PERSONALITY Science

Open Peer-Review for "Why Has Personality Psychology Played an Outsized Role in the Credibility Revolution?" (Atherton et al.)

Note: *3 reviewers were invited in Round 1, of which 2 accepted and provided reviews. In that round, 1 reviewer wished to remain anonymous. 1 reviewer wished to have their review openly published alongside the final article, with their name attached to it. The authors of the article agreed to have any available pre-publication peer-reviews published. There was only one round of reviews.*

Daniel Lakens:

In this paper the authors review the role personality psychology has played in the current discussions of how to improve science. The authors seem confident personality psychology is punching above its weight and has an outsized role. I appreciate the role of personality psychologists, but it does not strike me as outsized per se. In my country, The Netherlands, psychology psychologists played almost no role in the credibility revolution. When the authors say that personality psychology played a 'leading role' I am just pushed a bit too far to allow this to be just a speculative statement without a citation.

The analysis that personality psychology had the person situation debate, putting the existence of the fields at stake, which prepared it to look critically at itself is at first glance plausible – except that if we realize a much more severe 'crisis' in social psychology instigated by Gergen (what the older generation refers to as *the* crisis, like the British mean the first world war when they say *the* war) and yet social psychology was not prepared. The correct analysis should perhaps not stress the existence of earlier crises (that is similar across fields) but the fact that the field actually addressed them. That requires some explanation in itself. Compared to Gergen's social psychology as history perspective, similar questions needed to be addressed (how stable is social behavior?) – but they do not seem to have lead to research lines with competing predictions. Maybe because the time scale made this difficult, maybe because social psychologist were not interested in resolving the matter, maybe because they solved it by embracing more implicit cognitive processes to study that seemed more stable – I really don't know nor have I reflected on it. But the discussion might need a slightly different focus because other fields also had crises.

It is not clear to me how having a crisis leads to a field doing things like not putting scientists on a pedestal. Indeed, for me as an outsider it seems Mischel *is* on a pedestal because of the crisis. Why would it lead the field to value data sharing or replications? I don't see a clear logical connection, so clarifying that would be appreciated.I Find the subsequent analysis "the fact that personality psychology is a relatively small field, has little money at stake (in terms of grants or external partnerships), isn't popular within scientific psychology" a better reason – but then you don't need a crisis. If the real reasons are the ones highlighted on the end of page 8, how does the crisis narrative fit in?

On page 9 the authors state that "personality psychologists were collecting larger samples, sharing data, reporting null effects, and replicating across studies before these became widely-accepted as being best practices across the rest of psychology". If this is true, you need neither the crisis nor the lack of status. If this was a norm from the start, maybe that is the explanation (and maybe having resources or ease of data collection are important, e.g., mainly survey based research). To summarize, in this section it is not clear what the authors argue is the cause and what the consequence, and how they relate. I understand this is a speculative section, that is not may issue – it is just not clear what their analysis is. Drawing this section out as a DAG might help structuring ones thought.

The idea that personality psychology was already in a good position depends a bit on what you define as a good position. Factors not mentioned are representativeness of the sample, measurement standardization, collaboration, innovation, practical applicability – or any other dimensions researchers would judge good science on. There is more to science than the limited subset of factors improved by the current methodological discussions.

The section on effect sizes interpretation is the least convincing part of the paper. I strongly disagree with the authors that personality psychology has either made any useful contributions to how the field uses effect sizes, nor does an adequate job itself. Renorming useless benchmarks just gives you different useless benchmarks – the fact remains they have no function, no meaningful interpretation, and the field needs to do better. The statement that small effect can accumulate is gratuitous. They can – but personality psychology has not done much (if so, please cite the work) to specify how and when it accumulates (and when not). Most importantly, other fields have done somewhat better, such as the way clinical psychology has thought about which effect sizes are *meaningful*, for example for individual improvement in therapy. So. I remain unconvinced by this section, and even worse, I am slightly worried the authors believe their subfield has done anything that is worth congratulating themselves about. Effect size interpretation is largely absent beyond completely tautological interpretations (i.e., the effect size was small because it was 0.11 and 0.11 is small because we have agreed to refer to it as small). A lot of work is needed here, and it seems better to admit this. I would consider moving this section to the weaknesses part of the paper.

I also disagree that the focus on effect sizes has much to do with the size of samples – indeed, the other reasons mentioned in this paragraph are more convincing (e.g., low cost of data collection). It is good that the field uses larger samples, and even if cost is smaller, some models of science suggest that fields should lower sample sizes to a minimum – which has not happened in personality psych. So yes, say the larger samples are good – but again, the analysis of why this is the case can be clearer.

The last point on real-world relevance, and ability to publish null results due to the focus on a standardized measure like the big five is nice and contains good points.

The re-use of existing data is not a unique challenge for personality psych – and indeed, the focus on making re-use possible means we will hopefully have this problem everywhere. It is also not a problem about data – it is at best a problem of theory. If you can combine data to tell any story, and combine items to create any measure, the problems are not alpha inflation, but measurement and theory. If you want to test things, the severity of the test comes from more than alpha control. It is misleading to say open data is a challenge – just make sure all open data has the big five, and as peer reviewers don't let people publish just do stories. I see the point the authors want to make – but it can be made better. Similarly, the short section on pre-registration is not really worked out well. "The fact that researchers cannot easily anticipate all of the choices they will need to make and problems they will encounter introduces flexibility despite their best intentions." Is true for almost all research, except those studies for which the derivation chain has been rigorously established. Again, the problem is not the alpha, but developing and testing theories. Collaboration might be a better solution than preregistration. Similarly, the section on replication is actually a section on theory.

