Reliability in Psychology: Means Versus Ends

While riffing on our field's "replication crisis" in a recent *Observer Forum* piece, "Taking Responsibility for Our Field's Reputation," two of my favorite psychologists posed a challenge: "Anyone who views the field's problems as exaggerated needs to explain … 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" (Pashler & de Ruiter, 2017, p. 10). Darn it: I had been hoping to keep my head down and continue watching this conversation play out from the sidelines, but I must admit that I fall into this camp: Though an enthusiastic supporter of several of Pashler and de Ruiter's specific exciting proposals, I do also worry that we have been exaggerating the scope of our problems in at least one important and underdiscussed way. Pashler and de Ruiter argue that I have a responsibility to explain, which seems fair.

A Vision of Psychological Reliability

I identify in part as a vision scientist, and from this perspective our field's crisis can feel a bit odd. After all, one of the core problems fueling this turmoil (and the first one mentioned in the quote above) is low power: We have too often drawn conclusions from lightweight studies testing barely a dozen subjects. But visual psychophysics is built (both historically and still today) on a foundation of experiments that routinely feature just two or three observers (often with an implicit feeling that one or two of them may be gratuitous). Yet vision science doesn't seem to be suffering to the same degree that some other subfields of our discipline are. At least, I would characterize the reaction of many of my psychophysicist colleagues more in terms of puzzlement than panic. It may be a bit harder to highlight high-profile replication failures in this area. (Should we be worried that visual crowding may not be real? Or motion adaptation? Or the Muller-Lyer illusion? Or motion-induced blindness? These phenomena have been studied extensively, but rarely with direct replications across labs.) And we now know, as those of us who teach lab classes have long suspected, that at least some effects in the wider neighborhood of cognitive psychology are rather impressively robust (e.g. Zwaan et al., in press).

No doubt there are several independent reasons for this state of affairs. For one, in vision science those two or three observers are often completing hundreds or thousands of trials each — something not always taken into account when calculating "power." Second, such effects are often large and robust (with "$p < .001$" not being at all uncommon). Third, statistics aren't our only source of evidence in the study of perception: We also sometimes rely on compelling subjective demonstrations. (In our lab, we strive not just for $p < .001$, but also for $p <$ "Holy cow: look at that!") Here, however, I'd like to focus on a fourth reason — one that underlies why our field's replicability crisis strikes me as perhaps inflated in some discussions.

Preregistration and "Postregistration"

How might some studies manage to avoid collapsing even with (what is in one sense) low power and few cross-lab direct replications? You might think that the answer would have to involve preregistration. After all, preregistration is nearly universally hailed these days as The Answer, and it is rewarded accordingly. It is "the *only* way for authors to irrefutably demonstrate that their key analyses were not *p*-hacked" (Simmons et al., in press; emphasis added). It is required for what Pashler and de Ruiter call Class 1, the "highest credibility category" into which we can place a research finding. It can earn you visible respect in the form of a nifty badge, initially at *Psychological Science*, and now in other APS journals. It can even net you a share of $1 million, as in the "Preregistration Challenge" from the Open Science Framework (cos.io/prereg/).

Still, preregistration seems to be spreading into vision science a bit more slowly than in some other subfields — and of course none of the older foundational work in psychophysics was preregistered. So what has kept it from collapsing?

An important part of the answer, I think, is that studies in this area of our field are frequently "postregistered." This is a term that I like to use for papers that include internal replications of their primary effects — in separate samples of identical size, explored via identical analyses, with identical exclusion criteria, etc. This is not at all uncommon for papers in our field, which often feature multiple independent experiments which each replicate the basic effect in question, often while also controlling for a different possible confound or comparing the basic effect to a different variant.

Such internal replications provide an independent test of nearly all of the "researcher degrees of freedom" that may otherwise plague us. Worried that a sample size was *p*-hacked? You may worry less if multiple internal replications are constrained to have the same sample size. Worried about suspiciously baroque analyses or exclusion criteria? You may worry less if the internal replications are constrained to have the same analyses and exclusion criteria. Of course, this is even more true when there isn't much nuance at all — e.g., when simply comparing two distributions with a single test, without any exclusions. I present these as subjective impressions, but of course this help can be quantified: If an initial sample size or analysis plan suffers from *p*-hacking, what is the probability that a second (and

AT RANDOM

"Whether we like it or not, these people really do control our society. The kids who test in the top 1% tend to become our eminent scientists and academics, our Fortune 500 CEOs and federal judges, senators and billionaires."

-Psychological scientist **Jonathan Wai**, Duke University, talking with *Nature* about longitudinal data demonstrating a link between early cognitive ability and adult achievement.

third, and fourth) internal independent replication that is constrained to be identical in these respects will also demonstrate the effect?

