PERSONALITY Science

PsychOpen GOLD

Comment [Invited Paper]

Check for updates

Open Commentaries to "Ten Steps Toward a Better Personality Science: How Quality may be Rewarded More in Research Evaluation" (Leising et al.)

Personality Science, 2022, Vol. 3, Article e9227, https://doi.org/10.5964/ps.9227

Published (VoR): 2022-05-06

Handling Editor: Mario Gollwitzer, Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Related: This article is part of the Theme Bundle "Rewarding Research Quality in Personality Science", consisting of a target article (https://doi.org/10.5964/ps.6029), open peer commentaries (https://doi.org/10.5964/ps.9227), and a rejoinder (https://doi.org/10.5964/ps.7961).

Abstract

The current collection comprises several comments to "Leising, D., Thielmann, I. Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science: How quality may be rewarded more in research evaluation" (https://doi.org/10.5964/ps.6029). In their target paper, Leising et al. presented a number of steps that personality researchers (and others) may take to improve the scientific standards in their field. Leising et al. answer to these comments in a rejoinder (https://doi.org/10.5964/ps.7961).

Keywords

personality science, scientific standards, research quality, research evaluation

1 – Comment

What the Hell Is Good Science? Introduction to the Theme Bundle

Mario Gollwitzer[1]

[1] *Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany.*

Reading the target paper of this "Theme Bundle" (Leising et al., 2022), steering it through an extensive peer-review process, collecting commentaries and reading them too, and finally writing the present text belonged to the most pleasurable professional experiences I had in 2021. The target paper is clearly thought-provoking, intellectually rich, and provocative in many aspects. It was foreseeable right from the start, and even more so after receiving the reviewers' comments to the original version of the paper, that the text would stir a range of reactions among its readership and motivate many people to voice their consent or dissent. This is why the editor-in-chief at Personality Science, John Rauthmann, and I decided to solicit both invited and open commentaries and give the authors of the target paper the opportunity to respond to these. The package that resulted is, in my view, a "must-read" for every scholar interested in personality science and psychology more broadly. Although many commentators disagree with some (or many) of the recommendations made by Leising et al. (2022; i.e., the "ten steps towards better personality science"), they all agree that personality science can (and should) improve, that (implicit or explicit) incentive structures are relevant for such improvement, and that it is very difficult to define what "improve" actually means. There is no ultimate criterion for "good" science – the "ought" state. Yet, fueled by the replication debate in psychology, we feel that there is a discrepancy between the ought and the is.

Leising et al. (2022) argue that consensus-building (Steps 1-5) and methodological rigor including openness and transparency (Steps 6-10) may be a way to reduce the is-ought discrepancy, and especially the "consensus-building" argument provoked some well-grounded criticism among the reviewers of the original paper and other commentators (e.g., see the comments by Asendorpf & Gebauer; Corker; Denissen & Sijtsma; Fernandes & Aharoni; Hagemann; Hilbig et al.; Hogan et al.). A problem that these authors have with the "consensus-building" idea is that 1) consensus is prone to biases (e.g., it favors the mainstream), that 2) the common ground for consensus-building may be too shaky (i.e., a "garbage-in-garbage-out" problem; see the comment by Corker), and that 3) consensus penalizes dissent and, thus, a healthy evolution of science. Some commentators add an important flavor to the first point by arguing that each consensus-building process is likely to be influenced by power and status hierarchies in the scientific community (see the comments by Adler; Beck et al.; Fedorenko et al.; Galang & Morales; Klimstra; McLean & Syed). Without proper safeguards against a violation

of procedural justice standards (e.g., Leventhal, 1980), those who are high in power and status will define the consensus. The process itself will lack inclusiveness, and the outcome will lack representativeness – a point that Naomi Oreskes (2019) also makes in her book "Why trust science?", a book that is of prime relevance for our present discourse.

Other comments focus on the reward scheme that Leising et al. are proposing (e.g., Asendorpf & Gebauer; Beck et al.; Friedman; Schmitt). The argument here is that such reward schemes disqualify certain scientific approaches to personality (e.g., qualitative research; see Klimstra; McLean & Syed) and that it is the wrong criteria that are eventually being rewarded, such as submissiveness, closed-mindedness, and compliance with the mainstream (see Asendorpf & Gebauer). As good as Leising et al.'s intentions are, the idea of a standardized reward scheme for "good science" may ultimately aggravate the problems we have with quantitative indicators for scientific "quality" (such as the infamous h-index) instead of solving them. I agree with these arguments, and I would like to add an aspect that one of the reviewers of the initial manuscript (Rainer Bromme, who explicitly agreed on having his name disclosed here) had touched upon. Referring again to Oreskes (2019) as well as microbiologist/science philosopher Ludwik Fleck (1980), Bromme noted that the social process of doing science is far more complex than implied by the authors of the target article. Most researchers are (hopefully) not merely interested in increasing their h-index or accumulating reward points – they are intrinsically motivated to work on something important, to learn with and from other people, and to find out something relevant. For the sake of making all this possible, they even behave altruistically: they agree to review other people's work, they serve as committee chairs, and they devote their time to self-administration (see the comment by Schmitt).

As any other social system, academia already has implicit ("built-in") incentive structures that reward altruism and punish egoism. Cynics may snort and provide exemplars of narcissistic, egotistic individuals who went far in their academic careers, but I am convinced that these exemplars are exceptions to the rule. Most scholars would probably argue that the partners they particularly enjoy collaborating with are helpful, supportive, diligent, and trustworthy. Mutual trust is likely a much stronger predictor of "doing good science" than extrinsic rewards (Altenmüller & Gollwitzer, 2022), so instead of reinventing grading schemes and replacing one quantitative indicator by another, we might want to think more carefully about a system in which both self-correction (and openness to each other's criticism) and mutual trust-building in science can be optimized.

That said, there are many ways in which doing good science can be facilitated by methodology and infrastructure: this includes not only platforms (repositories) for sharing our data, materials, and papers so that they can be much more easily accessed by others ("open science"), but also databases that can provide us with up-to-date knowledge

about the psychometric properties (including measurement invariance) of methods and measures, as Horstmann and Ziegler as well as Mazei et al. mention in their comments.

A final point mentioned in some comments is that many of the "ten steps" described by Leising et al. are exclusively applicable to quantitative, variable-centered approaches to personality science and to "contexts of verification" (see Hogan et al.). Some commentators rightfully note that this should by no means imply that these approaches are in any way more valid or esteemed than, for instance, person-centered, idiographic, and/or qualitative approaches (see Dunlop, Klimstra). I believe that Leising et al. agree with this, but it deserves being mentioned explicitly here.

In sum, I hope that readers will enjoy digesting this Theme Bundle and that the present conversation will be the start of a lively, engaged discussion within personality science and beyond. To repeat, it was an extreme pleasure and honor being guest editor of this edition, and I thank the editors of Personality Science, the authors of the target article, and all commentators for this delightful opportunity.

References

Altenmüller, M. S., & Gollwitzer, M. (2022). Prosociality in science. *Current Opinion in Psychology, 43*, 284–288. https://doi.org/10.1016/j.copsyc.2021.08.011

Fleck, L. (1980). *Entstehung und Entwicklung einer wissenschaftlichen Tatsache* [Genesis and development of a scientific fact]. Suhrkamp. (Original work published 1935)

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27-54). Plenum. https://doi.org/10.1007/978-1-4613-3087-5_2

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

2 – Comment

Good Intentions – Unfortunate Side Effects

Jens B. Asendorpf[1], Jochen E. Gebauer[2,3]

[1] *Department of Psychology, Humboldt University Berlin, Berlin, Germany.* **[2]** *Department of Psychology, University of Mannheim, Mannheim, Germany.* **[3]** *Department of Psychology, University of Copenhagen, Copenhagen, Denmark.*

Great innovation only happens when people aren't
afraid to do things differently.
-Georg Cantor

The authors of the target article had good intentions. They wanted to advance personality science by implementing 10 "steps" designed to boost the trustworthiness of research.

We are sympathetic with most of these steps. We strongly disagree, however, with 1) the authors' overly strong focus on consensus about important research questions, and 2) their proposed reward scheme for "good scientific work." This scheme directly translates into a social reward system for young scientists, thereby shaping the norms of scientific culture. The scheme rewards swimming with the mainstream and penalizes swimming off or against the mainstream. In effect, minority views, creativity, and innovation will be discouraged. We assume that the members of the task force do not intended to penalize minority views, creativity, and innovation and, thus, we encourage them to consider not only the potential merits of their proposal but also potential, negative side-effects on creativity and innovation.

Contrast the target article's reward scheme with the first three evaluation criteria for a successful ERC grant proposal (European Research Council, 2021, p. 33):

1. To what extent does the proposed research address important challenges?
2. To what extent are the objectives ambitious and beyond the state of the art (e.g., novel concepts and approaches or development between or across disciplines)?
3. To what extent is the research high-risk gain (i.e., if successful the payoffs will be very significant, but there is a high risk that the research project does not entirely fulfill its aims)?

The target article's reward scheme would reward research that fulfills those criteria with zero points (unless the applicant was not the first to identify the "important challenge," but borrowed it from some consensus list of important challenges). A small variation on a mainstream topic devoid of any new idea, by contrast, would gain 10 points if related to a consensus in a certain area, published open access and pre-registered with sufficient power – simple criteria that can be easily achieved. The target article briefly mentions creativity and innovation in passing, but these two criteria did not make it in the authors' scheme. They defend their neglect of those cardinal criteria with the difficulty of judging creativity and innovation. But difficulty cannot be an excuse for

proposing an unbalanced reward system, especially if the authors expect their reward scheme to shape scientific culture, setting norms for what kind of research should be admired and what kind should be frowned upon.

As the past 15 years have shown, quick and dirty measures of research quality such as the Hirsch index can gain a life of their own and mislead young scientists. Many tried to maximize their h-index through heavy conference tourism, excessive networking, and inflation of co-authorships and cross-citations, all of which kept them away from their core tasks as scientists.

A Personality Perspective

Adopting the proposed reward scheme without in-built corrections for creative and innovative contributions would result in stagnation. The proposed scheme would attract extrinsically motivated students high in agreeableness and low in openness to new experience who are naturally inclined to swim with the mainstream (Eck & Gebauer, 2021) and, thus, fulfil the ten steps at the cost of creativity and innovation. At the same time, the proposed reward scheme would discourage students who are intrinsically motivated for science because of high openness to experience and a strong curiosity motive. As these students like to challenge the status quo, they are often judged as low in agreeableness, which increases the risk to be discouraged even more, leave science, and create a start-up.

Empirical studies of entrepreneurs (founders of new businesses) have shown that they can be characterized by strong curiosity motivation, strong intrinsic motivation, strong need for autonomy, identification with a long-term goal, tolerance to ambiguity, and a characteristic Big Five profile consisting of high openness, high conscientiousness, low neuroticism and low agreeableness (Rauch & Frese, 2007; Zhao & Seibert, 2006). Science need such people as much as striving economies do. Science cannot compete with venture capital regarding salary but this is not critical if science provides a platform for expressing curiosity and enthusiasm and tolerates people's corners and edges.

An Evolutionary Perspective

Any evolution, whether genetic (Darwin, 1859) or cultural (Boyd & Richerson, 1985), including the evolution of the sciences, is based on variation and selection. Successful adaptation to an ever-changing environment requires sufficient variation, including variants that are not well adapted at present. If the conditions for successful adaptation change, a few of them will be better adapted in the future than the presently well-adapted. Genetic evolution requires a broad repertory of genetic variants maintained by, for example, mutation and sexual recombination. Cultural evolution requires a broad repertory of cultural variants maintained by, for example, novel ideas, new combinations of ideas through communication, and misunderstanding in communication. Without such

variation around the mainstream, a research area tends to reiterate its status quo and risks to end up in a dead end.

References

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

Eck, J., & Gebauer, J. E. (2021). A sociocultural norm perspective on Big Five prediction. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000387

European Research Council. (2021, July 12). *ERC Work Programme 2021*. European Commission. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2021/wp_horizon-erc-2021_en.pdf

Rauch, A., & Frese, M. (2007). Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of Work and Organizational Psychology, 16*(4), 353–385. https://doi.org/10.1080/13594320701595438

Zhao, H., & Seibert, S. E. (2006). The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *The Journal of Applied Psychology, 91*(2), 259–271. https://doi.org/10.1037/0021-9010.91.2.259

3 – Comment

Consensus and Diversity—A Comment on Leising et al.

Dirk Hagemann[1]

[1] *Department of Psychology, Universität Heidelberg, Heidelberg, Germany.*

Leising et al. (2022) report the consensual opinion of five psychologists with a clear mission: "Scientists need to be increasingly rewarded for doing good work, not just lots of work." Because nobody can object to the proposition "quality matters", this mission statement is indisputable. However, when it comes to specify what "good work" really is and how good work may be encouraged, things become more complicated and less clear-cut. The authors base their consideration on writings of Karl Popper, including key concepts such as cumulative progress, formalization, and of course, falsification. There is much to say about why science is not only an accumulative business (Kuhn, 1962), why formalization of theory and hypotheses has limitations (Heylighen, 1999), and

why physicists do not believe in falsification (Singham, 2020). Due to space limitations, however, I will restrict my comment to what the authors call the building of "consensus".

According to Leising et al. (2022), fostering "consensus" regarding research goals, terminology, measurement practices, data handling, and the current state of theory and evidence are the first five steps toward a better personality science. When they introduce "consensus" as an essential element of good science in their section "What is good research?", they refer to work of Oreskes (2020). Because the formation of "scientific consensus" has been a target of intense studies in the sociology, history, and philosophy of science (Oreskes, 2020; Shwed & Bearman, 2010), it is worth to have a closer look at this concept.

In the field of science studies, "scientific consensus" is a term that refers to the transformation of empirical propositions into facts by a scientific community (e.g., Shwed & Bearman, 2010). In the typical case, some experts analyze the status of a scientific domain and report their conclusions. One prominent example is a report by Oreskes (2004), who analyzed the validity of an IPCC report from 2001 that stated a consensus on anthropogenic climate change. The repeated questioning of this consensus by some policymakers motivated her research. Oreskes (2004) classified 928 abstracts that contained the keyword "global climate change" into several groups, ranging from endorsement over indifference up to rejection of the consensus position. She found no single abstract that disagreed with the consensus position and concluded that "scientists publishing in the peer-reviewed literature agree with IPCC" (Oreskes, 2004, p. 1686). This "scientific consensus" is why the anthropogenic climate change may be considered to be a scientific "fact" (Shwed & Bearman, 2010) or scientific "knowledge" (Oreskes, 2020). Aside the proposition "human activity is producing global climate change ", other propositions that have become target for consensus research are "smoking causes cancer", "vaccinations cause autism", or "there are gravitational waves", just to give some examples (for these and further cases, see Oreskes, 2020; Shwed & Bearman, 2010). In any case where Oreskes (2020) discusses the "scientific consensus", she refers to empirical propositions that reflect a positive or reliable knowledge, which is to say that these propositions can be "true" or "wrong" because they represent empirical features of the real world (of course, consensus does not imply an eternal truth because scientists may err; c.f. "instability of scientific truth" in Oreskes, 2020, p. 74).

