Where Are The Self-Correcting Mechanisms In Science?

Simine Vazire[1] & Alex O. Holcombe[2]

[1] Melbourne School of Psychological Sciences, University of Melbourne

[2] School of Psychology, The University of Sydney

ABSTRACT

It is often said that science is self-correcting, but the replication crisis suggests that, at least in some fields, self-correction mechanisms have fallen short of what we might hope for. How can we know whether a particular scientific field has effective self-correction mechanisms, that is, whether its findings are credible? The usual processes that supposedly provide mechanisms for scientific self-correction – mainly peer review and disciplinary committees – have been inadequate. We argue for more verifiable indicators of a field's commitment to self-correction. These include transparency, which is already a target of many reform efforts, and critical appraisal, which has received less attention. Only by obtaining Measurements of Observable Self-Correction (MOSCs) can we begin to evaluate the claim that "science is self-correcting." We expect the validity of this claim to vary across fields and subfields, and suggest that some fields, such as psychology and biomedicine, fall far short of an appropriate level of transparency and, especially, critical appraisal. Fields without robust, verifiable mechanisms for transparency and critical appraisal cannot reasonably be said to be self-correcting, and thus do not warrant the credibility often imputed to science as a whole.

Where Are The Self-Correcting Mechanisms In Science?

"When we observe scientists, we find that they have developed a variety of practices for vetting knowledge – for identifying problems in their theories and experiments and attempting to correct them."

Oreskes (2019, p. 64)

"In reality, mechanisms to correct bad science are slow, unreliably enforced, capricious, run with only the barest nod towards formal policy, confer no reward and sometimes punitive elements for a complainant who might use them"

Heathers (2019)

The assertion that "science is self-correcting" is frequently made when people argue that science is trustworthy. When serious scientific errors come to light in science, people explain that such errors are part of the normal and healthy process of science: science sometimes fumbles, but it can be counted on to correct itself. Indeed, even the finding that the most prestigious journals in psychology seem to routinely produce many false positives can be explained away as "science correcting itself". In the introduction to the 2016 volume of the *Annual Review of Psychology*, published just six months after the Reproducibility Project: Psychology results (Open Science Collaboration, 2015) were published, in which over half of findings published in top psychology journals could not be replicated, Fiske, Schacter, and Taylor (2016) concluded that there is no crisis:

> First came a few tragic and well-publicized frauds; fortunately, they are rare—though never absent from science conducted by humans—and they were caught. Now the main concern is some well-publicized failures to replicate [...]. All this is normal science, not crisis. A replication failure is not a scientific problem; it is an opportunity to find limiting conditions and contextual effects. Of course studies don't always replicate.

A few years later, in a report titled *Reproducibility and Replicability in Science,* the National Academy of Sciences (2019) study committee[1] suggested that "The advent of new scientific knowledge that displaces or reframes previous knowledge should not be interpreted as weakness in science". While this could perhaps be justified, this (and the rest of the NAS report) leaves open the question of what a breakdown in science *would* look like. How can we tell the difference between an efficient self-correcting system and one that is making unforced errors and needs fixing?

Many appeals to a self-correcting system seem vacuous in that details regarding the effectiveness of the system and the underlying processes are rarely, if ever, given. Case in point is a blog post entitled “Why science is self-correcting” by the cognitive scientist Art Markman (Markman, 2010). Markman wrote, “There's no point in scientific misconduct; it is always found.” He went on to claim that “The field is able to separate the good results from the bad fairly quickly”, with a lack of detail that seems common and that we find frustrating.

How can we know whether self-correction in science, or in a particular science, is in fact efficient and reliable? The degree to which false claims and findings are identified and speedily corrected likely varies across scientific disciplines and subdisciplines. Knowing how self-correcting a particular science is is crucial to understanding how much we ought to trust that field - or how much credibility to ascribe to the claims made by that scientific community. Thus, claiming that all of science is self-correcting, if only parts of it have reasonably effective self-correction mechanisms, may eventually undermine the credibility of even the most robust sciences, as it links their reputation to that of other sciences with lower credibility. For this reason, it is important that the public be able to appropriately calibrate their trust in any given science, and to do so they must be able to evaluate the degree to which any given scientific community has effective mechanisms for self-correction. To do this, we must make the self-correcting mechanisms of science visible.