Critically, this sort of "postregistration" can help to ensure reliability e*ven when the study wasn't explicitly preregistered.* Indeed, in some ways this underdiscussed solution to our problems may even be better than preregistration (though of course they are not mutually exclusive!). Such constraints can ensure that the internal replications cannot have been *p*-hacked, whereas there is nothing to stop an unscrupulous researcher from preregistering several different variants of a study (e.g., with different sample sizes) and then only linking to the one that ended up working. (Tools such as aspredicted.org have some built-in protections against this, but that still can't stop someone from preregistering different versions across different sites.) And this approach can also save time and words in an exposition. With just a single study, you may need the preregistration, perhaps along with a careful and explicit autobiographical motivation for how you generated your sample size (as *Psychological Science* requires). But with multiple internal replications, you don't need to worry so much about where the sample size came from, as long as it is identical in all of the internal replications.

This approach also helps to protect against the file-drawer problem: With just one experiment, perhaps a researcher actually ran four variants (each separately preregistered?), and then only reported the one that worked. This concern becomes less realistic for a study with several internal replications, all with the same parameters. (Do you really think that the researcher ran 16 variants and then only reported the quartet that worked?)

Multiple Paths to Credibility?

The reason I worry that our field's problems may have been exaggerated in some contexts is thus that so many recent discussions have focused only on one sort of solution to the underlying problems of low power and *p*-hacking, with the implication that studies that have not been preregistered can't count as having the "highest credibility" (per Pashler & de Ruiter's proposal). And this thought also fuels suggestions that somehow the entire literature pre-2010 should be viewed with skepticism, given that approximately none of it was preregistered. But preregistration simply isn't the only way for Odysseus to tie himself to the mast and thus avoid the sirens of *p*-hacking: He can also constrain himself to employ the same methods and analyses and sample sizes (etc.) in multiple internal replications when publishing papers. (Ironically, this point was explicitly noted in some of the earlier discussions of the replicability crisis, but they seem to have been forgotten. For example: "Even if we got a study to work only after 44 attempts, there is still just a 5% chance of it working again under the null: replication *p* values are kosher"; Simonsohn, 2012, p. 597.)

This sort of practice is relatively common in vision science and cognitive psychology. It far predates our field's current turmoil, and the relative frequency of this practice may help to explain the possibly uneven profile of reliability across our field (e.g., Zwaan et al., in press). When you recognize the utility and the frequency of this approach, things may not look quite so bleak — at least in some subfields. In any case, this has been a genuine attempt to accept the charge that those of us who feel that our field's problems may have been exaggerated have a responsibility to explain why. I hope that these thoughts don't count as "quick and facile defenses" that "carry no weight" (Pashler & de Ruiter, 2017, p. 10) and that are part of the problem; instead, I hope that they too might be part of the (multifaceted) solution. At any rate, it bears remembering that many of our field's findings are not in fact "one-shot." Rather, they are multiple-shot findings, even within individual papers, with those shots all sharing many of the same key properties.

A Badge-Oriented Coda: Rewarding Means Versus Ends

To be clear: None of this provides any reason not to preregister a study. As many have pointed out, the cost of doing so (for producers of science) is close to nil, and the advantages are legion (e.g., Lindsay, Simons, & Lilienfeld, 2016; Wagenmakers & Dutilh, 2016). But this is all the more reason not to oversell preregistration by claiming that any study that isn't preregistered is automatically more suspect, or that (for consumers of science) this approach is the only sign of credibility.

In the end, what we should care about is reliability, regardless of the specific means by which we got there. So while I applaud the many benefits of "preregistered" badges in our journals (see Lindsay et al., 2016), I also find them misguided in a way. I hope that we don't start taking the absence of such a badge as necessarily reflecting unreliability, and what I really wish we had was a badge for the ends, and not just one particular means: <*This study — in one way or another! — has a built-in guard against researcher degrees of freedom.*> If those sorts of badges existed, perhaps their frequency might help us to more accurately characterize the reliability of our field — even if preregistration is still the only way to earn your share of a million bucks?

-Brian Scholl

References

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *Observer, 29*, 14–17.

Pashler, H., & de Ruiter, J. P. (2017). Taking responsibility for our field's reputation. *Observer, 30,* 8–10.

Simmons, J., Nelson, L., & Simonsohn, U. (in press). False-positive citations. *Perspectives on Psychological Science.*

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press). *Perspectives on Psychological Science, 7*, 597–599.

Wagenmakers, E.-J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *Observer, 29*, 13–14.

Zwaan, R., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (in press). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review.*

Author Note

For helpful comments on previous drafts — albeit definitely of the that-doesn't-mean-that-they-endorse-any-of-this variety — I thank Chris Chabris, J. P. de Ruiter, Steve Lindsay, Hal Pashler, Joe Simmons, Dan Simons, and the members of the Yale Perception and Cognition Laboratory.