence a matter of public record in some way (e.g., formal correction, statement in a later paper noting the finding is not supported)?	Yes No
If you responded “yes” to Question 8 above: In what way is your loss of confidence a matter of public record? (Please select all that apply).	Retraction of the paper Formal correction to the paper Statement in a later publication that the finding or claim is not empirically supported Conference presentation Social media posting (e.g., blog post, Twitter, Facebook) In some other way [text box provided for elaboration]
If your loss of confidence is not a matter of public record please indicate why not.	Worry about how the disclosure would be perceived Don't want to hurt co-authors' feelings Feel that it's not my place to disclose any issues Don't know of a suitable venue for such a disclosure Haven't had the time given other higher priorities I am not sufficiently sure about my loss of confidence to proceed Do not think it's appropriate to second-guess one's previous work publicly Do not think it's necessary because the finding has not attracted much attention I think my personal beliefs are irrelevant/provide no added value I do not think it's necessary because other researchers corrected the scientific record (e.g., independent non-replication or published re-analysis) It has/had never occurred to me to make it public I’m not sure what the appropriate way would be to communicate my loss of confidence [text box provided for elaboration] Other reason [text box provided for elaboration]
Demographic variables: What is your age? What is your gender? What region of the world do you currently work primarily in? Are you currently affiliated with a university or research institute? If working at a university, what department are you affiliated with? If working at a university, what is your job rank?	