Individual scientists do not typically self-correct on their own

We have argued above that public trust in science is likely linked to the perception that science has efficient self-correcting mechanisms - and that trust in any given scientific field is warranted only to the extent that we can identify robust self-correction mechanisms in that field. Where should we look for these self-correcting mechanisms? Across the board, both scientists and non-scientists seem to agree that the self-correcting mechanisms in science, and therefore the trustworthiness of science, do not come from individual scientists’ honesty and trustworthiness.

According to a recent Pew survey of US adults (Funk, Hefferon, Kennedy, & Johnson, 2019), in 2019, 86% of poll respondents said they have at least “a fair amount” of confidence in scientists to act in the public interest, which is more than for business leaders, the news media, and elected officials. However, only 11-16% of respondents in the same survey agree that scientists “admit and take responsibility for their mistakes” (when asked specifically about medical, environmental, and nutrition scientists). This cynicism seems to be shared by scientists themselves, who appear to have a fairly negative view of their fellow scientists’ integrity. In a survey of early- and mid-career scientists with funding from the US National Institutes of Health, 61-76% of

scientists reported that other scientists fail to adhere to Mertonian norms such as disinterestedness and openness (Anderson, Martinson, & De Vries, 2007).

While it is difficult or impossible to estimate the prevalence of dishonesty or incompetence in science, it is clear that the number of papers with errors exceeds, likely by multiple orders of magnitude, the number that are corrected, either by retraction or formal correction. For example, by one estimate, approximately half of psychology journal articles published in eight well-respected journals between 1985 and 2013 include at least one report of a statistic and associated degrees of freedom that is inconsistent with the reported *p*-value. In about one in eight papers, this discrepancy was deemed to be a "gross error", in that while the reported *p*-value was significant, the recalculated *p*-value based on the reported degrees of freedom and test statistic was not, or vice versa (Nuijten et al., 2016). Similarly, in an analysis of genetics articles published between 2005 and 2015, a team of researchers found that about 1 in 5 articles had Microsoft Excel files with gene lists that are affected by gene name errors (e.g., gene names that Excel automatically converts to dates; Ziemann et al., 2016).

So why do members of the public, and presumably scientists themselves, continue to trust science despite this justified cynicism about individual scientists' commitment to self-correction? Many historians, sociologists, and philosophers of science have suggested that trust is warranted despite individual scientists' fallibility because self-correction, and the production of knowledge, is a social process rather than the result of individual scientists' humility or rationality (Campbell, 1988; Haack, 2003; Kuhn, 1962; Longino, 1990; Oreskes, 2019; Sztompka, 2007).

According to this view, healthy scientific communities are structured in ways that encourage, or even ensure, self-correction at the group level. The opening quote by Oreskes illustrates a common refrain, echoed in statements by other scholars of science and by scientists themselves, including methodologist Donald Campbell:

> "The resulting dependability of reports (such as it is, and I judge it usually to be high in the physical sciences) comes from a social process rather than from dependence upon the honesty and competence of any single experimenter. Somehow in the social system of science a systematic norm of distrust (Merton's [1973] "organized skepticism") combined with ambitiousness leads people to monitor each other for improved validity. Organized distrust produces trustworthy reports." (Campbell, 1988, p. 324)

The word "somehow" in the above passage glosses over a number of issues. How does the social system of science achieve this? How do we know that this is so?

Others have made similar claims with equally little evidence - it seems that the obvious success of science (or at least of the physical sciences) is taken as incontrovertible evidence that science must have developed efficient self-corrective mechanisms somehow. This can also be found in the more informal discourse of journal editors who ask that we "let the self-correcting mechanisms of science take their course" (McConnell, 2020), as if this happens magically.