The section on page 16 where the authors ask "What distinguishes a small but important effect from one that is negligible?" has a very obvious answer: A p-value in a minimal effect test. The authors seem very close to realizing that is a field moves away from NHST to estimation, it has a problem (that is solved by not moving away from p-values). And again, as I mentioned above, the field has the problem that is has not been able so far to *really* interpret effect sizes by stating what the smallest effect size of interest is.

On page 19 the authors state "groups whose percentages in our personality undergraduate programs, graduate programs, faculty positions, and professional societies are lower than their percentages in

the general population". With respect to undergraduate and graduate students, this is surprising to me, because in my country, the percentage of people who identify as LGBT have on average a higher education level than the average population, (https://www.scp.nl/binaries/scp/documenten/monitors/2018/11/21/lhbt-monitor-2018/LHBT-monitor+2018.pdf) and as many women as men receive higher education (and for the younger generation, more women than men go to higher education) https://www.cbs.nl/nl-nl/nieuws/2019/10/evenveel-vrouwen-als-mannen-met-hbo-of-wo-diploma. I understand these numbers will differ per country. This statement would benefit from some sort of Constraints on Generalizability statement, and some citations would be good to support this claim for undergraduates. The authors are correct that measuring these percentages at the level of departments and scientific organizations is difficult, but SIPS has collected some numbers (http://improvingpsych.org/2020/06/16/sips-demographics-report/), and these percentages are again higher than those in the population (e.g., compared to the US: https://en.wikipedia.org/wiki/LGBT_demographics_of_the_United_States#2021). Some context and references in these sections would make the point the authors are making clearer.

On page 19 the authors state "The most overt examples of hostility towards women and underrepresented minorities stem from norms of professional behavior that are common to psychology more broadly, which tolerate harassment and discriminatory behavior in many professional contexts". I am very surprised to read this. Where is discriminatory behavior tolerated? Am I misunderstanding this statement? Do the authors mean that, even though most institutions have rules against this, it still occurs? Do the authors argue that psychology is relatively more hostile to women and URM compared to other scientific fields, or non-university jobs, and therefore they prefer to work elsewhere?

On page 12 the authors are positive about cross-cultural work ("descriptive studies document how personality traits vary across nations (e.g., Rentfrow et al., 2013)") but on page 21 they are not. It seems the statement on page 12 should be nuanced more than it is.

It is surprising the section on page 22 does not discuss the opportunities we have realized in online conferences. The authors write "by attending conferences such as the American Arab Middle Eastern and North African Psychological Association's" but in these times a better recommendation would be to stop flying across the world, and embrace the recent experiences we have had with online conferences – it could be acknowledged this already allowed more diversity (at a lower environmental cost) in conferences in the last months. This is also important for young families.

In the conclusion, the authors again state that personality psychologists lead the credibility revolution. I don't want to force the authors to move beyond their speculations per se, but I wonder what would happen if the 50 most cited meta-science papers were selected, and the field of origin of the author(s) would be coded. How many would be personality psychologists? Or would the authors prefer to use a different definition of 'leading'?

The sentence "Some may wonder whether they should follow in our footsteps -- what is the cost of increasing rigor?" is very peculiar. How is a cognitive psychologist supposed to read this, for example? Or even worse: A perception researcher?

The statement: "We may indeed be more boring, but what we gain from rigor, in terms of scientific truth, far outweighs that cost." lacks nuance. One of the criticism on CERN is that it is incredibly rigorous, but also has produced very little since the Higgs boson at great cost. The question is not if we need rigor, but how much we need. It's like a car: if you drive 2 miles an hour you won't get into an accident but you also will not get anywhere. A real cost benefit analysis would be a bit more complex than the authors have provided here – and as I assume all are personality psychologists, I recommend you check you own biases and exert some self-criticism. The choice is not as dichotomous as in the sentence "If psychologists eschew rigor for magical thinking and fail to embrace transparency, the entire enterprise of psychological research may decline in its influence altogether, even to the point of irrelevance." It's not a choice between no speed limit versus 50 mph, but a choice between 20 mph or

50 mph. We can not really make this cost-benefit analysis – but one needs to consider the possibility that personality psychology is too boring.

Minor:

The last paragraph on page 15 did not really add anything – it was not clear to me what the point of it was. We have fully reproducible manuscripts, and these will become the default at some point in time. What is the problem beyond this?

Check for double spaces.

Although the section on causality is fair, I think I can guess who wrote it, and it has substantial overlap with other papers that have recently been written on this (e.g., the discussion on Granger causality was already a bit niche in the other paper it appeared it, but to repeat it in this paper is probably not the level at which DAG's should be discussed. This section also has a lot more references than other sections, but that was actually good.

On page 21 it can be acknowledged that some organizations have been documenting demographics to measure diversity, as have some review portals – the recommendations the authors make is not that novel, and work has started on this. The challenge in this respect is often privacy law (e.g., at organizations).

The sentence "For example, in our teaching and knowledge mobilization, we could take care to avoid characterizing certain traits as wholly adaptive (or maladaptive) without communicating that social and contextual factors are necessary for understanding what is considered "adaptive" or not" was not clear to me – I think some extra context is needed to clarify it.

The final sentence "So, let's get to work." would have been more appropriate a decade ago.

When you upload supplementary material such as an excel spreadsheet, it is advisable to remove personal information that office documents otherwise come with.