It is obviously not possible to extend this concept of "scientific consensus" to research goals, terminology, measurement practices, and data handling. Research goals may be interesting, terminology may be useful, and measurement practice and data handling may be valid. However, the attribute "true" is not in stock because none of these items can ever be scientific facts in the sense of an empirical proposition (see Shwed & Bearman, 2010). No matter if we belief that the first four steps of Leising et al. (2022) are useful for making personality science better or not, Oreskes (2020) does not provide the epistemological grounding.

Quite the contrary may be true. In order to integrate empirical findings from different studies into a reliable scientific fact or trustworthy knowledge, the integrated evidence and the involved scientific community needs to satisfy several criteria of good science. Oreskes (2020, pp. 143-144) lists five of them: (1) “Do the individuals in the community bring to bear different perspectives? Do they represent a range of perspectives in terms of ideas, theoretic commitment, methodological preferences, and personal values?” (2) “Have different methods be applied and diverse lines of evidence considered?” (3) “Has there been ample opportunity for dissenting views to be heard, considered, and weighted?” (4) “Is the community open to new information and able to be self-critical?” (4) “Is the community demographically diverse: in terms of age, gender, race, ethnicity, sexuality, country of origin, and the like?” Only when the answers to these questions is "yes" should we trust the consensus.

The simple reason why Oreskes (2020) demands diversity to such a large degree is that each individual research finding and each individual scientist is inevitably biased—and diversity is the method to counteract it. Of course, these criteria of good science echo many tenets in the methodology of psychology, such as using different operations and methods to counteract the mono-operation bias and the mono-method bias and to establish convergent and discriminant validity (Shadish et al., 2002). In any case, these criteria of Oreskes (2020) appear rather to be an anti-thesis to the first four steps of Leising et al. (2022) than an epistemological support for building “consensus” beyond the empirical facts.

To conclude my thoughts about the proposal of Leising et al. (2022), I think these authors have hijacked Oreskes' (2020) notion of "consensus" to give their proposal the air of epistemological grounding, but they really missed the point: Diversity of concepts and methods is the prerequisite for a consensus about empirical findings to be trustworthy. Their ten steps and in particular their reward scheme emphasize agreement and consensus but downplay or ignore the values of disagreement and diversity. Therefore, the proposal of Leising et al. (2022) will need further elaboration before it may help to establish a better personality science.

References

Heylighen, F. (1999). Advantages and limitations of formal expressions. *Foundations of Science, 4*(1), 25–56. https://doi.org/10.1023/A:1009686703349

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

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2,* Article e6029. https://doi.org/10.5964/ps.6029

Oreskes, N. (2004). The scientific consensus on climate change. *Science, 306*(5702), 1686. https://doi.org/10.1126/science.1103618

Oreskes, N. (2020). *Why trust science?* Princeton University Press.

Shadish, W. R., Cook, T. C., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin Company.

Shwed, U., & Bearman, P. S. (2010). The temporal structure of scientific consensus formation. *American Sociological Review, 75*(6), 817–840. https://doi.org/10.1177/0003122410388488

Singham, M. (2020). *The great paradox of science.* Oxford University Press.

4 – Comment

Consensus Is not the Cure; It's Part of the Disease

Benjamin E. Hilbig[1], Morten Moshagen[2], Ingo Zettler[3]

[1] *Department of Psychology, University of Koblenz-Landau, Landau, Germany.* **[2]** *Institute of Psychology and Education, Ulm University, Ulm, Germany.* **[3]** *Department of Psychology and Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Copenhagen, Denmark.*

Leising, Thielmann, Glöckner, Gärtner, and Schönbrodt (2022) suggest better (personality) science can be achieved through greater consensus-building and improving the credibility of empirical research. Whereas the latter is a perfectly reasonable goal (even if some specifics remain debatable), the former strikes us as questionable to put it mildly. Specifically, Leising et al. (2022) call for and seek to heavy-handedly reward more consensus on (a) what we know and what we ought to find out next (their Steps 1 and 5) as well as (b) terminology, measurement, and analytical approaches (their Steps 2-4). Crucially, although they acknowledge that consensus must be seen as preliminary and envision consensus on the basis of a fair, transparent, and evidence-based process, their essential proposal – manifested as a reward-point-table – asks only whether there is self-ascribed[1] consensus.

First off, it is worth considering the intentions apparently feeding this call for consensus by Leising et al. (2022). Undeniably, lack of theoretical integration and too few serious efforts to pit competing theories against each other with the best available data and analyses techniques hinder progress. More obviously still, inconsistencies and idiosyncrasies in terminology, measurement, and analytical strategies can hinder the establishment of reliable evidence. Personality science traditionally rewards construct inflation (continuous invention of ever more constructs and operationalizations with

1) Without a well-defined population it is actually not even possible to determine whether one has the votes necessary and sufficient to call something consensus. This is not nitpicking but a hurdle that Leising et al.'s suggestion would need to overcome to avoid remaining entirely impractical and ultimately arbitrary.

dubious distinctiveness or incremental usefulness) over theoretical specification or integration and has not sought let alone found any effective means of preventing jingle-jangle-fallacies. Vague definitions render traits barely testable (beyond some correlation with something similar) and allow for multiple non-equivalent operationalizations of what is (often inductively) given the same label. There is no denying that (personality) science is beset by systematic problems along these lines. As such, the diagnosis by Leising et al. (2022) is spot-on.

But consensus is not the cure. Consensus is, by definition, a matter of majority belief or preference. It is ultimately whatever most agree to, for whatever reason, and thus not necessarily aligned with any logic or evidence. In science, however, authority is not wielded by some (self-proclaimed) majority but by logic and evidence, albeit preliminary and uncertain. To promote the role of consensus, Leising et al. (2022) go so far as to state that "facts are claims about which agreement has been reached among scientists in the respective field" (manuscript p. 5). They are not. Facts are consistent, independently verified, and reliable empirical observations – or, albeit more tentatively, explanations unequivocally supported by such observations. Climate change is not a fact because most climate scientists believe in it, but because of evidence[2] .

Indeed, in line with Kuhn's very argument, major scientific progress typically requires upending the consensus (normal science), rather than cementing its power through still more rewards. The history of science is replete with examples that clearly demonstrate how the consensus produces drag or inertia, flat-out denial, or worse. The consensus shouted down Wegener's theory of continental drift despite consistent and strong evidence; the behaviorist consensus upheld proliferation of impoverished associationist theorizing for decades despite consistent and strong evidence for the role of insight and expectation/anticipation in learning and problem solving; and if the replication crisis has taught us anything then how much it takes to question let alone upend (allegedly) "consensual" knowledge.

The deck is already stacked against dissent and counterevidence to widely held views. For an individual scientist's career, normal science is a safer bet whereas rocking the boat is risky at best. Anomalies are often kept under the rug by the gatekeepers and stewards of normal science (editors and reviewers) or, if occasionally published, met with some

2) Note that we in no way question Oreskes' main arguments. For one, whenever public policy is hindered by a small, vocal minority who simply claims that there is no consensus among scientists so as to hinder certain policies, it is undoubtedly necessary and responsible to point out that this claim is false (as Oreskes did in her seminal 2004 paper). Moreover, it is out of the question that consensus in a diverse, non-defensive, and self-critical scientific community ought to be a prime reason for non-experts (e.g. politicians) to trust science (rather than special interest groups) when making hugely consequential, life-or-death decisions (as is Oreskes core claim in her 2020 book that Leising et al. so prominently cite). Since non-experts cannot and should not be expected to form a judgment on the extent and (un)certainty of the evidence, their best bet will often be to heuristically approximate it through the extent of consensus among the experts. None of these valid arguments, however, imply that science becomes any "better" by seeking and promoting consensus within the scientific community.

combination of pushback, defensive downplaying, and denial. In (personality) science, the symptoms as diagnosed by Leising et al. (2022) are upheld by the consensus, albeit often implicit. If anything, our field needs to shield critical arguments and "inconvenient" evidence from the consensus, not hand the consensus bigger guns.

By all means, let us promote and reward publications that serve theory specification and integration, that actively seek to apply Occam's razor so the field may focus on the constructs, models, and operationalizations we (really) need, that expose shortcomings in our terminology and frameworks and offer up alternatives, that develop a single, authoritative means of measuring a (well-defined) construct – or any combination of these. But let us never vote on it.

Any of the above can be achieved by any number of individuals. Their faction size does not and should not matter, only their logic and evidence. Anyone who would rather be a majority whip is free to go into politics. Anyone who would rather abide by shared beliefs is free to practice religion. These are the dominions of consensus, lending legitimacy to politics and authority to religion. But unlike science, neither politics nor religion are tasked with nor remotely suited for closing in on the truth. The latter goal, distal and formidable though it is, will best be served by a (personality) science that, above all, rewards challenging theories and alleged facts – consensual or not – through logic and evidence.

References

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Oreskes, N. (2004). The scientific consensus on climate change. *Science, 306*(5702), 1686–1686. https://doi.org/10.1126/science.1103618

Oreskes, N. (2020). Why trust science? Princeton, New Jersey: Princeton University Press.

5 – Comment

Six Reactions to 10 Steps Toward a Better Personality Science

Robert Hogan[1], Peter Harms[2], Ryne A. Sherman[1]

[1] *Hogan Assessment System, Tulsa, OK, USA.* **[2]** *Department of Management, University of Alabama, Tuscaloosa, AL, USA.*

We are deeply invested in the personality research enterprise; we found the "10 Steps" paper to be interesting and informative, but it also raised six issues as follows:

1. **History of science.** The 10 Steps paper (Leising et al. 2022) implicitly assumes a version of history of science that is widely shared but wrong. The assumption is that scientists make fundamental discoveries and then engineers turn them into applications, but the reverse is more often true. Starting with Archimedes of Syracuse (287-212), engineers solve practical problems, and scientists then study those solutions. Knowledge flows from applications to science; few discoveries in pure science ever led to applications. Personality psychology started in Austrian psychiatry and German military psychology; it began as an effort to solve problems about people and organizations and we believe that, to the degree that personality research ignores applications, it is doomed to irrelevance.
2. **Context of discovery vs. context of verification.** Reichenbach (1951) famously distinguished between "the context of discovery" and "the context of verification." The former concerns where ideas come from; the latter concerns how they are evaluated. The former is more important than the latter—with no ideas to verify, there can be no research. The 10 Steps paper focuses almost exclusively on the context of verification—which is a kind of under-laborer activity. Future progress lies hidden somewhere in the context of discovery.
3. **Anchoring ourselves to the FFM**. The 10 Steps paper assumes that the Five-Factor Model represents the true (phenotypic and genotypic) structure of personality, but we believe this view is insupportable. There is plenty of evidence showing that there are many personality factors relevant to important life outcomes beyond the FFM (e.g., Ashton et al., 2004; Mõttus et al., 2020; Paunonen & Jackson, 2000; Wood et al., 2010). In our own research, starting with pools of personality descriptors and using predictive validity to inform our model, we find evidence of 7 primary factors. Consider the following: leadership is the most important problem in human affairs. When good leadership is in place, countries and institutions flourish; when bad leadership is in place (Belarus, Venezuela, Myanmar), everyone suffers. Leadership is a function of personality and the two most important personality variables

predicting leadership are Ambition and Humility, neither of which can be readily identified in the FFM.

4. **Dangers of consensus.** When we prioritize consensus, we get stuck on suboptimal solutions like the Big Five, the Dark Triad (Paulhus & Williams, 2002), and the Motivation to Lead scale (Chan & Drasgow, 2001). We then spend vast amounts of time trying to improve these models. As noted above, it is (or should be) obvious that there is content outside of these consensus-based taxonomies—which themselves are under-theorized (or completely atheoretical). But as the 10 Steps paper suggests, once we settle on such suboptimal solutions, the only viable path for future research is endless grinds looking for equivalence or ways to excuse nonequivalence. Consensus focuses attention on the context of verification and shuts down further conceptual inquiry. But, as Alfred North Whitehead once quipped, "To set limits of speculation is treason to the future." Consensus drives "normal science" but constrains new paradigm shifts and sets limits to speculation. Finally, the 10 Steps paper does not lay out who will make the consensus picks or by what means. Designating certain individuals as arbiters of truth, regardless of their expertise, risks endangering the freedom and egalitarian norms that define the scientific enterprise.
5. **Measuring entities vs. predicting outcomes.** Modern personality research (and the 10 Steps paper) assumes that the goal of personality assessment is to measure entities—usually traits—based on the view that traits define the structure of personality. We believe that these views (i.e., that the goal of assessment is to measure traits and that traits define the structure of personality) lead to hopeless metaphysical confusion. The problem starts by defining traits as both (a) recurring consistencies in behavior, and (b) unobserved latent entities. In our view, the goal of assessment is to predict recurring consistencies in behavior (traits or, in everyday language, "reputation") and important life outcomes. The effort to measure traits defined as unobserved latent entities leads to two further problems: (1) How do we know when we have finally measured these latent entities (and no others), and (2) what will we do then? If we start with the (applied) view that the goal of assessment is to predict recurring consistencies in important behavior and outcomes, we can avoid speculating about unobserved latent entities. We can also leave the process of investigating neuropsychic entities to real neuroscientists while we study useful outcomes.
6. **Methodological idealism.** During the Middle Ages, theologians were preoccupied with finding airtight arguments for the existence of God. This led to a careful evaluation of reasoning processes and the identification of logical fallacies. One of the most famous was the fallacy of dogmatic methodism: the idea that, if one applies the appropriate methods to analyzing a problem, truth will inexorably emerge. If you

are unable to find the truth, then you used the wrong methods. The 10 Steps paper suffers from methodological idealism.

In conclusion, although we support many of the ideals and goals of the 10 Steps paper, we are somewhat skeptical of their proposed solutions. The authors suggest that the first step should be to establish common goals for the field. We agree, but we would argue that the grand problem for our discipline concerns predicting significant behavioral outcomes (i.e., longevity, status, relationship quality, well-being). If we keep this goal in mind, then progress will follow.

References

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blass, L., Boies, K., & De Raad, B. (2004). A six-factor structure of personality descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Chan, K.-Y., & Drasgow, F. (2001). Toward a theory of individual differences and leadership: Understanding the motivation to lead. *The Journal of Applied Psychology, 86*(3), 481–498. https://doi.org/10.1037/0021-9010.86.3.481

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Mõttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality, 34*(6), 1175–1201.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism and psychopathy. *Journal of Research in Personality, 36*((6)), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the Big Five? Plenty! *Journal of Personality, 68*(5), 821–835. https://doi.org/10.1111/1467-6494.00117

Reichenbach, H. (1951). *The rise of scientific philosophy*. University of California Press.