In some cases, scholars appealing to social processes of verification and correction do articulate what those processes look like:

> "the existence of strong and rigid self-policing, and self-controlling mechanisms built into scientific communities preventing access by incompetent or dishonest people, allows it to be taken for granted that those who pass through the checks of gatekeepers can be trusted (of course, again, this requires 'auxiliary trust' in the integrity and reliability of gatekeepers). More concretely this involves the peer review of publications or grant applications, standardized procedures for obtaining academic degrees and titles, doctoral committees, tenure committees, complex and collective procedures for awarding scientific prizes, juries, disciplinary committees guarding the ethos of science: in extreme cases even courts of law." (Sztompka, 2007, p. 217)

If pressed, most scholars appealing to these social mechanisms of self-correction might, like Sztompka, point to journal-based peer review and disciplinary committees charged with investigating research misconduct, but it is hard to know because it is rare for scholars who claim that science is self-correcting to articulate what the self-correcting processes are. One notable exception is Jamieson and colleagues' (2019), who outline some concrete standards that make science trustworthy. Consistent with our assessment, their paper identifies journals and university committees as key stakeholders in enforcing and signaling these standards (see Table 1), though they also discuss the role of individual authors and critics in the scientific community.

Traditional scientific institutions are ineffective self-correction mechanisms

Journal-based peer review is thought by many to be the most important error-detection feature of science. However, numerous studies suggest that journal-based peer review does not catch the majority of errors. A study at the British Medical Journal sent test papers with nine major errors to over six hundred peer reviewers. Reviewers were given different levels of training to assess whether intervention could help, but for all groups, the mean number of errors detected was three or fewer (Schroter et al., 2008).

Given the enormous number of errors that are not detected by journal peer review, clearly additional, post-publication mechanisms are needed for a field to claim that they prioritize self-correction, and therefore that their published findings are highly credible. In theory, journals themselves provide one potential avenue for such post-publication corrections, via published commentaries, critiques, and corrections. However, anecdotal reports suggest that the bar for publishing reports of errors in a scientific paper can be very high (Heathers, 2015; Pickett, 2020; Friedman & colleagues, in press; Goldacre et al., 2020).

Likewise, university and publisher- or journal-based committees charged with investigating potential research misconduct are notorious for being unresponsive to evidence, which was one reason the Office of Scientific Integrity, now the Office for Research Integrity, was created in 1989 (Gustson, 1994). In both cases, the institutions' interests are out of line with those of science - journals and universities have strong disincentives to admit that their own research is faulty, and the independent Office of Research Integrity that was created in response is woefully under-resourced (Retraction Watch, 2014). Heathers (2015), Pickett (2020), and Friedman and colleagues (in press) have provided first-person accounts of the difficulty of bringing errors or misconduct to light through such formal channels as journals or university disciplinary committees. Clearly, the traditional safeguards are not always effective at ensuring that self-correction is prioritized. Indeed, they have often impeded correction even when scientists have come forward, at sometimes significant cost to themselves, with strong evidence of errors or misconduct.

If peer-reviewed journals and disciplinary committees cannot be counted upon to reliably detect and correct errors in science, different mechanisms are needed. Luckily, more avenues are open to us now than were in previous centuries, thanks to advances in computing. Data and figures can be re-analyzed much more efficiently than in the past, and reports of errors can be shared on the internet without any gatekeeping. However, the internet is a big place, so for these reports to come to the attention of researchers on the topic, a database must link the comments to the paper (Eagleman & Holcombe, 2003). PubPeer is one website that does this, and thriving communities in some fields, such as cancer biology, have begun to embrace it as a place to discuss possible errors and data issues in specific papers. Multiple venues have flowered in the wake of the pandemic and the associated need to rapidly review new research (Holcombe, 2020). These examples illustrate the potential for more robust and transparent mechanisms for self-correction than the opaque, conflicted, and highly time-consuming journal peer review processes and misconduct investigation processes.

Towards better indicators of self-correction mechanisms

As we have argued above, public trust in science does not seem based on trust in individual scientists' commitment to self-correction, but likely on a belief that scientific communities are structured in such a way that self-correction operates at the group level, and is prioritized by the group. However, the few institutions that are sometimes identified as sources of self-correction in science - namely, peer reviewed journals and university misconduct committees - have not delivered. Nevertheless, so far, the public seems to be willing to take it for granted that these social mechanisms for self-correction exist and are effective. However, to maintain public trust in science, and to allow the public to calibrate their trust in different scientific communities according to the presence and strength of these mechanisms, it is important to identify these mechanisms and make them visible and verifiable to those outside of a particular scientific community. While confidence in science has remained high during the social, behavioral, and health sciences' replication crisis, reassuring the public with appeals to invisible self-correction mechanisms (or with handwaving at journal peer review or misconduct committees) may not be effective forever.

