Taking Responsibility for Our Field's Reputation

August 31, 2017

To put it bluntly, academic psychology's public reputation seems to be in free fall.

When the press coverage of the "replicability crisis" in psychological science first began a few years ago, reporters generally broached the topic in a respectful and delicate fashion, hinting at problems but not trumpeting them.

That has changed noticeably in the past year or two. Science reporters who used to assume that peer review was a hallmark of "scientific literacy" are now openly stating that peer review is unreliable — not just in psychological science but across the scientific disciplines.

And the public is getting the message, if the comments sections of online science news stories are any indication.

News reports of our field's replication initiatives illustrate some troubling sentiments. Some typical remarks: "Psychology is not a science, I repeat not a science. It is similar to beliefs like religion."[1] and "Surprise, surprise, [psychologists are] quacks that produce garbage they claim is science."[2]

The perception that psychological science is more afflicted with replication problems than other fields may be completely unwarranted (for example, early results from the Cancer Reproducibility Project suggest at least as big problems there). And there is little doubt that psychological researchers (and APS itself) have done much more than their share to address these problems head-on. But be that as it may, our reputation is in need of a lift.

Looking Backward to Go Forward

So what can we do? Discussion in the field has been almost entirely focused on the question of how to improve the quality of research *going forward*. The past few years have seen a creative outpouring of progressive research models such as Preregistered Replications and Many Labs Projects.

That's great, but our sole basis for our science being taken seriously by anyone besides ourselves is the credibility of our literature — our cumulative work product — rather than our good intentions for the future.

The numerous unsuccessful attempts to replicate findings — including some rather landmark results — means the textbooks we require our students to buy and the lectures we deliver probably describe as many false findings as true ones. But while tentative and incomplete theoretical understanding is a normal step in the scientific process, a literature strewn with empirical claims based, in part, on botched research and faulty peer review is something else entirely. And over the coming years, that situation is likely to sink into the minds of the educated public and wreck our reputation as a science.

So What Can We Do to Bolster Our Credibility?

We contend that our reputational problem *can actually be fixed rather rapidly and decisively* if we embrace an aggressive (and painfully thorough) commitment to honest labeling. We see this commitment as entailing two changes in our practices, which we can crudely label *truth in packaging* and *investigator accountability*.

Truth in Packaging

As a first key step, all reviews or summaries produced by psychological scientists, whether in textbooks or review-oriented journals or books (including mass-market books), need to *explicitly and conservatively label the degree of support enjoyed by any research finding that is mentioned*. The highest credibility category — call it "Class 1" — must be reserved for findings that have been confirmed in one or more preregistered replication, where publication bias, HARKing (hypothesizing after the results are known), and *p*-hacking can all be confidently excluded.

A few years ago, most of us would have assumed that *being confirmed in a number of conceptual replications* was an admirably high and thoroughly reassuring level of confirmation (confirmation with evidence of generalizability thrown in for good measure, as it were). However, we now see that this is incorrect. A discipline's reliance on conceptual rather than direct replication interacts with publication bias to vitiate its literature more effectively than anything else we know of (Pashler & Harris, 2013). Over and over again, we see literatures that are resplendent with varied and imaginative conceptual replications, and yet somehow *no result in particular* ever seems to replicate when a direct replication is undertaken. Clever new meta-analytic tools will not rescue us either. For instance, precognition research, claiming support for ideas that would violate the laws of physics, has been given a bill of good health by the "*p*-curve" technique (Simonsohn, Nelson, & Simmons, 2014; Bem, Tressoldi, Rabeyron, & Duggan, 2015).

This Honest Labeling proposal is admittedly fraught with unappetizing consequences. Many of us — including the present authors — believe that there are large swaths of research within big parts of psychology where statistical power is good (usually because of general reliance upon repeated measures and within-subject designs). In these areas, we suspect, most of the findings reported probably stand up. Still, a field that means to take itself seriously needs to brand any result lacking in preregistered replication (including many of our own) as Class 2 (strongly suggested in the literature but not scientifically confirmed). All research that is based on single studies and/or low-powered designs we would call Class 3 evidence, or, to borrow a phrase from Harold Jeffreys' classic *Theory of Probability*, "worth no more than a bare mention."

Ironically, our proposal to classify a single experiment that has reached conventional levels of significance as a preliminary finding corresponds well with how English statistician Ronald A. Fisher himself, in his 1935 book *The Design of Experiments*, suggested people should view this test, which is to ignore results that are not even able to jump over this low bar.

He said:
"It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to *ignore all results which fail to reach this standard*, and, by this means, to

eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results" (p. 13, italics ours).

So in Fisher's view, 5% significance is far from a sufficient condition for accepting a hypothesis as true. For serious credibility, he required replicability, saying, "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." That corresponds closely with what we propose for Class 1.

The positive consequences of adopting this disciplinary self-control will be immediate: Results labeled as having only second- or even third-class empirical support will constitute an obvious invitation for researchers to set about replicating these findings. We probably will discover that a great many results are real, which could produce a torrent of good news. These can then be confidently cited in textbooks and press releases. But we will also find that many results are more problematic than we ever suspected (see, for example, the recent failures to confirm the well-known "blocking effect" in Pavlovian conditioning; Maes et al., 2016).