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality, 44*(2), 258–272. https://doi.org/10.1016/j.jrp.2010.02.003

6 – Comment

Improving Research Quality: The Roles of the Timing and Scope of Changes in the Incentive Structure and the Quality of Committee Work

Manfred Schmitt [1]

[1] *Department of Psychology, Universität Koblenz-Landau, Koblenz/Landau, Germany.*

I enjoyed reading this thoughtful and thought-provoking target article (Leising et al., 2022), and I wholeheartedly agree with the authors' overall goal of improving the quality of personality research. Who would not agree with this goal? I also appreciate the steps they propose to move toward it. None of them are new, and some have been requested for decades. The authors are aware of this. The value added by the target article to previous requests is that they are all compiled into one paper, which I hope will generate synergistic impact. I also agree with the authors' belief that a change in incentives will be the most effective way to improve research quality. Again, the proposal to reward research quality rather than the number of publications is not new. But if the steps to better research quality have been known for so long, why are we still seeing so many papers of mediocre quality? In my view, two factors have contributed importantly to previous failures at improving research quality: (1) the timing and scope of changes in the incentive structure and (2) the insufficient quality of committee work. Considering these factors will be crucial for preventing continued failures.

The Timing and Scope of Changes in the Incentive Structure

Replacing the currently predominant quantity standard with quality standards comes at a cost. The authors elaborate on two of them: the time and energy needed to conduct high-quality research. They only briefly mention two risks that I consider major: bias and unfairness. Individual researchers, research groups, departments, and fields (personality) that commit to high-quality research will be penalized if other researchers, research groups, departments, and fields do not also make the same shift simultaneously. Otherwise, bias in performance assessments and unfairness in hiring, promotion decisions, and the allocation of grant money cannot be avoided. The issue is similar to the one faced by attempts to deflate high school and university grades, which become inflated over time if they serve as admission criteria for higher education or master programs. Deflating grades in some schools or bachelor programs but not in others will disadvantage students from the former and give an advantage to students from the latter. The same would happen if the shift from quantity to quality occurred in only some parts of an academic community. I am afraid that it will not be easy to implement the proposed changes

everywhere, as mandatory, and at the same time. Therefore, advocating concerted efforts in one field (personality) might not work. Rather, beginning at a specific point in time, an entire discipline (psychology) needs to commit itself to the proposed changes and pass them as binding. In addition, I wonder about the necessity of implementing the proposed changes internationally. Researchers apply for jobs in other countries, and researchers from several countries compete for the same grant money (e.g., ERC funding). What is needed in addition to the suggestions made in the target article are multidisciplinary and international task forces, which first need to reach a consensus on quality standards. The target article could serve as an excellent resource for this consensus-building process. Next, a road map for the implementation of the new standards needs to be developed. This road map would need to be formally accepted by journals, academic societies, universities, funding agencies, and research governors on the level of states and nations. This is a huge task that has a long way to go.

The Quality of Committee Work

High-quality research makes little difference for career success if it is not recognized and appreciated as such by review boards and search and promotion committees. The perception and valid appreciation of research quality by assessors and decision-makers is more likely to be achieved if the research-related reward structure is flanked by changes in the rewards for committee work. I am saying this because I have come across many failures when shifts such as the ones recommended in the target article were attempted. I have chaired dozens of search, evaluation, and promotion committees and served as a reviewer on a large number of selection, promotion, and funding committees. Almost all the committees I chaired or served on began their work with the agreement to prioritize quality over quantity. When it came to making decisions (ranking applicants or grant proposals), this agreement was often forgotten. When time was short or ran out, which happened regularly, committee members—myself included I must shamefully confess—began counting publications, citations, awards, and grant money. We did not do this because we thought it would result in the best possible decisions but simply because we did not have enough time to read dozens of papers and assess their quality. Counting papers is easy and quick, whereas evaluating their quality is difficult and time-consuming. Because junior and senior researchers alike are chronically short on time, falling back on quantitative heuristics is tempting and happens. How can this be avoided? Self-assessment of research quality according to the proposed reward scheme may work to some extent. Yet, quality criteria, such as creativity, relevance, fit with the research profile of a department, and other criteria, require additional assessment by reviewers and committee members. My feeling is that replacing quantitative heuristics with profound evaluations of research quality will happen only if we radically change the procedural rules of committees and the reward structure for committee work. In order to be fair and effective, rewards for committee work will have to increase dramatical-

ly and go far beyond the rewards we are used to. In addition to rewards, many committee members need more training and guidance than they currently have. Students and young researchers become members of powerful committees without adequate preparation for their task. Assessing research quality requires substantial expertise and skill. These need to be taught and learned. Both tasks—teaching and learning to assess research quality—become attractive when they are rewarded.

References

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

7 – Comment

A New Academic Incentive Structure: Does It Fit the Psychology of Human Motives?

Jaap J. A. Denissen[1], Klaas Sijtsma[2]

[1] *Department of Developmental Psychology, Utrecht University, Utrecht, The Netherlands.* **[2]** *Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands.*

> Science would be ruined if (like sports) it were to put competition above everything else.
> —Benoit Mandelbrot (cited in Gleick, 1987, p. 70).

Leising, Thielmann, Glöckner, Gärtner, and Schönbrodt (2022) diagnose an important problem that plagues psychology: An incentive structure that places much value on researchers' quantity of scientific output. The authors propose a novel structure: Scientists can earn points by investing in more consensus-based, cumulative, open, and replicable science. The authors predict that such structure might result in an improvement of the scientific enterprise because it moves away from the current incentivization of speed and competition, which plausibly underlies much of the crisis of faith in the replicability of research findings. Like any prediction, this one also is liable to empirical scrutiny.

An incentive structure not only needs to produce the right collective outcomes but also fit the psychology of human motives. There are, broadly speaking, two major and partly overlapping basic motive taxonomies. Self-determination theory features a need for affiliation (which it labels relatedness), achievement (which it labels competence), and autonomy. Classical motive theory also features the first two of these needs, but adds the need for power, called the need for status by Anderson et al. (2015). Both tax-

onomies allow for individual differences, though perhaps to differing extents (Soenens, Vansteenkiste, & Van Petegem, 2015). Here, we take an eclectic, broad-brush approach and assume four basic motives (affiliation, achievement, status, and autonomy) that differ between individuals. So how does the proposal of Leising et al. (2022) stack up?

Through a motivational lens, Leising et al.'s (2022) analysis might be grounded in a working hypothesis that the status motive has become too dominant in science. In their search for fame and recognition, scientists do not cooperate as much as they should, and invest in quantity and speed rather than replicability and quality. In extreme cases, this can produce tension with the need for achievement, which is fulfilled if people truly master something new and complicated – for which there might be too little time in a hyper-competitive regime. The status quote also threatens the need for affiliation if it sacrifices cooperation for competition. Leising et al. (2022) do not entirely ignore the status motive: In their proposal, researchers can still compete for points and related prestige, but they have to do so by acting responsibly and collaboratively. This requirement appeals to submission to textbook methodology and ethical principles; that is, fair play.

A first question is whether it is indeed possible to channel the status motive towards collaborative goals. While there is research that shows that "communal narcissism" can indeed be focused on prosocial outcomes (Gebauer et al., 2012), the proof of the pudding is in the eating. In addition, we can only guess to what extent truly pioneering researchers have been driven by a need for status (e.g., in the case to discover the structure of DNA)? Can we expect the same pace of risky and time-sensitive breakthroughs under a more collectivist incentive structure, which rewards people for relatively safe endeavors? The benefits of the newly proposed system might still outweigh the drawbacks, and any change of system has of course winners and losers. Still, the issue requires careful analysis and empirical evidence.

If our motivational analysis is correct, the greatest question mark is the need for autonomy. While it is true that hierarchical power structures also limit the autonomy of lower-ranked scientists, the newly proposed system appears more collectivistic. That is, it establishes a set of "consensus statements" regarding methods and theories, and scientists are highly rewarded with credit points if they conform. However, what about researchers, who prefer to work alone and/or are ahead of their time? The authors do discuss how innovation would still be possible in such a system, but without supporting evidence, we cannot be sure. Some skepticism is informed by a motivational insight from economics: Any quality indicator of work can trigger strategic behaviors that render the counting metrics themselves obsolete (this is known as Goodhart's law; Chrystal, Mizen, & Mizen, 2003). We would be concerned if collectivistic approaches suppress diversity of opinion, which runs counter to ideas about science being a Darwinian-like process in which the generation of diverse ideas is followed by consensus-based selection (Simonton, 2003). In our view, there should be ways to reward contrarian scientists that

consistently "run against the grain" – and do so with determination but without breaking the rules of the game.

This relates to our final point, which is about individual differences. We have speculated that the status quo seems well attuned to the need for status and autonomy (at least for the powerful). Switching to the new system might create a focus on affiliation (cooperation) and perhaps on research that is performed with greater precision, which would be welcomed by many. Perhaps this is a good trade-off. Still, it might be even better to create diversity in incentive structures, so that individual differences of researchers can be put to productive use. After all, it has been suggested that team diversity contributes to better outcomes (Rock & Grant, 2016). The challenge would then be to create an incentive structure that values the contribution of everyone with useful skills and potential, whether they are quirky contrarians or agreeable consensus builders, avid reviewers or bold explorers, shy background types or active spotlight seekers. Whether such a diversity-promoting incentive structure is compatible with a score-keeping system as proposed by Leising et al. (2022), without doubt also raising discussion and perhaps even conflict among contestants, remains to be seen. A personnel assessment system that discourages questionable behavior but does not get in the way of autonomy (initiative, courage, and academic freedom) sets a daunting task, and we think that personality psychologists are well equipped to contribute to solving it!

References

Anderson, C., Hildreth, J. A. D., & Howland, L. (2015). Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological Bulletin, 141*(3), 574–601. https://doi.org/10.1037/a0038781

Chrystal, K. A., Mizen, P. D., & Mizen, P. D. (2003). Goodhart's law: Its origins, meaning and implications for monetary policy. In Paul Mizen (Ed.), *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart* (Vol. 1, pp. 221-243). Edward Elgar Publishing.

Gebauer, J. E., Sedikides, C., Verplanken, B., & Maio, G. R. (2012). Communal narcissism. *Journal of Personality and Social Psychology, 103*(5), 854–878. https://doi.org/10.1037/a0029629

Gleick, J. (1987). *Chaos: Making a new science*. Vintage.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Rock, D., & Grant, H. (2016). Why diverse teams are smarter. *Harvard Business Review, 4*, 2–5.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin, 129*(4), 475–494. https://doi.org/10.1037/0033-2909.129.4.475

Soenens, B., Vansteenkiste, M., & Van Petegem, S. (2015). Let us not throw out the baby with the bathwater: Applying the principle of universalism without uniformity to autonomy-supportive and controlling parenting. *Child Development Perspectives, 9*(1), 44–49. https://doi.org/10.1111/cdep.12103

8 – Comment

In Unity There Is Strength but in Divergence, Unexpected Leaps

Sharlene Fernandes[1], Eyal Aharoni[1]

[1] *Department of Psychology, Georgia State University, Atlanta, GA, USA.*

Leising et al. (2022) suggest strategies to improve the quality of research in personality psychology. These include the need to develop consensus regarding research goals, terminology, measurements, data handling, and the current standing on theory and evidence. The authors also suggest that the development of such consensus should be rewarded, and they provide concrete criteria for rewarding good scientific practices. However, some of these strategies might not be feasible, and at worst, might do more damage than good. Below, we critique some of Leising et al.'s (2022) suggestions for implementing a consensus-building approach.

Differing Frameworks and Goals for Evaluating Scientific Evidence

Academic discourse is bound to be rife with disagreements. In the target article, the authors suggest that experts in the field can develop a consensus on which established literature they rely on as well as terminology and measurements that researchers use for a specific subject matter. They also suggest that past evidence can be evaluated objectively using logic and mathematical modeling. However, researchers with different approaches and scientific goals might, justifiably, rely on different pieces of the scientific puzzle to support their claims. For example, when evaluating a personality questionnaire, a researcher trained in psychometrics might care more about the psychometric model or structure of the data while a researcher with clinical expertise might rely more on the measure's diagnostic and predictive utility in deciding whether a questionnaire has sound validity. Such differential weighing of scientific evidence is inevitable and adds value by providing a holistic understanding of the construct in question.

Influences Beyond Scientific Interests

Although there is value in collaboration and consensus, it would not necessarily immunize scientists from the biases that the authors are trying to avoid. Psychological science is by default, a human endeavor embedded in a sociopolitical context. By suggesting that, through collaboration, researchers can reach a consensus on common views, methods, and terms, the authors presume that the objective features of psychological science can be fully distilled from social processes. However, scientific priorities and perspectives can never be fully protected from social and political motivations. For example, scholarship suppression or even academic boycotting can occur when academics present evidence that is not harmonious with the moral views the public holds at a given time (see Stevens et al., 2020 for specific examples). Moreover, when collaborators have differing viewpoints, they may negotiate in order to reach some agreement. Such negotiation can be affected by researchers' self-interest, cognitive biases, and extraneous factors like researchers' reputation or popularity. Therefore, pressures to concur can exacerbate systematic bias in scientific literature.

Construct Diversity

The jingle-jangle fallacy, defined as the tendency to use the same term to denote different things or different terms to mean the same thing, as explained by the authors is indeed problematic. Psychological constructs are abstract, hypothesized ideas of a latent (hidden) variable. Legitimate conceptual differences in how researchers define psychological constructs make it possible and sometimes necessary for variations of the same construct to exist in parallel. For example, there are several operationalizations and related measures for the construct of psychopathy and it is scientifically meaningless to ascertain which one is most representative because each variation of the construct has its own assumptions and nomological network resulting in different but non-comparable evidence for convergent and discriminant validity (see Lilienfeld et al., 2015). In such situations, conceptualizations of a construct might be different to the extent that a single measure is not enough to capture all operationalizations but overlap enough that each variation does not warrant a unique label.

Moreover, while some personality traits and related constructs are timeless and universal, most are likely to be bound by the specific context in which it was developed (Gooding, 2000). The authors propose that for psychologists to formalize various conceptual theories, they should use mathematical modeling or collaborate with mathematicians. Even if such an effort can bring researchers closer to the "true score", some psychological constructs, particularly new ones, lack the specification to permit the calculation of exact values, but this does not necessarily mean they lack value.

Ambiguity About Gold Standards of Measurement

Leising et al. (2022) propose that in order to develop consensus, using different measurements for the same construct should be discouraged. The authors recommend that experts in the field should create a consensus measure or recommend a few measures that researchers must use if they are examining a specific construct of interest. But if all models are ultimately wrong (Box, 1979), Leising et al.'s (2022) measurement mandate might prove too restrictive. Moreover, for highly heterogeneous constructs, a single gold standard of measurement might prove impossible. Even if, as per the authors' suggestion, a consensus document is created, it would be difficult to determine the practicalities of what evidence and how much of it warrants the inclusion of a new measure, especially because experts in the field might operate within diverse frameworks.

Researchers in personality psychology might be able to achieve intellectual consensus in so far as they share a similar scientific and technical background and have a common framework for approaching the scientific problem. Consensus achieved through this medium might come at a cost of reducing intellectual diversity. We applaud the authors for highlighting valid concerns that without consensus, academics might be susceptible to engaging in bad scientific practices like using outdated theories, deeply flawed measurement tools, or creating new measurements that overlap almost entirely with existing ones. Consortiums and other collaborative efforts as recommended in the target article and rigorous standards maintained by peer reviewers and journal editors are steps in the right direction to promote the development of partial consensus while allowing for scientific diversity and innovation. But for all researchers in any field to achieve consensus regarding terminology, measurement, and research goals seems not only overly optimistic but also, at times, counterproductive to adopting good scientific practices. We believe that encouraging disagreement in academic discourse is just as worthwhile as building consensus and is crucial for scientific advancement.