Ideally, self-correction mechanisms in science should be visible, quantifiable, and easy for outsiders to evaluate. Below we propose a preliminary list of qualities that are: a) characteristics of scientific communities rather than individuals, b) potentially measurable, and c) potential indicators of a commitment to self-correction. This list is meant as a work in progress, to be expanded, revised, and adapted.

Measurements of Observable Self-Correction (MOSCs)

Self-correction requires two steps: the errors need to be out in the open (transparency), and there needs to be active checking of each other's work (critical appraisal). Indeed, in their paper "Signaling the Trustworthiness of Science", Jamieson et al. (2019) discuss both transparency-related norms, and a "culture of critique" as fundamental to maintaining the trustworthiness of science. Consistent with this, we have split the MOSCs into two categories: transparency and critical appraisal. While this distinction is not as clear in reality as we present it here, we believe it is useful. In our experience, many scientists have accepted the value of transparency, but stop there. This is problematic because while transparency is necessary for credibility, it is not enough. Even the most transparently-reported science can be very flawed; the flaws will just be out in the open. However, there is no guarantee that those flaws will be identified and corrected, even with full transparency. While transparency is certainly better than secrecy, self-correction cannot happen if that transparency is not then used to detect and correct errors (Gelman, 2017; Vazire, 2020).

We have noticed that some researchers, even those who have bought in to the value of transparency, sometimes balk at the idea that other researchers may use their transparency against them, to point out errors or flaws. We argue that this is exactly

what transparency is for, and why transparent scientific reports are more credible than opaque reports - it is because the scientist engaging in transparency is giving her critics ammunition that we trust her research more. A scientific community that encourages transparency while discouraging criticism is not one that prioritizes self-correction. Transparency enables critical appraisal, and what philosopher of science Helen Longino (1990) calls “transformative interrogation”, but transparency alone is not enough to ensure it happens.

TABLE 1. MOSCs

Indicators that a field is transparent	**Indicators of critical appraisal in a field**
Open data	Error detection
Open code	Post publication peer review
Open materials and methods	Bias detection (at aggregate level)
Contributorship	Computational reproducibility checks
Transparent peer review	Empirical replicability checks
Pre-registration/pre-analysis plans	Publication of negative/null results
Registered Reports	Strong theory and predictions
Level playing field/no barriers to entry	Diversity

The importance of transparency, including the facets listed in Table 1, has been explained by others (Asendorpf et al., 2013; Christensen, 2020; Nosek et al., 2018). Because of the extensive amount of attention that has been devoted to transparency, many may assume that we are well on our way to adequate levels of transparency

across the sciences. However, it is important to recognize that transparency remains very uneven across fields (e.g., Tenney et al., 2020). As part of the Reproducibility Project cancer biology, Errington and colleagues (2014) set out to examine the methods of 51 papers which reported the results of 197 experiments. Unfortunately, last year Errington (2019) reported that for none of those 197 experiments were they able to design a full protocol based purely on the paper, without communicating with the authors. We are optimistic that the situation is not so dire in most fields, but there is clearly a spectrum. Metascientific efforts to evaluate levels of transparency across various scientific disciplines are under way (e.g., Christensen et al., 2020).

Because there is already an extensive literature on the value of transparent and open practices in science, we won't discuss those MOSCs here beyond listing the ones we are aware of in the left column Table 1. Instead, we focus on an aspect of self-correction that has received comparably little attention, our second pillar (right column of Table 1): critical appraisal.

Critical Appraisal MOSCs

Error detection

A field's commitment to self-correction can be gauged in part by how many "unforced errors" appear in the published record - errors that are easily preventable. New tools and algorithms have emerged which can be applied to detect such errors, such as Statcheck and GRIM, which provide some checks of whether the different numbers reported in a paper are consistent with each other (Nuijten et al., 2014; Brown & Heathers, 2017). Some fields, such as parts of high energy physics, have a tradition of setting up their full analysis pipelines prior to feeding it any real data, and running tests on the pipeline to check for errors. Software tools such as Docker (Boettiger, 2015; Clyburne-Sherin et al, 2019) allow anyone to archive a fully working version of their analysis code. A field that prioritizes self-correction is one that invests in developing and using these kinds of tools, and that is shown to have few such errors when these tools are applied to its published literature.