We believe that the broad adoption of open labeling will create an incentive to focus resources on finding out what we think we know is real and what is unreal. And it stands to substantially stop the precipitous drop in the public reputation of psychological research.

Personal Accountability

Our field also needs to persuade individual scientists to respond more positively, as many have done, when other researchers fail to replicate their original results. Blaming a failed replication on unknown moderating factors should not be the acceptable response. If the original investigators know how to get the effect, let them step forward, repeat it, and show us all how to do it.

We propose a standard of accountability common in many professional fields. The code of conduct for professional engineers, for example, holds them accountable for the structures they design. And prior to 2010, the (rare) published failures to replicate seemed to breed a sense of obligation on the part of original investigators to try to recreate their phenomena. In the past few years, however, that sort of accountability seems to have diminished.

An ethic of personal responsibility should, in our view, be recognized as a fundamental ethical standard for psychological scientists, as well. Thus, it should be included in the ethics courses that are now becoming a standard part of graduate instruction. The goal here is not to embarrass those whose results turned out to be flimsier or narrower than suspected. Rather, it is to allow the field to efficiently determine if failures reflect methodological changes in the replication attempt or error in the original study. Figuring that out quickly is obviously Job No. 1 if we actually care about getting it right as a field.

To sum up, we believe that by embracing and codifying two explicit new standards of *truth in labeling* and *responsibility in authorship*, we can do a great deal to shore up the reputation of academic psychology. Of course, the exact mechanics needed to implement such a commitment are substantial, and the scientific community would need to discuss them extensively, but the proposed commitments are well within our individual and institutional resources. We hope to see the field that we love so much

working in this way to hold on to its reputation as a respectable scientific enterprise.

Is This Really Necessary?

Some well-known and outspoken colleagues seem to believe that the radically self-critical measures we advocate here are quite unnecessary — that we can make all this go away by deploying a few quick and facile defenses. "Move along, nothing to see here," would seem to be a preferred approach for some. In fact, the evidence for widespread problems is not restricted to the report of the now-famous Reproducibility Project (Open Science Collaboration, 2015), which found that less than a third of the findings could be replicated with standard statistical approaches. More sophisticated analyses of those data back up the bad news (Etz & Vandekerckhove, 2016; Morey & Lakens, 2016), and independently, a large swath of replications appearing in a special issue of the journal *Social Psychology* (Marsman et al., 2017) find more than "anecdotal" levels of evidence for only 7 out of 60 significant findings from the original articles.

Others lament the difficulty of establishing solid psychological findings and attribute the failure of our replications to hypothetical ad- and posthoc "moderator" variables.

These kinds of arguments carry no weight for a number of reasons. One is the deadly combination of publication bias and low power that indisputably characterizes much of behavioral research. Operating in tandem, these twin defects are in and of themselves fully sufficient to guarantee that our literature will be replete with imaginary findings, as statistician John Ioannidis showed in his famous 2005 article (see also Szucz & Ioaniddis, submitted). Second, the situation is really far worse than what Ioannidis reckoned, since we now know that our field has many corrupted data-analysis practices (e.g., *p*-hacking) that greatly exacerbate the impact of publication bias and low power (Simmons, Nelson, & Simonsohn, 2011).

Anyone who views the field's problems as exaggerated needs to explain (preferably supported by convincing simulations) how we could possibly be getting reliable one-shot findings given the malign combination of low power, publication bias, *p*-hacking, and the evidently low bar of our conventional threshold of 5% significance. Continued efforts along these lines will simply make us look defensive and deceptive.

By contrast, embracing open self-criticism and responsibility as investigators and as a field, as we propose here, will enable us to earn and ultimately enjoy the reputation we seek.

–**Hal Pashler**
University of California, San Diego
-J. P. de Ruiter
Tufts University

Notes

[1] smithsonianmag.com/science-nature/scientists-replicated-100-psychology-studies-and-fewer-half-got-same-results-180956426

[2] theatlantic.com/science/archive/2016/03/psychologys-replication-crisis-cant-be-wished-away/472272/#comment-2557372082

References and Further Reading

Bartlett, T. (2013, January 30). The power of suggestion. *The Chronicle of Higher Education*. Retrieved from www.chronicle.com/article/Power-of-Suggestion/136907/

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research, 4:1188*.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS one, 11*, e0149794.

Ferguson, C. J., Brown, J. M., & Torres, A. V. (2016). Education or indoctrination? The accuracy of introductory psychology textbooks in covering controversial topics and urban legends about psychology. *Current Psychology*, 1–9.

Fisher, R. A. (1935). *The design of experiments*. New York, NY: Macmillan.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*, e1002106.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*, e124.

Jeffreys, H. (1961). *Theory of probability*. Oxford, United Kingdom: Clarendon Press.

Maes, E., Yannick, B., Matías, J., Palloni, A., Krypotos, A. M., D'Hooge, R., … Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General, 145*, e49–e71.

Marsman, M., Schönbrodt, F., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'Replications of Important Results in Social Psychology.' *Royal Society Open Science, 4*: 160426.

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. Manuscript submitted for publication. Available at: github.com/richarddmorey/psychology_resolution/blob/master/paper/response.pdf

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666–681.

Szucs, D., & Ioannidis, J. P. A. (submitted). When null hypothesis significance testing is unsuitable for research: A reassessment. Available at www.biorxiv.org/content/biorxiv/early/2016/12/20/095570.full.pdf