References

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Gooding, D. C. (2000). Experiment. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 117-126). Blackwell.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2,* Article e6029. https://doi.org/10.5964/ps.6029

Lilienfeld, S. O., Watts, A. L., Francis Smith, S., Berg, J. M., & Latzman, R. D. (2015). Psychopathy deconstructed and reconstructed: Identifying and assembling the personality building blocks of Cleckley's Chimera. *Journal of Personality, 83*(6), 593–610. https://doi.org/10.1111/jopy.12118

Stevens, S. T., Jussim, L., & Honeycutt, N. (2020). Scholarship suppression: Theoretical perspectives and emerging trends. *Societies, 10*(4), Article 82. https://doi.org/10.3390/soc10040082

9 – Comment

Consensus-Finding and Legitimacy: Commentary on Leising et al.

Adrianne John R. Galang[1,2], Marie Rose H. Morales[1,2]

[1] *Department of Psychology, University of the Philippines Diliman, Quezon City, Philippines.* **[2]** *UP Personality and Individual Differences Research Laboratory, University of the Philippines Diliman, Quezon City, Philippines.*

Leising et al. (2022) anticipate that, with proper incentives, various groups will work together to articulate possible versions of consensus, and that these positions will then be considered and debated by the field. For some aspects of our science this is already happening: the past decade provides a good example of how a consensus was built around the idea that a significant portion of our results in psychology were not replicable. It helps that the evidence was so provocative and compelling (e.g., Open Science Collaboration, 2015), and yet even then there was initial disagreement about the extent of the replication crisis (e.g., Gilbert et al., 2016). Because of these reforms, some kinds of consensus relating to better reporting standards, robust analyses, controlling for post-hoc theorizing, and greater transparency in methods overall are now less controversial.

In contrast, we believe that moving towards a consensus on substantial theoretical matters, without a thorough understanding of the current diversity of positions, might be problematic for three reasons: it will be less persuasive; it might discourage lines of investigation that might have led to a more robust convergence; once we have it, we might find the consensus less flexible than we would want.

Outside the numerous "grand" theoretical positions outlined in the standard textbooks, there is a whole menagerie of mid-level theories, and it is a genuine question whether the full range of ideas deployed by contemporary personality scholars are adequately documented in the usual reviews of the literature. Theorizing, of course, is not a matter of popularity, but consensus building is a social process. Legitimacy is accrued if the procedure is seen as having involved a wide sample of colleagues from the field. So it matters what the starting conditions are for what kind of consensus emerges, and our contention is that maybe we do not yet have a good handle on the full range of ideas at play.

And because such breadth is necessary, it also matters who is involved in the process of consensus building. The validity and legitimacy of any consensus will depend upon the extent of collaboration between researchers from the global north and those from the global south. Psychological studies, including those in personality science, have always been mostly reliant on WEIRD (Westernized, Educated, Industrialized, Rich, and Democratic) samples, a problem adequately discussed elsewhere (Henrich, Heine, & Norenzayan, 2010; Rad, Martingano, & Ginges, 2018). The much touted "consensus" on the Big 5 illustrates the challenges for future concordances. Findings from large samples around the world generally corroborate a five factor structure, but not without significant exceptions (Church, 2016). It is a credit to the field that, in our own case, a less-WEIRD population like the Philippines features prominently as part of the cross-cultural evidence base for the Big 5 (e.g., del Pilar, 2017; Katigbak et al., 2002), but this is just a demonstration that a non-English Big 5 instrument can be constructed in a non-WEIRD culture. Future attempts at convergence should do better by encouraging the search and development of viable alternatives from non-WEIRD scholars to test current frameworks against; to include more theorists rather than just more populations.

A premature narrowing of conceptual and methodological horizons might itself be detrimental to progress. If the rationality of a research program depends on its ability, in the long-run, to identify lines of investigation that are more fruitful than others, there is persuasive work from Mayo-Wilson, Zollman, and Danks (2011) formally demonstrating that a research strategy that is rational for an individual investigator (or individual lab) might not necessarily be rational for the group (or the field) if adopted by all of its members. The upshot of this is that investigators can still converge by independently pursuing what they believe to be the best course of action, as long as the results of their investigations are freely and transparently shared within the field. This is not inconsistent with the aims of the target article, but more explicitly bottom-up: a consensus paper or convention might be thought of as merely the articulation of an already existing convergence rather than an attempt to legislate it into being. This is "consensus-finding".

Leising et al. (2022) indicate that they are aware of the problems of reaching a sub-optimal consensus. They defuse these concerns by appealing to the idea that the consensus view should be constantly updated as the field develops. We find this view too optimistic. A consensus is political in nature, even when the content of the consensus is scientific. Franz (1997) documented the emergence of an international consensus on climate change and she observed that momentum towards a consensus gathered pace within a period when there was no significant change in the documented scientific findings. This implies that agreement was driven by factors outside of the data, factors arguably of a political nature. Because of this a scientific consensus, once established, might also be difficult to abandon since a change of direction could incur costs in terms of credibility and alliances. The cited example of defining the term "planet" is inapt: the change had no substantial impact on the theories and methods of astronomers, being

mainly a cultural wrangle (Messeri, 2010). We cannot imagine that a vote to define the word "trait" would pass so sedately.

Given what we've said above, we advocate for the following counter-proposals:

- Reward efforts at "consensus-finding": efforts to document and organize the full-range of contemporary positions on fundamental topics in the field. For example, Rauthmann's (2022) project of surveying personality scientists as to the definition of "trait", or an ethnography of laboratories, similar to Harp-Rushing (2020).
- Reward conscious efforts to incentivize, solicit, and amplify voices from a broad range of scholars, in support of "consensus-finding".
- Reward the pursuit of an explicitly limited consensus centered around explananda and validity of evidence, before seeking consensus on more extensive theoretical models. This will allow space for competing frameworks and bottom-up convergence, while establishing a common reference from which to evaluate them.

References

Church, A. T. (2016). Personality traits across cultures. *Current Opinion in Psychology, 8*, 22–30. https://doi.org/10.1016/j.copsyc.2015.09.014

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

del Pilar, G. H. (2017). The development of the Masaklaw na Panukat ng Loob (Mapa ng Loob). *Philippine Journal of Psychology, 50*(1), 103–141.

Franz, W. E. (1997, September). *The development of an international agenda for climate change: Connecting science to policy* [Monograph; IIASA Interim Report, IR-97-034]. http://pure.iiasa.ac.at/id/eprint/5257/

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science.". *Science, 351*(6277), 1037. https://doi.org/10.1126/science.aad7243

Harp-Rushing, K. O. (2020). *Fixing science: Innovation, disruption, maintenance, and repair at a North American open science non-profit* [Doctoral dissertation, University of California, Riverside]. https://escholarship.org/uc/item/8rs6279m

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29. https://doi.org/10.1038/466029a

Katigbak, M. S., Church, A. T., Guanzon-Lapeña, M. A., Carlota, A. J., & del Pilar, G. H. (2002). Are indigenous personality dimensions culture specific? Philippine inventories and the five-factor model. *Journal of Personality and Social Psychology, 82*(1), 89–101. https://doi.org/10.1037/0022-3514.82.1.89

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Mayo-Wilson, C., Zollman, K. J. S., & Danks, D. (2011). The Independence Thesis: When individual and social epistemology diverge. *Philosophy of Science, 78*(4), 653–677. https://doi.org/10.1086/661777

Messeri, L. R. (2010). The problem with Pluto: Conflicting cosmologies and the classification of planets. *Social Studies of Science, 40*(2), 187–214. https://doi.org/10.1177/0306312709347809

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America, 115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Rauthmann, J. F. (2022). *What are personality traits? Surveying perspectives, definitions, and persisting issues.* Manuscript in preparation.

10 – Comment

There Is no Viable Path to Consensus Based on the Current Research Literature

Katherine S. Corker[1]

[1] *Department of Psychology, Grand Valley State University, Allendale, MI, USA.*

Ten years into the replication crisis, psychologists have learned that we need to do better. Many initiatives have been proposed and implemented to improve research practices, but a neglected domain for improvement has been changing incentives. The target article (Leising et al., 2022) provides a concrete proposal for reshaping how we evaluate research, which they propose will better align incentives and improve research quality. I suspect other commentaries will focus their critiques on the particulars of the proposed point system for scoring research quality, which, although laudable for its specificity, contains some shortcomings related to its ability to properly reward research beyond very common research designs.

I will instead focus on where the target paper differs from other papers in this genre: its focus on fostering consensus. Leising et al. (2022) propose that "the field cannot go on without greater and more explicit consensus as to (1) what the most important issues to be investigated are, (2) how things shall be named, (3) how things shall be measured, (4) how data shall be analyzed, and (5) what the current state of theory and knowledge is" (p. 8). Given this focus on consensus, I was surprised that the authors gave only scant attention to meta-analysis, which is often upheld as a consensus building tool (Chan

& Arvey, 2012). Here, I explain why meta-analysis has often failed to create consensus, illuminating a more general flaw in the Leising et al. (2022) vision of consensus building, and I describe what we should do instead to achieve a more cumulative psychology.

Meta-analysis is a tool for providing summaries of sets of studies, both in terms of central tendency (what a typical study looks like) and variability (how much studies vary from one another). Furthermore, meta-analysis is the quantitative part of a broader research synthesis method known as systematic review. Systematic reviews are designed to take stock of all current knowledge on particular research questions, and they involve a repeatable, comprehensive search strategy, coupled with evidence quality evaluation. Systematic reviews may include meta-analyses, but they can also be qualitative (see Corker, 2022 for additional discussion on strengths and weaknesses of both methods).

As such, systematic reviews and meta-analyses seem to be promising candidates for establishing consensus in a given research area. Indeed, Chan and Arvey (2012) said as much: "Meta-analysis may contribute to the advancement of knowledge and normal science … by facilitating the building of consensus in a field or topic" (p. 85). The idea is that a review will provide a birds-eye overview of current evidence on a topic, showing where consensus vs. disagreement exists in the knowledge base. Interestingly, although consensus would seem to be an unabashed good, historically there have been worries that meta-analyses would provide too definitive an answer to a question. If strong consensus was revealed through a meta-analysis, further research on a topic might be quelched: "Meta-analyses may declare a 'winner' or 'loser' before an ongoing trial is over" (Feinstein, 1995, p. 77).

Although promising as a consensus building tool, the reality of meta-analysis is less sanguine. Anyone who has attempted to perform a meta-analysis or systematic review will report how frustrating it is to discover how idiosyncratically researchers work. In most areas of psychology, there is little overlap between studies in terms of measurement, methodology, and even theoretical approach. Crucially, the lack of commonalities between studies sometimes makes synthesis well-nigh impossible. How can the results of such studies even begin to be combined?

Leising et al. (2022) lament this exact state of affairs and propose that we reward those who build consensus positions and use consensual methods in their research to encourage the proliferation of this kind of work. A researcher eager to adopt their recommendations might therefore gather several research groups together to engage in a meta-analysis or systematic review, with the goal of uncovering the group's collective understanding of knowledge on a research question. But our eager colleague would almost immediately be met with a problem: in the presence of such diversity in measurement, methods, and theory, exactly where is consensus supposed to come from?

Put succinctly, my argument is the following. Given the current state of affairs in psychology – in which 50% or more of our results fail to replicate, the vast majority of measures are ad hoc or of unknown validity, theories are so vague as to be mostly

unfalsifiable, and there is little to no overlap in approach between studies in supposedly related topical areas – neither meta-analysis nor any other currently known consensus generating procedure is going to be sufficient to allow us to fashion gold from lead.

If true, where does that leave us? Before we "quit the academy and make an honest living selling shoes" (Meehl, 1990, p. 237), we might consider the alternative course proposed by Paul Rozin twenty years ago. Rozin (2001) argued that detailed, precise descriptions of reliably occurring social phenomena should be the basis for later experimentation and model testing. In contrast with biology and other developed sciences, he argued, psychologists have often jumped straight to hypothesis testing and skipped over the rich observational research that could inform and shape those experiments.

Certainly, it could be argued that personality psychology has valued this kind of descriptive work more than some other subfields of psychology. Indeed, the existence of the Big Five as an organizing framework is a testament to how generative careful descriptive work can be. But given our current levels of progress and understanding, we could surely do even more such work. Is my argument a call for totally atheoretical, inductive work? No, because even descriptive work is theory-laden and must begin somewhere (Oreskes, 2019). But if theories are explanations of available facts, then to strengthen our theories – and therefore our consensual understanding – it would be nice to first have some reliable facts to explain. But attempting to jump straight to consensus on the basis of our current (highly flawed) research literature, as Leising et al. (2022) recommend, is likely to fail.

References

Chan, M. E., & Arvey, R. D. (2012). Meta-analysis and the development of knowledge. *Perspectives on Psychological Science, 7*(1), 79–92. https://doi.org/10.1177/1745691611429355

Corker, K. S. (2022). Strengths and weaknesses of meta-analysis. In L. Jussim, J. A. Krosnick, & S. T. Stevens (Eds.), *Research integrity: Best practices for the social and behavioral sciences.* Oxford University Press.

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology, 48*(1), 71–79. https://doi.org/10.1016/0895-4356(94)00110-C

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2,* Article e6029. https://doi.org/10.5964/ps.6029

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Oreskes, N. (2019). Why trust science? Perspectives from the history and philosophy of science. In S. Macedo (Ed.), *Why trust science?* (pp. 15-68). Princeton University Press.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

11 – Comment

Efforts to Improve Personality Psychology Must Prioritize the What, Who, and Why, Not Only the How

Jonathan M. Adler[1]

[1] *Olin College of Engineering, Needham, MA, USA.*

Efforts to improve research in personality psychology often prioritize the way science is conducted and overlook the importance of what gets studied, by whom, and for what reasons. Leising et al.'s (2022) proposal of ten steps towards a better personality psychology are certainly laudable recommendations for innovation in our field but, like many contemporary scientific reform efforts, they emphasize the *how* and largely overlook the *what, who, where, and why* of research. While progress in the conduct of scientific research is absolutely important, we should not pursue this progress without attending to the vital matter of science's broader aims: improving our world.

What Gets Studied?

Leising et al. (2022) open their agenda with a bold effort to define "good research." They arrive at this definition: "Good research works toward understanding and predicting important phenomena" (p. 3). Their ten steps make major progress in identifying processes that might lead towards improved science, but they do not directly tackle the question of how to define the focus of inquiry: "important phenomena." Leising et al. (2022) appropriately note that "scientific facts are claims about which agreement has been reached among scientists in the respective field" (pp. 4-5). If we prioritize the process of building consensus among today's scientists about today's corpus of knowledge we risk foreclosing on a body of scientific facts that reify the outputs of the very power structures that Leising et al. (2022) so deftly critique. As one example, people with disabilities are almost completely absent from the personality literature, despite being the largest minority group in the United States (Adler et al., 2021). Instead, an agenda of consensus building must be complimented by an agenda of broadening and diversifying the topics personality psychologists have studied.

Who Participates in Science (and Where)?