Post publication peer review

A field should not have a very high bar for publication of critical comments, and such comments should be easy to find. This is important because, as James Heathers (2015) wrote, "Formal critical correspondence is the ONLY way to make a visible and public correction to the official version of a journal article. It is written, archived and maintained in such a manner that it is irreparably bound to the criticism, which also received the distinction of being 'published'." Anecdotally, however, there is a high bar in many fields (see the "Traditional scientific institutions are ineffective self-correction mechanisms" section above).

Less formal channels for post publication critique, such as annotations (e.g., using hypothes.is), social media posts, and comments on databases such as PubPeer, can also serve an important function, although they are unlikely to soon achieve the level of effectiveness of the publication of critical commentaries in formal outlets. Thus, the highest level of commitment to self-correction in the domain of post publication peer review would be an abundance of critical commentaries published in high profile journals. A less compelling but still important indicator would be the amount and quality of critical commentary happening in informal channels. Finally, even formally published, and damning, criticism seems to often have little effect on the citation rate of the criticized work.

Bias detection

Although bias can be difficult to detect in an individual study, meta-analytic and systematic reviews provide insights into the prevalence of bias in research literatures. These techniques have evolved rapidly in recent years to develop various indices of bias, such as *p*-curves (Simonsohn et al., 2014). These techniques are now regularly applied to groups of papers in systematic reviews and meta-analyses, and can also be used for other metascientific purposes such as examining levels of bias across journals, over time, or across different methods or practices (e.g., papers reporting primary analyses with vs. without covariates). The results help us to monitor how a field is doing at combating publication bias and other distorting practices, such as *p*-hacking. Moreover, by examining where bias is most severe, each field can begin to develop interventions for reducing bias and improving the credibility of its published literature. Therefore, the prevalence of such bias-detection efforts in a field, and the amount of bias detected, are indicators of a commitment to self-correction.

Computational reproducibility checks

The vast majority of scientific results involve numerical data analysis to calculate summary and test statistics. Both the analyses themselves and the reporting of the associated numbers are highly error-prone. Indeed, the gap between the data and the claims made in a paper can be so vast that Buckheit and Donoho (1995) famously wrote, “An article [...] in a scientific publication is not the scholarship itself, it is merely advertising for the scholarship.” (p. 5). Thus, to ensure that errors and discrepancies between the data and the report are caught, it is crucial that audits be performed to compare the reports to the underlying data.

The data and code underlying the quantitative results of scientific outputs have become increasingly available in many fields of science, thanks to open science reforms encouraging or mandating such sharing. This has made it possible for metascientists to systematically audit samples of published papers to evaluate whether the papers are

computationally reproducible - that is, whether the quantitative results can be reproduced from the original data (i.e., by re-analyzing the same dataset that the original report is based on). Such efforts have produced rather disappointing results. In one analysis of 37 papers published in the journal *Cognition* after the journal implemented a mandatory data sharing policy, the metaresearchers were only able to computationally reproduce all of the results in 11 out of 35 articles with reusable data, and a further 11 could be reproduced after getting assistance from the original authors (Hardwicke et al., 2018). For the remaining 13 articles, the full results could not be reproduced from the original data, even after obtaining assistance from the original authors.

There is a great deal of variance in how seriously journals and researchers take the problem of computational reproducibility. Many journals continue to leave it up to authors to decide whether to share their data and code with other researchers. Other journals require authors to submit reproducibility "packages" which are then checked by an external team, and all reproducibility problems must be fixed before the article can be published (*AJPS* verification policy: https://ajps.org/ajps-verification-policy/). Going even further, the journal *eLife* now publishes articles with interactive figures and live code that connect data to the analysis, all within the browser (Maciocci et al., 2019). We can therefore evaluate a scientific community's commitment to self-correction by evaluating its journals' policies, or conducting post-publication reproducibility checks and directly estimating the computational reproducibility of published findings.