Collaboration and dissent rightly feature prominently in Leising et al.'s (2022) ten steps for improving personality psychology. In addition to elevating these processes, we need

to consider the barriers to participating in our field that keep some potential scholars from ever becoming prospective collaborators or dissenters. For example, all three Invited Symposia at the most recent meeting of the Association for Research in Personality in July, 2021 focused on issues of who conducts research in our field and where. The opening symposium, titled "How Open Is Personality Psychology to Researchers from Marginalized Communities?" (King & Booker, 2021), focused on barriers to entry in the social practices of our field, in our pedagogical approaches, and in our intellectual traditions. The second symposium, titled "The Big Five Across Cultures" (Thalmayer, 2021), focused on the need to examine dispositional traits in different global contexts and the benefits of using emic methods designed to protect against narrowly reproducing Western (and especially American) conceptualizations. And the third symposium, titled "Where Are Race, Culture, and Ethnicity in Personality Research?' (Syed 2021), identified the overly-narrow ways in which personality psychology has conceptualized race, culture, and ethnicity, both within our studies and within our professional societies. Efforts to improve the conduct of personality psychology that do not simultaneously work to broaden participation in our field along demographic, geographic, and intellectual axes will fail to recruit the appropriate sets of collaborators and dissenters necessary to succeed.

Why Study Personality?

Leising et al. (2022) adopt a laudable reflexivity, acknowledging their own values that shaped their proposal. This is important not only for the quality of their proposal itself, but also for the modeling it offers to the field, demonstrating that all scientific agendas are created in the context of values, which too often remain implicit. Leising et al. (2022) name transparency, collaboration, efficiency, and accountability as the dominant values that influenced their recommendations. They also acknowledge that they "neglect the importance of originality, innovation and relevance in good research" (p. 41). In arguing for an expanded set of priorities for improving personality psychology, I would like to elevate and sharpen "relevance" as an ethical imperative for our field (in my read, originality and innovation are ways of pursuing new knowledge, whereas relevance is itself an absolutely vital goal). Despite the proliferation of journal articles in personality psychology, the resources that produce personality research (time, effort, money, etc.) are finite. As such, personality psychologists must not only pursue research questions that will be informative, but those that actually matter. Our world is on fire – literally and metaphorically. There are far too many emergencies that require our expertise as personality psychologists to continue conducting our science in the way we have been. Across our field we need to turn our attention to the huge range of issues that plague our world. This does not mean that all personality research must be applied research (though I do believe we need substantially more applied research in our field), but that alongside new standards for good research we must adopt a "So What Criterion" – an

expectation that our research serve pressing contemporary issues. I shudder at Leising et al.'s (2022) notion that these priorities "will still have to be assessed by journal reviewers or hiring/tenure committees, in much the same way that they are already being assessed at present" (p. 41). Instead, I see no reason that we cannot strive towards prioritization and consensus about which issues are most pressing and leverage the same kinds of incentives that Leising et al. (2022) describe in encouraging personality psychologists to study them. A new ethics of how we conduct research must extend to what broader purpose our research serves.

As we take steps to improve personality science we must not only attend to the way we conduct our scholarship – the how – but also to what we study, who the people conducting personality psychology are (including where they live), and why our research matters to the broader world. Doing so will not only improve our research, but also center ethics in the practice of personality psychology in the service of meaningful impact. And, just as there is urgency in improving the *how* of personality psychology, there is equal imperative for expanding beyond *how* to include *what*, *who*, *where*, and *why* – the *when* is now.

References

Adler, J. M., Lakmazaheri, A., O'Brien, E., Palmer, A., Reid, M., & Tawes, E. (2021). Identity integration in people with acquired disabilities: A qualitative study. *Journal of Personality, 89*(1), 84–112. https://doi.org/10.1111/jopy.12533

King, L. A., & Booker, J. (2021, July 9). How open is personality psychology to researchers from marginalized communities? [Conference presentation]. Association for Research in Personality Conference. https://www.youtube.com/playlist?list=PLc31LGrEuscPHbIOtLxV30aWUgljforXQ

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Syed, M. (2021, July 23). Where are race, culture, and ethnicity in personality research? [Conference presentation]. Association for Research in Personality Conference. https://www.youtube.com/playlist?list=PLc31LGrEuscPLUFMDJLuqLqpBUL3Fk8hx

Thalmayer, A. G. (2021, July 16). The Big Five across cultures [Conference presentation]. Association for Research in Personality Conference. https://www.youtube.com/playlist?list=PLtBIkCgRptvn3j7rgo9Z26aQEv0jz_por

12 – Comment

CRediT Where Credit Is Due: A Comment on Leising et al.

Emorie D. Beck [1], Clifford I. Workman [2], Alexander P. Christensen [2]

[1] *Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*
[2] *Penn Center for Neuroaesthetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.*

Leising et al. (2022) propose a 10-step checklist that they argue will facilitate "a better personality science." Although we agree with many of the proposed steps, whether the checklist separates "good research" from bad is an empirical matter. We therefore consider whether the checklist would have caught one of the replication crisis's most infamous papers—namely, Bem's (2011) "Feeling the future" in the *Journal of Personality and Social Psychology*. Table 1 demonstrates the difficulty we faced in coming to a consensus on the score to assign Bem's paper, which is the first of our several criticisms.

Table 1

A Worked Example of Bem (2011) Using the 10-Step Checklist Proposed by Leising et al. (2022)

Criteria		Max Score	CIW	EDB	AC
0	Paper gets published in a peer reviewed outlet	1.0	1.0	1.0	1.0
1a	Presents broad consensus regarding important research goals	5.0	0.0	0.0	0.0
1b	Addresses important research goals that were outlined in consensus document	0.5	0.5	0.5	0.5
2a	Presents broad consensus regarding terminology	5.0	0.0	5.0	0.0
2b	Uses terminology from consensus document	0.5	0.5	0.5	0.5
3a	Presents broad consensus regarding measurement practices	5.0	0.0	0.0	0.0
3b	Uses measurement practices from consensus document	0.5	0.5	0.5	0.5
4a	Presents broad consensus regarding data pre-processing and/or analysis	5.0	0.0	0.0	0.0
4b	Uses consensus practices regarding data pre-processing and/or analysis	0.5	0.5	0.5	0.5
5a	Presents broad consensus regarding state of knowledge and/or theory development	5.0	5.0	0.0	0.0
5b	Builds directly on consensus document regarding state of knowledge and/or theory development	0.5	0.0	0.5	0.5
6a	Includes algebraic or formal-logic formulation of theory being tested, and how it relates to measured variables	2.0	0.0	0.0	0.0
6b	Includes account of how the tested formal theory relates to previous formulations of the same or related theories	1.0	0.0	0.0	0.0

Criteria		Max Score	CIW	EDB	AC
7a	Strictly separates explorative from confirmatory analyses, with the latter being pre-registered at the same level of specificity at which the results are later reported	1.0	0.0	0.0	0.0
7b	Is a registered report	2.0	0.0	0.0	0.0
8	Includes at least one direct replication attempt (of others' or one's own results), with a new sample and at least equal power as previous study	1.0	1.0	1.0	1.0
9a	Includes pre-registered a priori power analysis / sample size planning based on specific and realistic expected effect size estimate	0.5	0.5	0.5	0.5
9b	Has an expected type I error rate of ≤ .05 and type II error rate of ≤ .20, based on realistic effect size estimates	1.0	1.0	0.0	0.0
9c	Demonstrates representativeness of participant samples(s) in regard to the population of interest	3.0	0.0	1.5	1.0
9d	Demonstrates representativeness of stimuli in regard to the environmental conditions of interest	3.0	0.0	3.0	3.0
10a	Data is made open	0.5	0.5[a]	0.5	0.0
10b	Open data is accompanied by meta-data that (at least) documents all variables in the data set in a manner that enables new analyses without requiring further interactions with the people who collected the data (see FAIR principles)	1.0	0.0	0.0	1.0
10c	Code is made open (and well documented)	0.5	0.0	0.5[b]	0.0
10d	Materials are made open (and well documented)	0.5	0.0	0.0	0.0
10e	All data, materials and code from a project are found in a single directory online	0.5	0.0	0.0	0.0

[a]Available from https://replicationindex.com/2018/01/20/bemcorrespondence/
[b]Available through contact with Bem (2011).

The midpoint of our scores was 12.0 (building consensus: 3.3, using consensus: 2.3, formalization: 0.0, preregistration: 0.0, replication: 1.0, informativeness: 3.5, and open science: 0.8). Does this score separate the bad from the good, the reproducible from the irreproducible? These questions are difficult to answer for at least two reasons. Since most published research hasn't been scored, individual scores are difficult to contextualize. However, even if most published research was scored, and a consensus between scorers was reached, we contend that conceptual problems built into the checklist render scores difficult to interpret, and that the scope of the checklist misses important things.

Consensus Statements

A critical component of Leising et al.'s (2022) steps toward improving scientific standards in personality center around consensus building. There are several critical ways in which the methods for building consensus in psychology could have unintended negative consequences by assuming: (1) our science is sufficiently mature for consensus to emerge and (2) that consensus building will change incentives in ways that do not reward well-known, eminent, and productive individuals (in terms of publication numbers).

Eminence and Advancement

First, it is unclear whether and how early career researchers (ECRs) and researchers from underrepresented backgrounds (RUBs) will have a role in consensus building. The proposed system rewards individuals for collaborating with others to build consensus. If past initiatives and expert meetings in personality are a representative sample, however, then consensus will be driven by a small group of mid- to late-career researchers from the United States and Western Europe. In part, this arises from eminence with many eminent personality scholars having little contact with other researchers outside of western nations. Thus, how ECRs and RUBs will play a role in consensus building results in an unfair penalization of location and rank.

Second, there is an inherent tension between consensus and innovation. Given the overlap between eminent scholars with those in reviewer, editor, and other positions of influence, consensus statements are prone to enabling undue gatekeeping against challenges to consensus. Or, at minimum, publishing contradictory statements and research becomes prohibitively difficult, particularly for ECRs and RUBs, because researchers who disagree with the consensus may adhere for the sake of the proposed standards of "quality."

Finally, scholars require adequate training in the requisite domain(s) to create consensus documents in them. Yet the current academic system rewards individual contributions, particularly empirical ones, more than team-science-based or theoretical contributions. For consensus documents to guide research, personality science needs to (1) embrace team-science by rewarding many types of contributions, and (2) train students in theory building as much as statistics.

A CRediT Alternative

Creating a better science requires a shift in academia's reward structures. The current system rewards producing more publications with little reference to contributions to those publications. Even with Leising et al.'s (2022) ten steps toward a better science, researchers who produce more "quality" research with minimal contribution will be most rewarded. The contributor roles taxonomy (CRediT[3])—fourteen high-level roles that specify the researcher's contribution to a publication—offers a simple yet effective way of

Figure 1

Example of CRediT Section in a CV

Dr. Emorie D Beck, Ph.D. CV June 2021

Contributorship Roles (CRediT)

Paper	Year	Concept	Data Cur.	Analysis	Funding	Invest.	Method.	Admin.	Resources	Software	Supervis.	Validation	Vis.	Draft	Review
Invited Journal Articles															
Beck & Christensen	2021	X	–	–	X	X	X	X	–				X	X	X
Jayawickreme et al.	2021	X	–	–	–	–	–		–	–		–	–	X	X
Beck & Jackson	2020a	X	X	X	X		X	X		–		X	X	X	X
Nosek et al.	2019	X	–	–	–	–			–	–			X		X
Journal Articles															
Beck & Jackson	2021a	X	X	X			X	X	X	–		X	X	X	X
Bollich-Ziegler et al.	2021	X	X	X			X	X	X	–			X	X	X
Malle et al.	2021	X		–			X		X	–					X
Beck & Jackson	2021b	X	X	X	X	X	X	X	X	–		X	X	X	X
Saef et al.	2021	X	X	X	–	–	–	X	–	–		–	X	X	X
Hill et al.	2021	X	X	X			X	X	X	–		X	X	X	X
Jackson & Beck	2021a	X	X	X	X	X	X	X	X	–			X	X	X
Jackson & Beck	2021b	X	X	X			X	X	X	–		X	X	X	X
Frumkin et al.	2020	X	X	X			X		X	–			X	X	X
Beck & Jackson	2020b	X	X	X			X	X	X	–		X	X	X	X
Piccirillo et al.	2019	X	X	–	–		–	X	X	–		–	X	X	X
Beck & Jackson	2017	X	X	X			X	X		–			X	X	X

Note. "X" indicates contributorship role, "–" indicates not applicable for a given project, and blanks indicate roles not undertaken. "Concept" = Conceptualization; "Data Cur." = Data Curation; "Invest" = Investigation; "Method" = Methodology; "Admin" = Project Administration; "Supervis." = Supervision; "Vis." = Visualization; "Draft" = Writing: Original Draft; "Review" = Writing: Review & Editing.

weighing the quality of researcher contributions rather than quantity alone. We propose adding a CRediT section to CVs to move us toward contribution-based standards (see Figure 1 for an example).

Besides making a researcher's contributions clear, the CRediT section offers insights into expertise and team science. A quantitative role may include "analysis." A supervisory role may include "conceptualization" and "funding." "Writing: original draft" indicates

3) https://casrai.org/credit/

that the researcher contributed to their content area. Different combinations of contributions reflect different yet important roles in team science.

Theory Training

The target article is the latest to join in calling for more formal theory (e.g., Borsboom et al., 2021; Gray, 2017; Oberauer & Lewandowsky, 2019). While we agree that better theory can help move psychology forward (Oude Maatman, 2021), psychology trainees aren't trained to think theoretically (Bosch, 2018). If psychology is to improve theory, then psychologists must be trained in theory (Smaldino, 2019). Trainees receive instruction on basic and advanced statistics, yet they often do not receive training on basic theory. Instead, students take courses that are within their area of specialization (Bosch, 2018). Learning about extant theorizing is a far cry, however, from learning how to theorize.

Borsboom and colleagues (2021) outline a theory construction course that provides an example for training programs. Students learn to distinguish between data, phenomena, and theories. Then, students choose a topic in psychology to identify robust phenomena. Finally, students use software for simulations to create models that test theoretical propositions. This program could be split into separate courses where students learn to understand the difference between modeling and theory (Haslbeck et al., 2019), simulate models to test theories (Robinaugh et al., 2021), and investigate incompatibilities and underdetermination in theory (Oude Maatman, 2021).

Conclusion

Leising et al.'s (2022) rubric fails to address underlying systemic issues in psychology. Consensus building, while important, will only reify eminence and gatekeeping. Rating research based on a "quality-based" checklist as opposed to a contribution-based rubric perpetuates the score counting that is endemic to the current reward system. Finally, formal theories cannot be achieved without formal training in theory. These systemic issues are pervasive and cannot be fixed without changes to the reward structure. Without changing the evaluation system of academia, we cannot change the reward system that supports it. Including the CRediT taxonomy in CVs offers a simple yet effective way of weighing the quality of researcher contributions rather than quantity alone.

Funding: E.D.B. was supported by National Institute on Aging grants R01-AG067622 and R01-AG018436. C.I.W. was supported by the National Institute of Dental & Craniofacial Research grant F32DE029407. A.P.C. was supported by a grant funded by the Templeton Religion Trust. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institutes of Health or the John Templeton Foundation.

Author Contributions: E.D.B., C.I.W., and A.P.C. all fulfilled the CRediT roles of Conceptualization, Investigation, Method, Project Administration, Visualization, Writing: Original Draft, and Writing: Review and Editing. Author order was determined by five random shuffles of the authors' names in Excel.

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407–425. https://doi.org/10.1037/a0021524

Borsboom, D., van der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science, 16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bosch, G. (2018). Train PhD students to be thinkers not just specialists. *Nature, 554*(7690), 277. https://doi.org/10.1038/d41586-018-01853-1

Gray, K. (2017). How to map theory: Reliable methods are fruitless without rigorous theory. *Perspectives on Psychological Science, 12*(5), 731–741. https://doi.org/10.1177/1745691617691949

Haslbeck, J., Ryan, O., Robinaugh, D., Waldorp, L., & Borsboom, D. (2019). *Modeling psychopathology: From data models to formal theories.* PsyArXiv. https://doi.org/10.39227/osf.io/jgm7f

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. Advance online publication. https://doi.org/10.5964/ps.6029

Oude Maatman, F. (2021). *Psychology's theory crisis, and why formal modelling cannot solve it.* PsyArXiv. https://doi.org/10.39227/osf.io/puqvs

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Robinaugh, D. J., Haslbeck, J. M., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science, 16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature, 575*(7783), 9. https://doi.org/10.1038/d41586-019-03350-5

13 – Comment

A Different Road Towards a Better Personality Science

Kate C. McLean[1], Moin Syed[2]

[1] *Department of Psychology, Western Washington University, Bellingham, WA, USA.* **[2]** *Department of Psychology, University of Minnesota, Minneapolis, MN, USA.*

There is no question that there is a need for a better personality science. Less clear, however, are the answers to what we think "better" actually means and how exactly we get there. At the 2021 conference of the *Association for Research in Personality*, all of the invited sessions focused on how to create a more diverse, equitable, and globally-minded personality science. The conference is largely focused on North American representation and perspectives, but there were other voices as well, and one of the invited sessions pertained to diversity and inclusion in personality research in Europe. Collectively, the sessions made plain that there is a hunger for change, and a *need* for change that, if not addressed, could derail the field. Attending that conference and reading the target article by Leising et al. (2022) left us with the sense that we have entered into two different understandings of reality.

Sessions at the conference focused heavily on reflecting on and improving our methods to generate a stronger knowledge base, just as Leising et al. (2022) did. The point of departure, however, is that improving our methods are not enough. Like many of our colleagues at the conference, we emphasize the need to understand how our methods are intertwined with the tremendous lack of inclusivity and diversity in our field, and how they reify existing power structures. Improving our methodological approach can not only lead to more robust conclusions, it can also make us think more deeply about what we do and how we do it. Perhaps most importantly, doing so will help us to understand and acknowledge the limits of what we do. This kind of humility and reflection is at the core of "steps towards a better personality science," though as we elaborate, this might be characterized as more of a sea change than steps.

Although we are fans of common ground and understand the need to speak the same language and share tools to work together, we found the emphasis on consensus decision-making much too strong and much too premature. First, as we elaborate below, we do not yet have the appropriate data or methodological development to make these decisions. Second, there was no proposal to integrate the necessary work of reflecting on the biases that would impact such decisions. Third, there was no discussion of who would be in the room to make these consensus decisions, or of the power dynamics intrinsic to our existing systems that are perpetuated with such practices. The risk of Leising et al.'s (2022) proposal is that decisions made by a select few will not only contain unacknowledged bias, but will also systematically marginalize and restrict the very kind of work that is needed to understand our own biases in measurement and the

experiences and voices that have been left out of our science. The risk is the potential for no meaningful change at all.

This point is not hypothetical, as it is indeed already what we see in the field. The majority of our common tools have not been subject to sufficient tests of whether they are appropriate across diverse groups of people. In her talk at the conference, Monisha Pasupathi (2021) eloquently argued that if one takes diversity in the human condition seriously, we will need to understand that our common constructs, methods, and measures are not appropriate or relevant for all people. We are not physicists. We are not evolutionary biologists. We are scientists who study humans in all their messy complexities, which include their messy and complex cultural-historical contexts. We will have to realize the limits of what we are doing and do something different, which means we need to slow down and deepen our understanding, rather than prioritize rushed consensus based on limited knowledge.

The overarching problem is that the paper is targeted at the wrong level. Before we can make the large-scale movements towards the better science Leising et al. (2022) have envisioned, we need to have a much broader scope on where we have gone wrong - beyond the under-powered and unreplicated studies that have given rise to the current crisis. We need to understand the power dynamics and inequitable systems that supported and celebrated those scholars and programs of research that were based on faulty data and designs. We need to understand the historical context of why we value what we value (e.g., quantitative over qualitative; Syed, 2021a). We need to understand how our biases and values as researchers shape why we study what we study (e.g., Dunn et al., 2021). We need to take a good hard look at what is missing in our science (Syed, 2021b). We need to do the work.

Engaging in this kind of reflection highlights how the current emphasis on "good science" comes from an entirely quantitative perspective. Descriptive and qualitative work are the very kinds of tools that will help us to understand the potential limits of our theories and measures, the appropriateness of our measures for different populations, and the variety in the human condition that we should be capturing if we are doing our job. Unfortunately, the scoring system proposed by Leising et al. (2022) places this kind of descriptive and qualitative work at an even greater disadvantage. Qualitative research - mentioned only very briefly in passing - would receive the very lowest of scores, not even scorable on multiple components of the system.

We also find it ironic that the authors argue against the quantity-focused form of assessment that dominates the field, instead advocating for an approach that is more nuanced and grounded in the actual practices of the field (even if incomplete). Perhaps the same appreciation of the bias of quantitative work, and the potential benefits of qualitative work, can be applied to our science.

In short, we don't offer tidy solutions because the first step is serious reflective work. We need to truly examine the science we have created, and let some of it go. Who knows what we might discover.

References

Dunn, E. W., Chen, L., Proulx, J. D. E., Ehrlinger, J., & Savalei, V. (2021). Can researchers' personal characteristics shape their statistical inferences? *Personality and Social Psychology Bulletin, 47*(6), 969–984. https://doi.org/10.1177/0146167220950522

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Pasupathi, M. (2021, July). It's past time to acknowledge that all psychology is cultural psychology [Conference session]. Virtual Association for Research in Personality [online].

Syed, M. (2021a). It's 2 x 2 designs all the way down: Social psychology's over-reliance on experiments needlessly restricts diversity in the field. Invited presentation at the Society for Personality and Social Psychology Annual Conference [online]. https://osf.io/gc3mu/

Syed, M. (2021b). *Where are race, ethnicity, and culture in personality research?* PsyArXiv. https://doi.org/10.39227/osf.io/m57ph

14 – Comment

Consensus in Context: Clear Disagreement Can Be the First Step Toward Agreement

Erick J. Fedorenko[1§], Patrick V. Barnwell[1§], Richard J. Contrada[1]

[1] *Department of Psychology, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA.*

[§]*These authors contributed equally to this work.*

Consensus

Scientific consensus is an instrumental goal, desired end state, and marker for progress. Yet, consensus is only as good as what is agreed upon. Although Leising et al. (2022) place considerable stock in consensus, there is welcome nuance to their position, with more implied by agreement with Oreskes (2020), that we would like to see further elaborated.

As Leising et al. (2022) note, good science involves rigorously testing multiple theories (Chamberlin, 1965; Platt, 1964). The same holds for efforts to define and foster that

which constitutes good science. The alternative is to form consensus around a single pet position, or lowest common denominator across several, in a popularity contest or bargaining process. To avoid such outcomes, participants with conflicting viewpoints must participate. In addition, when a progress report on obtaining consensus is issued, at least one minority report could be appended. Although egos would likely make it difficult for some to participate, and for some to admit they were wrong, science stands to benefit from continuing debate.

Which viewpoints to represent? We assume that the list of five domains addressed by Leising et al. (2022), which includes measurement, but not experimental manipulations, interventions, and situational contexts, is not intended to focus on personality *dispositions* to the exclusion of *social-cognitive processes* (Mischel & Shoda, 1995) and *person-situation transactions* (Buss, 1987). And that omission of sampling is not intended to exclude consideration of age, gender, race, ethnicity, health status, and culture.

But consensus among even a diverse group of personality researchers would be limited. We favor representation of other subareas of psychology, and disciplines outside psychology. Otherwise, there is risk of creating an echo chamber and a concession that personality only matters to personality researchers. How can personality science enhance understanding and address issues related to mental and physical health, social interactions, and organizations, and benefit from taking on these problems? How can it contribute to and be enriched by disciplines other than psychology, from genetics and neuroscience to cultural anthropology?

How would consensus look? It may turn out to be all too familiar, highly limited, even trivial—like many peer-reviewed publications. Papers often provide a selective literature review to fit a narrative. There is usually rote listing of caveats, each paired with a counterargument and doubling down on the narrative. Would the consensus statement of a panel of personality scientists, and subsequent papers that endorse and follow it, look much different?

Are there models for less mundane forms of consensus-building? Is the means by which the Big Five Trait model was arrived at (John & Srivastava, 1999) a good example? Or the mechanisms for creating each new version of the DSM (Kendler & Solomon, 2016)? What have we learned from such efforts?

Given that safeguards are needed against arriving at consensus views that are not valid, we suggest a step *prior* to consensus-building: participants with *opposing* views collaborating to describe the *absence* of consensus, *sharpening our understanding of areas of disagreement*, and suggesting *different ways to test different perspectives* against each other. The ensuing effort to arrive at consensus would likely benefit.

Quality Assessment

Is consensus about quality achievable? Many criteria are not mentioned by Leising et al. (2022) and there are alternative ways to structure them: Contributions may be methodo-

logical, theoretical, empirical, and/or application-related; they may have heuristic value, show integration across perspectives, and/or clarify distinctions between them; there is critique, innovation, and creativity. Even with an agreed upon list, would we agree on measurement and prioritization? And if citation counts are faulty and subject to misuse, which quality markers would be any different?

Our impression is that research quality *is* given significant weight in hiring, promotion, the granting of scientific awards, and publication. Perhaps its weight should be greater. But who would assign points for quality? Would all journals participate? Would they score quality in the same way? Would any proposal to implement a quality metric garner more agreement than exists for the proposition that quantity matters? That more of a good thing is better than less of a good thing?

The peer-review publication system, as much as any other mechanism, controls the quality of scientific literature. Is the absence of discussion of how the peer-review process might be modified to enhance the quality of published research a concession that it cannot be changed? Or that there is no reason to change it? With regard to the reward for grant funding, how can the application review process be improved, so that use of grant funding as reward has a greater positive impact on the science itself?

Do personality researchers want to be rewarded for quality? How can competing interests, like research quantity, grant funding, and notoriety, be overcome? Would they embrace a point system if they can still publish in top tier journals, which seems to be of great importance to many? Or, would top tier journals require a higher point threshold, which would bind the two? Top-tier journals often prioritize flash and novelty over quality, are highly read, and generate revenue. How can the role of the corporate bottom line be reformed or removed?

And if the field makes all these changes, would the result be welcomed by psychology departments, deans, and university presidents? Research quantity no longer mattering, no reward for acquiring research grants, and no more prestigious journals, all replaced by a digital quality badge?

Overall, we commend Leising et al. (2022) for fostering discussion among researchers interested in improving research practices. We do not dispute the need for a shift in personality science, perhaps all sciences, away from superficial measures of quantity, and towards an emphasis on quality. However, we feel it is important to ask the foregoing questions to direct attention to the thorny issues that impede such a paradigm shift, and hope they ultimately will be addressed by the personality field.

Author Note: Order of authorship for the first two authors was determined by coin toss.

References

Buss, D. M. (1987). Selection, evocation, and manipulation. *Journal of Personality and Social Psychology, 53*(6), 1214–1221. https://doi.org/10.1037/0022-3514.53.6.1214

Chamberlin, T. C. (1965). The Method of Multiple Working Hypotheses: With this method the dangers of parental affection for a favorite theory can be circumvented. *Science, 148*(3671), 754–759. https://doi.org/10.1126/science.148.3671.754

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Kendler, K. S., & Solomon, M. (2016). Expert consensus *v.* evidence-based approaches in the revision of the DSM. *Psychological Medicine, 46*(11), 2255–2262. https://doi.org/10.1017/S003329171600074X

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246

Oreskes, N. (2020). *Why trust science?* Princeton University Press.

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

15 – Comment

Three More Steps Toward a Better Quality Personality Science

Howard S. Friedman[1]

[1] *Department of Psychology, University of California, Riverside, Riverside, CA, USA.*

The target article (Leising et al., 2022) offers ten helpful steps toward a better personality science, but I believe several more steps will help advance the field even further. These suggestions come from first taking a step back to allow broader perspective.

The target article focuses mostly on current topics in open science and knowledge cumulation such as sharper theory and prediction, and better statistical power and reliability. The authors rightly propose appreciating quality over quantity, especially in determining the value of the work and of the scholar. Perhaps ironically, although the target article nicely refers to the "replication crisis," it would be helpful to go back nearly

half a century to the beginnings of the *Personality and Social Psychology Bulletin* (PSPB journal), which discussed some of the same issues as a "crisis" in our field.

In one article well-read nearly a half century ago, the then-prominent psychologist Zick Rubin discussed "On Measuring Productivity by the Length of One's Vita," in which he bemoaned the rating of Psychology departments by counting the number of articles published by their faculty (Rubin, 1978). He concluded his piece by saying, "I am not an opponent of quantification in psychology. It may even be useful for psychologists to develop, if they can, "objective" indices of the quality of psychological work. But, for heaven's sake, let's not measure ourselves by the number of articles we publish" (p. 198). Similarly, a founder of our field, Muzafer Sherif (1977), bemoaned the wheat versus chaff ratio in the skyrocketing volume of research and publication. Sherif went on to discuss many of the important issues now being reiterated by Leising et al. (2022) in the current era.

So I propose that we need another criterion added to the steps proposed in the target article. It might be termed: "Contribution to the comprehensive cumulation of knowledge and understanding." That is, published research should be more highly valued if it has an extensive knowledge of the century-long history of our field and builds deeply upon it, in an integrative fashion. Such an accomplishment does not simply mean citing relevant, important articles from the past, but also means keeping (or resurrecting) the best ideas and showing how they can now be further refined given new theories and new research methods.

A second step that would extend and complement the suggestions offered by Leising et al. (2022) is partly derivative from rightful concerns with the emphasis on the use of publication counts and large grant funding to decide which research projects (and which Psychology departments) are superior. This comparison process necessarily sets up a competition. Of course, some researchers and some departments are undoubtedly better at advancing the field of personality, and should be rewarded (or at least recognized) for their achievements. But the system does not have to be one of severe competition.

In part, this new structure would reward the incorporation of and cooperation with traditionally under-represented people and points of view, bringing both new perspectives and increased external validity. Further, if research publications are better appreciated if they involve multiple, cooperating laboratories—including cross-cultural considerations—then the quality and generality of personality research will likely improve. Thus *teamwork* could be an explicit criterion for evaluating faculty research.