Empirical replicability checks

An important part of error control is having a handle on the prevalence of false positive and false negative errors, as well as Type M and Type S errors - errors of magnitude and sign (direction) of effects (Gelman & Carlin, 2014). There is no straightforward way to estimate the prevalence of these errors, but they are among the most consequential (especially the rate of false positives among claimed "discoveries", i.e., the False Discovery Rate, given the preponderance of positive results in most scientific literatures - see next section). The best way to estimate the prevalence of these errors is a very expensive and time-consuming process: replicating studies from scratch by using similar methods to collect new data. Large-scale replication projects provide an indication of how easily one can collect new data and find the same result as previously published studies of a particular type and research area. Camerer et al. (2018) replicated 21 systematically-selected experimental social science studies published in Nature and Science between 2010 and 2015, with disappointing results. Such efforts should themselves be repeated to help assess whether fields are improving in whether their findings replicate. Replication studies of individual papers are increasingly welcomed by journals and the community, and this trend must continue if we want people to have confidence that most results in a field or literature are replicable. Thus,

the prevalence of replication studies, their appearance in high profile outlets, their impact, and the level of receptivity to replication projects and their results, are all indicators of a field's commitment to self-correction.

Publication of negative/null results

Various scientific fields have an abysmal record of publishing negative and null results (Sterling, Rosenbaum, and Weinkam, 1995; Fanelli, 2010; 2012; Greenwald, 1975; Hubbard & Armstrong, 1992; Franco, Malhotra, & Simonovits, 2014). Fortunately, many researchers already appear to see this as a way to track whether the credibility of specific fields is improving. Metascientists are publishing new techniques for estimating publication bias (e.g., Andrews & Kasy, 2019) and initiatives such as Registered Reports have been devised and adopted to combat it (Chambers et al., 2015). Preliminary results from analyses of the emerging Registered Reports literature suggest that Registered Reports are successful at allowing more negative results to get make it into the published literature (Allen & Mehler, 2019; Scheel, Schijen, & Lakens, 2020).

Strong theory and predictions

Famously, some theories in physics make precise predictions. Einstein's theory of general relativity predicted that light would bend twice as much as Newtonian theory predicted. When Einstein's prediction was confirmed by observation (Dyson, Eddington, & Davidson, 1920), his theory became much more popular; it became highly credible.

Construction of strong theories may be premature in many sciences, such as some areas of psychology, which arguably should first concentrate on characterizing phenomena (Perfors, 2020; Rozin, 2009). However, once an area does have strong theories, those theories can accumulate a track record of accurate predictions, and fields that use such theories will, appropriately, gain credibility. The adoption of preregistration can facilitate accumulating such a track record. Thus, indicators of a commitment to develop the foundations for strong theory include a balance of research along the continuum from descriptive/exploratory to hypothesis-testing/confirmatory, with clear labeling of where along this continuum any particular finding falls, and with hypothesis-testing work accompanied by detailed pre-registered plans. Other indicators could be developed to quantify the degree to which the precursors to strong theory are in place, and the degree to which theories make clear, precise predictions.

Diversity

Diversity within a field helps overcome hindrances such as shared biases and close ties among the researchers studying a particular topic (Longino, 1990). People with different theoretical commitments and knowledge bases are likely to catch different sorts of errors. Those attached to different theories will be motivated to check, or double-check,

different calculations. In contrast, uniformity in approaches and perspectives breeds group-think. As Cordelia Fine (2020) wrote, “Whether for reasons self-serving or benign, everyone comes laden with prior knowledge, background assumptions and frameworks. That’s why it takes a diverse village, so to speak, to nurture scientific objectivity.”

Diversity in personal and institutional ties may also be important. In some areas of psychology research, nearly all researchers who conduct studies with particular paradigms may know each other. The gatekeepers at certain journals may be exclusively drawn from narrow clans, which can result in the suppression of new perspectives and approaches. This would be difficult to quantify, but qualitative studies and network analyses might provide some indication of whether this is the case for a particular research area. Geographical and institutional diversity in journal editors can be an additional, albeit very imperfect, proxy.