That is, there sensibly could be less emphasis on individual comparisons and more attention to teamwork. This criterion would not have to apply to all faculty, because some sorts of scholarship are not as amenable to team projects. But overall lessened concern with constant comparative evaluation of individuals would likely help focus researchers on long-term quality.

Third and finally, the quality of research (and investigators) can be looked at not only in terms of instantly-applied criteria but also as to how impactful the work is on the field of personality science over time. Some research reveals instantly recognizable advances, but other research needs to pass the test of time, and the rewards can come later. To use an example from my own research: When, in 1993, we published an article finding that conscientiousness, in both childhood and adulthood, could predict longevity across the decades in longitudinal data (Friedman et al., 1993), no one knew whether this was a function of the particular cohort, or sample, or personality measures, or perhaps was even a chance finding. It was only after multiple other research teams turned to examining this finding and increasing their attention to the whole field of personality and health and longevity that the full impact of the initial findings could be appreciated, two decades later (Friedman & Kern, 2014).

Note that long-term retrospective judgment of research importance (e.g., 5 years, 10 years or 20 years later) would still impact promotions (advancement) of faculty. Promotions to tenure are usually based on 5 or more years of work, promotions to full professor are usually based on 10 or more years of work, and promotions to the very senior research positions (e.g. endowed chairs) are often based on 20 or more years of work. The point here is that more explicit emphasis on the long-term importance and value of personality research would encourage researchers to remember to take a long view of the possible long-term impact of high quality research.

In sum, the three steps sketched here are usefully included in an explicit enumeration of considerations of how to best reward high quality research in personality science.

References

Friedman, H. S., & Kern, M. L. (2014). Personality, well-being and health. *Annual Review of Psychology, 65*, 719–742. https://doi.org/10.1146/annurev-psych-010213-115123

Friedman, H. S., Tucker, J., Tomlinson-Keasey, C., Schwartz, J., Wingard, D., & Criqui, M. H. (1993). Does childhood personality predict longevity? *Journal of Personality and Social Psychology, 65*(1), 176–185. https://doi.org/10.1037/0022-3514.65.1.176

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Rubin, Z. (1978). On measuring productivity by the length of one's vita. *Personality and Social Psychology Bulletin, 4*(2), 197–198. https://doi.org/10.1177/014616727800400203

Sherif, M. (1977). Crisis in Social Psychology: Some remarks towards breaking through the crisis. *Personality and Social Psychology Bulletin, 3*(3), 368–382. https://doi.org/10.1177/014616727700300305

16 – Comment

The Plurality of Pathways in PERSONality Science

William L. Dunlop [1,2] †

[1] *Department of Psychology, University of California, Riverside, Riverside, CA, USA.* **[2]** *Department of Psychology, Aarhus University, Aarhus, Denmark.*

†Author deceased prior to publication of this paper.

Those who choose to read the target article by Leising et al. (2022) will be glad they did. I was struck by several of its features, not the least of which being the sheer amount of time, energy, and expertise required to pull this article off, and pull it off so well. Papers like this work best when they aspire for root-and-branch overhauls of the way things are, rather than minor revision. Our authors could be accused of little else. When presented as a unified whole, their ten steps signal nothing less than a seismic shift in the way research in personality psychology is to be conducted, evaluated, and rewarded.

The authors offer a tight, logical path forward for personality psychologists interested in conducting empirical research with the necessary credibility. Such specificity carries numerous benefits, many of which are outlined in detail by the authors themselves. It also leaves certain types of research, types of research that most assuredly should have a place in personality science, on the outside looking in. I am thinking here of personological pursuits examining whole persons and lives. Although researchers doing this type of work may have something to gain by engaging with our author' first five steps, the 'back five' appear to be written with someone else in mind.

Positioning the Person in Personality Psychology

I was unaware of the authors' manuscript until the finished product came across my desk. For this reason, I can only speculate as to how the admirable initiative they undertook ultimately led to the provision of these ten steps. One could imagine, though, it began by first considering the type of research most personality psychologists are conducting these days and then asking the question, *how can we make this research better*? This is, of course, both a reasonable and necessary question to ask, and ask routinely. An alternative approach would be to begin in a manner less situated in contemporary personality psychology and more aligned with the broad ambitions upon which the field was founded. Doing so may lead us to ask a question like, *what theories, method and standards are required to best understand the person?*

Let us not forget that the person represents the focus of personality psychology, or at least it should. Unlike many of the constructs routinely drawn from it, the person does not fit neatly within one any single conceptual and methodological orientation. A panoply of paradigms is required to capture the person in all its glory (see also,

Hopwood & Waugh, 2019; Wiggins, 2003). The pathway of steps proposed by Leising et al. (2022) is seemingly best positioned to guide certain personality researchers in their variable-centered pursuits (for review, see Donnellan & Robins, 2010) exploring relations among variables between people, in the interest of better "understanding and predicting important phenomenon" (Leising et al., 2022, p. 3). A different pathway must be traversed if we seek to understand whole persons and lives.

Zigging and Zagging

This latter route takes on less of a variable-centered, and more of a person-centered, character. Therein, willing travelers work to understand the complexities of a single life throughout time (e.g., Anderson & Dunlop, 2019; Elms, 1994; McAdams, 2011; McAdams & West, 1997; Singer, 2016). As Runyan (2005) noted, constructing a portrait of a whole person requires engaging in 'historical-interpretive' work. This work need be no less empirical than the type of research found along the pathway specified by Leising et al. (2022) once more, given their knowledge of, and interest in, people, personality psychologists are uniquely qualified to engage in the historical-interpretive work undertaken in the interest of making sense of a single life.

As a case study of the case study approach, consider McAdams' (2011) psychological portrait of George W. Bush. Drawing from what is generally known to be true about the structure of personality, this author explained many of the occurrences in Bush's life by way of reference to his trait profile (high extraversion and low openness to experience) and life narrative (one that was largely redemptive in nature). Had McAdams not been so well versed in personality science, he would not have been able to illustrate this life with such proficiency. Put differently, it is likely that only a personality psychologist, and a personality psychologist with strong personological leanings, would have been able to capture Bush's personality and life so adeptly.

Turning the Page in Personality Psychology

To summarize, for a certain type of research, and the certain type of researcher, the steps provided by Leising et al. (2022) represent a clear and worthwhile roadmap. As they note, undertaking some of these steps will require considerable energy and effort. Collectively, though, the benefits of progressing in this direction seem to outweigh the risks. At least it is worth a shot. As always, however, there is a rub. When adopting such a coherent and tight framework lines must be drawn between what is and what is not considered empirical research in personality science.

I believe the guidelines proposed carry merit for some, but not all, types of research that very much have a home in our field. To understand the person, we must navigate multiple pathways. Among these additional pathways, one of the most important takes as its focus the individual life through time. Conducting such 'historical-interpretive'

research requires following a set of steps distinct from those provided by Leising et al. (2022; e.g., Anderson & Dunlop, 2019; Elms, 1994). Doing anything less would be to jettison the study of whole persons and lives to our sister disciplines (e.g., anthropology, sociology). This may ultimately lead to a more coherent and unified personality science. This would also cause us to fall short of the ambitions of the grand theories that upon which the field was founded.

References

Anderson, J. W., & Dunlop, W. L. (2019). Executing psychobiography. In C.-H. Mayer & Z. Kovary (Eds.), *New trends in psychobiography* (pp. 11–33). Springer.

Donnellan, M. B., & Robins, R. W. (2010). Resilient, overcontrolled, and undercontrolled personality types: Issues and controversies. *Social and Personality Psychology Compass, 4*(11), 1070–1083. https://doi.org/10.1111/j.1751-9004.2010.00313.x

Elms, A. C. (1994). *Uncovering lives: The uneasy alliance of biography and psychology.* Oxford University Press.

Hopwood, C. J., & Waugh, M. H. (2019). *Personality assessment paradigms and methods: A collaborative reassessment of Madeline G.* Routledge.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

McAdams, D. P. (2011). *George W. Bush and the redemptive dream: A psychological portrait.* Oxford University Press.

McAdams, D. P., & West, S. G. (1997). Introduction: Personality psychology and the case study. *Journal of Personality, 65*(4), 757–783. https://doi.org/10.1111/j.1467-6494.1997.tb00533.x

Runyan, W. M. (2005). Evolving conceptions of psychobiography and the study of lives: Encounters with psychoanalysis, personality psychology, and historical science. In W. T. Schultz (Ed.), *Handbook of psychobiography* (pp. 19-41). Oxford University Press.

Singer, J. A. (2016). *The proper pirate: Robert Louis Stevenson's quest for identity.* Oxford University Press.

Wiggins, J. S. (2003). *Paradigms of personality assessment.* Guilford Press.

17 – Comment

The Importance of Acknowledging Multiple Research Paradigms and Diversity, Equity, and Inclusion (DEI) for Improving Personality Science

Theo A. Klimstra[1]

[1] *Department of Child Study and Human Development, Tufts University, Medford, MA, USA.*

Leising et al. (2022) present an ambition plan for improving personality science. Their paper begins with a nuanced picture of what "good" science should look like, including greater openness about its shortcomings. I agree that it would be great to see criticism towards our own past work being embraced as a strength rather than a weakness.

Leising et al. (2022) also raise important concerns about the current focus on numbers of publications, citations, and the amount of funding received. This often plays out favorably for researchers who had access to good-quality data early in their career and gives such researchers an unfair head start. I am among those who have benefitted from this broken system. Furthermore, I agree that these numbers are easily manipulated, for example by researchers who disproportionally suggest citations of their own articles when serving as editors and reviewers. Therefore, I agree that we need drastic changes.

However, I believe that the steps and reward scheme proposed by the authors do not address some other underlying problems of our field. I propose that to improve personality science, we need to reflect on the scientific paradigm dominating the field and – related to that – the lack of diversity (of all kinds) among personality researchers.

In personality science, there is a predominant focus on quantitative approaches. For example, the rise of the Five-Factor Model was clearly driven by a belief in the relative objectivity of such approaches (i.e., factor analysis; Goldberg, 1990). Leising et al.'s (2022) procedure of counting "good" qualities of articles, with more "important" qualities weighing heavier, further fits in this tradition. These practices align with post-positivism, representing a view that significant, substantial, and replicable effects as observed in statistical analyses represent an objective truth, or at least a close approximation thereof.

Post-positivism can be mistaken for being the only "real" scientific paradigm, which can lead to reduced interest in, knowledge about, and appreciation of alternative approaches to science. Perhaps for this reason, Leising et al. (2022) misrepresent qualitative research with their comment that '…we think that qualitative research may benefit substantially from stricter formalization. A prime example would be the formalization of individual belief systems.' First, "qualitative" methods represent a wide variety of methods. For many of those methods, there are guidelines for what should be reported and how (e.g., Levitt et al., 2018). Second, qualitative researchers often already make their individual belief systems explicit, for example by referring to theories that imply adher-

ence to a particular belief system (e.g., Critical Race Theory; Delgado & Stefancic, 2017). Therefore, qualitative researchers do not need recommendations on stricter formalization. Instead, a greater appreciation of scientific paradigms other than post-positivism, and methods other than quantitative ones (i.e., qualitative methods, mixed methods), may lead to better, more inclusive, personality science.

Other scientific paradigms, such as constructivism (e.g., personal construct theory; Kelly, 1955), have previously been used for the study of personality. More recently, research aligned with the transformative paradigm (e.g., Mertens, 2007) or using practices from different paradigms (the pragmatic approach; e.g., Feilzer, 2010) seem to become increasingly common. Reviews by Arshad and Chung (2022) and Atherton et al. (2021) are examples of this. The reward scheme of Leising et al. (2022) may underappreciate such work, especially when pursued using qualitative methods. For example, criterion 9b ('has an expected type I error rate of ≤ .05 and type II error rate of ≤ .20, based on realistic effect size estimates') can only be applied to quantitative research.

From paradigms other than post-positivism, Leising et al.'s (2022) call for consensus regarding research goals, terminology, measurement practices, data handling, and the current state of theory and evidence can also be perceived as problematic. For example, the transformative paradigm holds 'that realities are constructed and shaped by social, political, cultural, economic, and racial/ethnic values indicates that power and privilege are important determinants of which reality will be privileged in a research context' (Mertens, 2007, p. 212). This perspective directly points to potential problems caused by underrepresentation in personality science of most groups other than white male-identifying researchers from North American and Western European countries and represents the underlying philosophy of science guiding recent critical reviews of our field (e.g., Arshad & Chung, 2022). Arshad and Chung (2022) provide excellent recommendations for how to address these, and related, problems.

Systemic inequality beyond personality science may have affected even the very constructs we focus on. A good example of this could be the Big Five. Derived from lexical analyses on the English language (Goldberg, 1990), a still large set of adjectives were grouped into a more manageable set of broad dimensions using factor analyses (the Big Five). However, Goldberg's (1990) samples primarily included English-speaking college students. To get into college, one needs to fluently use language fitting with what is privileged in one's educational systems. Furthermore, what is considered "standard" language is decided by humans with language use by marginalized groups often considered (by the dominant group) as inferior and non-standard (e.g., Siegel, 2006). Therefore, words that are meaningful to many in describing personality likely were excluded from lexical analyses. Hence, a five-factor structure of personality might reflect a useful way of thinking about broad individual differences among a particular (relatively privileged) segment of the population in countries where the lexical method and subsequent factor analyses were employed, but Big Five measures may poorly resonate with others. This is

not to say that the Big Five are not useful but to think that they came about in a bias- and value-free manner is naive.

This commentary is meant as an illustration of how relying on consensus for evaluating "quality" could lead to perpetuated inequality. The past and (to a somewhat lesser degree) current climates have been rather non-inclusive in personality science (Atherton et al., 2021), which likely led to values upheld by a small group being privileged. From any other perspective than a post-positivist one, these omissions are problematic. Therefore, we can only make personality science better if we recognize and reflect on the scientific paradigms we use and become more inclusive.

References

Arshad, M. & Chung, J. M. (2022). Practical recommendations for considering culture, race, and ethnicity in personality psychology. *Social and Personality Psychology Compass, 16*, Article e12656. https://doi.org/10.1111/spc3.12656

Atherton, O. E., Chung, J. M., Harris, K., Rohrer, J. M., Condon, D. M., Cheung, F.,Vazire, S., Lucas, R. E., Donnellan, B., Mroczek, D., Soto, C. J., Antonoplis, S., Damian, R. I., Funder, D., Srivastava, S., Fraley, R. C., Jach, H., Roberts, B., Smillie, L., …Corker, K. S. (2021). Why has personality psychology played an outsized role in the credibility revolution? *Personality Science, 2*, Article e6001. https://doi.org/10.5964/ps.6001

Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction* (3rd ed.). New York University Press.