Conclusion

A high level of trust is justified when something has a long and consistent track record of claims being true. The replication crisis shows that this is unfortunately not the case for some areas of science. The American president Ronald Reagan, in negotiation with his Soviet counterpart Mikhail Gorbachev, frequently invoked a Russian proverb, “Doveryai, no proveryai”, which means “trust, but verify.” At one time, perhaps, there was no reason to suspect that the sciences could not be trusted. Today, in the aftermath of the replication crisis, that is no longer the case.

Now, if we want people to trust science, we need to justify their trust. Different sciences will likely have different levels of justification. For each science, we should have something that allows at least some assessment of whether trust is warranted. We can no longer blithely declare “science is self-correcting” and expect that to reassure science’s stakeholders. We need to determine where self-correction mechanisms are working well and working rapidly, to buttress public trust in science where it is warranted, to improve the self-correcting mechanisms where they are weak, and to prevent less robust sciences from dragging other sciences down with them.

References

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS biology*, *17*(5), e3000246.

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics*, *2*(4), 3-14.

Andrews, I., & Kasy, M. (2019). Identification of and Correction for Publication Bias. American *Economic Review, 109(8),* 2766–2794. https://doi.org/10.1257/aer.20180310

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

Boettiger, C. (2015). An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review, 49(1), 71-79.

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

Buckheit, J. B., & Donoho, D. L. (1995). Wavelab and reproducible research. In *Wavelets and statistics* (pp. 55-81). Springer, New York, NY.

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

Campbell, D. T. (1988). *Methodology and epistemology for social sciences: Selected papers*. University of Chicago Press.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex, 66,* A1–A2.

Christensen, G., Wang, Z., Levy Paluck, E., Swanson, N., Birke, D., Miguel, E., & Littman, R. (2020). Open science practices are on the rise: The State of Social Science (3S) Survey. MetaArXiv. https://osf.io/preprints/metaarxiv/5rksu/

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational Reproducibility via Containers in Psychology. *Meta-Psychology, 3*.

Tenney, E. R., Costa, E., Allard, A., & Vazire, S. (2020). Open science and reform practices in organizational behavior research over time (2011 to 2019). Manuscript in preparation.

Dyson, F. W.; Eddington, A. S.; Davidson C. (1920). "A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of 29 May 1919". *Philosophical Transactions of the Royal Society, 220A* (571–581): 291–333.

Eagleman, D. M., & Holcombe, A. O. (2003). Improving science through online commentary. *Nature, 423(6935),* 15–15.

Errington, T. (2019, September). Talk presented at the Metascience Symposium, Palo Alto, CA. Retrieved from https://www.metascience2019.org/presentations/tim-errington/

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science forum: An open investigation of the reproducibility of cancer biology research. *Eife*, *3*, e04333.

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.

Fine, C. (2020, July). Sexual dinosaurs: The charge of ‘feminist bias’ is used to besmirch anyone who questions sexist assumptions at work in neuroscience. *Aeon.* Retrieved from https://aeon.co/essays/trumped-up-charges-of-feminist-bias-are-bad-for-science

Fiske, S. T., Schacter, D. L., & Taylor, S. E. (2016). Introduction. *Annual Review of Psychology, 67*.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345(6203),* 1502–1505. https://doi.org/10.1126/science.1255484

Friedman, H., MacDonald, D. A., & Coyne, J. (in press). Working with psychology journal editors to correct problems in the scientific literature. *Canadian Psychologist.*

Funk, C., Hefferon, M., Kennedy, B., & Johnson, C. (2019). Trust and mistrust in American’s views of scientific experts. Pew Research Center. https://www.pewresearch.org/science/2019/08/02/trust-and-mistrust-in-americans-views-of-scientific-experts/

Gelman, A. (2017). Ethics and statistics: Honesty and transparency are not enough. *Chance*, *30*(1), 37-39.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651.

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118. https://doi.org/10.1186/s13063-019-3173-2

Greenwald, A. G. (1975), “Consequences of prejudice against the null hypothesis,” Psychological Bulletin, 82, 1- 20.

Guston, D. H. (1994). The demise of the social contract for science: Misconduct in science and the non-modern world. *The Centennial Review, 38(2),* 215–248. https://www.jstor.org/stable/23740126

Haack, S. (2003). *Defending Science - Within Reason: Between Scientism and Cynicism.* Prometheus Books.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Lenne, R. L. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society open science*, *5*(8), 180448.

Heathers, J. [@jamesheathers]. (2019, March 01). In reality, mechanisms to correct bad science are slow, unreliably enforced, capricious, run with only the barest nod towards formal policy, confer no reward and sometimes punitive elements for a complainant who might use them. [Tweet]. Retrieved from https://twitter.com/jamesheathers/status/1101161838308401157

Heathers, J. (2015, October 2). A General Introduction: Formal Criticism in Psychology. Medium. https://medium.com/@jamesheathers/a-general-introduction-formal-criticism-in-psychology-ba193a940ec8

Heesen, R., & Bright, L. K. (n.d.). Is Peer Review a Good Idea? *The British Journal for the Philosophy of Science.* https://doi.org/10.1093/bjps/axz029

Holcombe, A. O. (2020, May). As new venues for peer review flower, will journals catch up? Psychonomic Society featured content. Retrieved from: https://featuredcontent.psychonomic.org/as-new-venues-for-peer-review-flower-will-journals-catch-up/

Hubbard, R., & Armstrong, J. S. (1992), “Are null results becoming an endangered species in marketing?” *Marketing Letters, 3,* 127-136.

Jamieson, K. H., McNutt, M., Kiermer, V., & Sever, R. (2019). Signaling the trustworthiness of science. *Proceedings of the National Academy of Sciences*, 116(39), 19231–19236. https://doi.org/10.1073/pnas.1913039116

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago press.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Maciocci, G., Aufreiter, M. and Bentley, N. (2019). Introducing eLife’s first computationally reproducible article. *ELife* blog. https://elifesciences.org/labs/ad58f08d/introducing-elife-s-first-computationally-reproducible-article

Markman, A. Why science is self-correcting. (2010). Ulterior motives blog on Psychology Today. Retrieved July 28, 2020, from http://www.psychologytoday.com/blog/ulterior-motives/201008/why-science-is-self-correcting

McConnell, J. [@JohnSMcConnell] (2020, June 14). Retraction should be reserved for publication misconduct. Otherwise, let the self-correcting mechanisms of science take their course. [Tweet]. Retrieved from https://twitter.com/JohnSMcConnell/status/1271860082427547649

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115(11),* 2600–2606.

Nuijten MB, Hartgerink CHJ, van Assen MA, Epskamp S, Wicherts JM (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48: 1205–1226.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Oreskes, N. (2019). *Why trust science?*. Princeton University Press.

Perfors, A. (2020). Generating theories is hard, and none of our theories are good enough. Presentation for the Annual Meeting of the Society for Mathematical Psychology. https://virtual.mathpsych.org/presentation/109

Pickett, J. T. (2020). The Stewart Retractions: A Quantitative and Qualitative Analysis. *Econ Journal Watch, 17(1),* 152–190.

Retraction Watch (2014, March). In sharp resignation letter, former ORI director Wright criticizes bureaucracy, dysfunction. Retrieved from https://retractionwatch.com/2014/03/13/in-sharp-resignation-letter-former-ori-director-wright-criticizes-bureaucracy-dysfunction/

Rozin, P. (2009). What Kind of Empirical Research Should We Publish, Fund, and Reward? A Different Perspective. *Psychological Science, 4(4),* 435–439.

Scheel, A. M., Schijen, M., & Lakens, D. (2020). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. Retrieved from https://psyarxiv.com/p6e9c/

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine, 101(10),* 507–514. https://doi.org/10.1258/jrsm.2008.080062

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143(2),* 534.

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

Sztompka, P. (2007). Trust in science: Robert K. Merton's inspirations. *Journal of Classical Sociology*, *7*(2), 211-220.

Vazire, S. (2020, January). Do we want to be credible or incredible? *APS Observer*. Retrieved from https://www.psychologicalscience.org/observer/do-we-want-to-be-credible-or-incredible

Ziemann, M., Eren, Y., & El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome biology*, *17*(1), 1-3.