Feilzer, M. Y. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research, 4*(1), 6–16. https://doi.org/10.1177/1558689809349691

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Kelly, G. S. (1955). *The psychology of personal constructs.* Norton.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *The American Psychologist, 73*(1), 26–46. https://doi.org/10.1037/amp0000151

Mertens, D. M. (2007). Transformative paradigm: Mixed methods and social justice. *Journal of Mixed Methods Research, 1*(3), 212–225. https://doi.org/10.1177/1558689807302811

Siegel, J. (2006). Language ideologies and the education of speakers of marginalized language varieties: Adopting a critical awareness approach. *Linguistics and Education, 17*, 157–174. https://doi.org/10.1016/j.linged.2006.08.002

18 – Comment

Avoid Allowing the Ends to Justify the Means

Lauren Lesko[1], Cora Miller[2]

[1] *Department of Psychology, University of California, Los Angeles, CA, USA.* **[2]** *Institute for Society and Genetics, University of California, Los Angeles, CA, USA.*

Background

The points outlined in this article are at once both commonsensical and revolutionary. Over many years we, as researchers, have developed scientific practices of objectivity and reason. Nevertheless, we are humans with emotions and biases that impede our progression toward high-quality science. We need structures in place to ensure that all parties are treated equitably and quantity of research does not supersede quality. This article addresses ten excellent examples of the places where our human nature gets most stuck when aiming for quality science, with one notable exception: *respect for participants*. Increased costs (both temporal and monetary) reduce the quantity of research that can be produced, but contribute to the ideals with which the IRBs were founded and the Belmont Report was written: respect for persons, beneficence, and justice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978). "The essential conflict in research is the duty to avoid allowing the ends to justify the means" (Moon, 2009). Too often as researchers, we prioritize the ability to conduct more studies more rapidly above the fair treatment of participants.

Personality science is different from the oft-referenced idealized science of physics in that our objects of study are humans. When studying an electron, the researcher doesn't need to worry about taking advantage of the electron, making it unduly uncomfortable, or compensating it in any way. In the rush to quantity vs quality of publications in personality psychology, it is impossible to properly respect participants. Legal and IRB protections are not necessarily adequate to ensure the respectful treatment of participants. The world changes rapidly, and new repercussions of old research methods constantly emerge. For example, with the advent of online data collection and subsequent rise of digital data insecurity, IRBs have begun enforcing stricter standards of digital data protection. In the time before these stricter standards but after digital data collection had begun, was it responsible for researchers to use insecure data collection and storage methods, even if those methods were easier and cheaper? Imagine that a researcher

identifies a method of better respecting and protecting their human participants, but this method is not required of the researcher to perform, and it would entail extra work either for the researcher or their already overburdened students or assistants. If this researcher asked you for advice on whether they should abandon this method or pursue it, what would you say?

Why This Matters

Treating participants better creates higher quality research. As humans, we are biased towards short-term visions, but to improve the quality of the literature long-term, we must treat participants well. Our society's current wave of anti-science sentiment is likely rooted, at least in part, in the disconnect between academia's Ivory Tower and the common citizen. Increasing participant protections and increasing the information that participants are provided decreases the power differential between researcher and participant and increases the participant's respect for and understanding of scientific research. If we want our work to have tangible effects on the world, we need people to respect the information that research provides. If we want to continue these lines of work, we must continue to acquire funding and participants. Both of these will be much easier to ensure long-term if we treat participants with respect and beneficence.

Actionable Steps

Researchers should properly compensate their participants and increase data sharing with participants. Paying participants a minimum wage, let alone a living wage, for their time spent participating is neither required nor rewarded, but will increase the quality of our science. Data sharing is even less commonly practiced than paying participants a minimum wage. Some might argue that we are not qualified to 'diagnose' participants with personality types, and participants might take the word of science as law, but recent research suggests that participants understand more than we think they do, and they are unlikely to be alarmed at the results given to them (National Academies of Sciences, Engineering, and Medicine, 2018). The more we attempt and request to share data with participants, the more likely it will be for survey software programs like Qualtrics to build this feature in. This practice is successfully becoming more common in other fields (e.g., environmental research (Boronow et al., 2017), and even healthcare––as of May this year, patients in the US can now access their medical information online). In order to increase the quality of our science, we should treat participants with respect, and two ways to do so are properly compensating participants and sharing their data with them.

To reduce the costs associated with these practices, participant care should be actively considered as a sign of research quality. Hiring committees and tenure review committees should include a category of their rubric dedicated to the researcher's track record of avoiding "ends justify the means" research. Grant reviewers should acknowledge that

the purpose of all grants is to improve human lives, and when participants are humans, participants deserve adequate compensation for their time. This should be at least equivalent to minimum wage for the time required, and preferably at least equivalent to the living wage for local participants. Respect for participants should be included as a sign of research quality in addition to the other points highlighted in the article.

Conclusion

As researchers, we value the pursuit of knowledge very highly. An even grander goal, however, is the purpose that we pursue knowledge for: to improve the lives of our fellow humans. Our direct focus on the more immediate pursuit of knowledge can blur our vision of the ultimate goal of improving human life, often encouraging focus on rapid and plentiful production of knowledge at the cost of the values about which the Belmont Report was written: respect for persons, beneficence, and justice. Putting formal structures in place to ensure a greater focus on the wellbeing of people we recruit as participants will ultimately increase our society's respect for and understanding of science, which can only improve the quality of our science.

References

Boronow, K. E., Susmann, H. P., Gajos, K. Z., Rudel, R. A., Arnold, K. C., Brown, P., Morello-Frosch, R., Havas, L., & Brody, J. G. (2017). DERBI: A digital method to help researchers offer "right-to-know" personal exposure results. *Environmental Health Perspectives, 125*(2), A27–A33. https://doi.org/10.1289/EHP702

Moon, M. R. (2009). The history and role of institutional review boards: A useful tension. *AMA Journal of Ethics, 11*(4), 311–316. https://doi.org/10.1001/virtualmentor.2009.11.4.pfor1-0904

National Academies of Sciences, Engineering, and Medicine (2018). *Returning individual research results to participants: Guidance for a new research paradigm.* https://doi.org/10.17226/25094

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978, January 13). Tab 2, Box 33 [Meeting Transcript]. Georgetown University Archives.

19 – Comment

One Measure to Rule Them All? Commentary on "Ten Steps Toward a Better Personality Science"

Kai T. Horstmann[1], Matthias Ziegler[1]

[1] *Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany.*

Leising et al. (2022) suggested ten steps toward a better personality science. Without a doubt, following or implementing the spirit underlying these ten steps would further improve the quality of personality science, it would make life easier for a lot of (young) researchers, and would culminate in a more integrative and manageable research field. We agree with this spirit and most of these propositions, cannot comment on some, but disagree with one. Specifically, Leising et al. (2022) suggested that the field should pursue common measurement practices, especially with respect to the measurement of the Big Five personality traits. The authors write "there needs to be a single, standard way of assessing the Big Five personality factors, as well as standard ways of assessing everything else that is considered worth assessing." (p. 12). The authors suggest that this should be achieved to avoid jingle-jangle fallacies, to facilitate meta-analyses, and to allow for cumulative knowledge building. Although we strongly agree with the goals that should be achieved, we disagree with the proposed solution, specifically with the idea of one standard measure to assess personality. Instead, we suggest an alternative that could be implemented for any type of test that uses individual items (e.g., statements about oneself as in personality tests, stimuli in reaction time tasks, intelligence tests items).

Psychological tests that assess one variable are usually developed for a specific purpose (e.g., comparing groups vs. assessing individual levels), and for a specific population (e.g., younger people, older adults, from different cultures) (Ziegler, 2014). To achieve this, items are, for example, tailored towards the language level of the targeted participants or to reflect specific living circumstances. As a result, different measures fulfil the needs of different researchers. Using just one tool would leave a gap in our tool kit. To provide a metaphor: There is no one tool to measure distance. At home, where precision does not matter too much, one can use a simple yardstick, for longer distances, a laser distance meter can be used. Of course, this example neglects the problem of different scales of measurement in psychology which potentially prohibit comparing results derived with different measures. A risk we see when using one authoritative measure alone is an artificial reduction of the potential construct breadth. The field of personality research underwent a similar development before, when the original Big Five were developed, based on items and adjectives selected by specific researchers. The result was that the field, after years of otherwise very productive work, realized that many,

often darker aspects of personality, were not reflected (Thalmayer, Saucier, & Eigenhuis, 2011).

Instead of forcing the same measure upon everyone, we would suggest providing, in the long run, an online repository of measures *and* data. Such a data base should ideally contain individual items and their scores, plus background information on the sample characteristics and the survey method (e.g., rating scale used, language, introduction, online vs. offline). Similar to automatically updating websites such as curatescience.org, such a website could generate item and scale statistics for different samples, estimates of reliability, validity, and even measurement invariance across different populations. To this end, the website would need to match uploaded items to existing scales or other items, compute (latent) scale composites, and provide (latent) correlations to other scales, if available.

Such a website or repository would ideally provide three services: First, researchers could search for items and scales that have been used and validated in samples similar to the ones they intend to collect. Second, items (without data) could be uploaded, and be compared to existing items based on linguistic features, thereby avoiding the jingle-jangle fallacies that Leising et al. (2022) point out as a potential threat to the test score's validity (similar to https://rosenbusch.shinyapps.io/semantic_net/; Rosenbusch, Wanders, & Pit, 2020). Finally, researchers could upload their data *and* their items to share both at the same time in a well-documented format.

From our standpoint, such a dynamic approach has several advantages over one authoritative measure to rule them all:

1. It is continuously possible to detect jingle-jangle-fallacies. If someone develops new items, they can directly see where in the existing item-universe their new measure would fall.
2. If someone is convinced, they need to develop a new measure for an existing construct, they can still do this. This repository would immediately allow them to provide evidence that their new measure would outperform existing measures (e.g., be measurement invariant across groups for which previously no invariant measure existed).
3. In the long run, an online repository with well-documented data would allow developing measures for one construct that consisted of different (i.e., in terms of wording), yet equivalent (in terms of psychometric properties) items. Last but not least,
4. the ultimate goal proposed by Leising et al. (2022) to have one measure for one construct is still possible, and even more likely, as it will be data driven instead of a measure that is chosen for other reasons (e.g., authority of the authors, prestige of the journal, because it has always been used).

Finally, we envision an even more ambitious goal: For each scale, it could be possible to generate a unique identifier or code (similar to a DNA sequence or a DOI) that unambiguously represents the items that have been used in a scale, the rating type scale used, and even the instructions. Published in an article, this code could then be submitted to the website, generating the exact questionnaire that was used (e.g., for reuse in online survey platforms such as formr.org, or as a word-template), additional statistics, and also further publications in which the scale has been used.

We agree, this proposal is not modest, and it will require financial investment as well as the investment and commitment of our community to share their data. Luckily, however, this project could be launched right away with the existing open data from personality research (Horstmann, Arslan, & Greiff, 2020). If successful, it could serve as a beacon for the integration of different measures in personality psychology and beyond.

References

Horstmann, K. T., Arslan, R. C., & Greiff, S. (2020). Generating codebooks to ensure the independent use of research data. *European Journal of Psychological Assessment, 36*(5), 721–729. https://doi.org/10.1027/1015-5759/a000620

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods, 25*(3), 380–392. https://doi.org/10.1037/met0000244

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment, 23*(4), 995–1009. https://doi.org/10.1037/a0024165

Ziegler, M. (2014). Stop and state your intentions! *European Journal of Psychological Assessment, 30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

20 – Comment

Why We Sometimes Need Different Measures

Jens Mazei[1], Marc Mertes[1], Ann-Kathrin Torka[1], Joachim Hüffmeier[1]

[1] *Department of Social, Work, and Organizational Psychology, TU Dortmund University, Dortmund, Germany.*

We fully agree with the general tenor of the article by Leising et al. (2022)—that research *quality* is vital—as well as with most of the suggested actions (e.g., implementing Open

Science Practices and conducting more replications; e.g., Nelson et al., 2018). Leising et al. (2022) outlined central issues and specific next steps for increasing the rigor and relevance of our discipline—including those that concern the measurement of constructs (Step 3). In fact, measures are a key building block of psychological science—literally. With our measures, we collect data that are subjected to statistical analysis, which then provide results used to draw inferences about hypotheses, which ultimately inform theories to organize "knowledge" in psychology (e.g., Bacharach, 1989).

Yet, always pursuing "standard ways of assessing everything else that is considered worth assessing" (Leising et al., 2022, p. 12) can leave crucial (yet overlooked in their article) benefits untapped. Admittedly, pursuing standards in measurement can make perfect sense, for instance, when various measures of the same construct exist but differ in quality. Objectivity, reliability, and validity have to be examined, so that a single best measure, if it exists, can prevail (Leising et al., 2022; we also agree with the problems associated with unsystematically adapting measures; pp. 11–12). However, the opposite approach—systematically pursuing a "diversity" in measurement—has notable benefits unconsidered by Leising et al. (2022). Diversity in measurement can contribute to rigor and informativeness of psychological science—precisely the kind of goals addressed in their article. It depends on the aims of a research endeavor whether a standard measure is needed or not (see Leising et al., 2022, p. 13).

Benefits of Using Different Measures

Not all findings in psychological science are robust (e.g., Open Science Collaboration, 2015). "Shaky" findings can raise doubts about the knowledge in psychology, but also stimulate debate about actions to determine and increase its robustness. In addition to clearly helpful Open Science Practices (e.g., Nosek et al., 2015; Protzko et al., 2020), conducting replication studies is a key means to do so (as Leising et al., 2022, also point out).

LeBel et al. (2018) illustrate that replication studies can be more ("direct replications") or less ("conceptual replications") similar to an original study (Hüffmeier et al., 2016). Conducting a conceptual replication means systematically *varying* elements of the methodology—including the used measures (e.g., LeBel et al., 2018; Schmidt, 2009). Such variations enable us to examine potential boundary conditions of a finding, or, conversely, its generalizability (e.g., Hüffmeier et al., 2016; LeBel et al., 2018). The question of generalizability is critical in many areas of research, as certain constructs (e.g., aggression) can manifest themselves in different ways. Hence, pursuing a diversity in measurement, as is done in serial replications, is essential for testing the robustness of findings.

As another case in point, for at least some constructs, measures can only be imperfect, as they are "deficient" and/or "contaminated" (e.g., in the case of job performance; e.g., Iaffaldano & Muchinsky, 1985). A given measure may either not capture all critical elements of a construct (deficiency) and/or (additionally) capture aspects that are not

actually part of a construct (contamination; as Leising et al., 2022, point out, even measures of scholarly achievement—the number of publications or citations—are "strongly contaminated"; p. 33). Hence, if a construct cannot be adequately captured with one standard measure, it is again advisable to use multiple measures and derive conclusions about underlying propositions based on the emerging cumulative evidence.

Conclusion

Although we greatly appreciate the article by Leising et al. (2022) and support most of their suggested actions, for the reasons discussed above, we deem it important to carefully consider when and why working toward a "standard way" to measure a construct is truly advisable. Notably, Leising et al. (2022, p. 13) briefly discussed a benefit of using multiple measures: "reduc[ing] the risk of a 'mono-method' bias." Yet, given the importance of measures in our discipline, one should not give short shrift to additional crucial benefits of developing and using multiple measures for the same construct. Doing so provides an important opportunity to strengthen psychological science and actually facilitates another crucial step discussed by Leising et al. (2022): conducting replication studies.

References

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*(4), 496–515. https://doi.org/10.2307/258555

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology, 66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97*(2), 251–273. https://doi.org/10.1037/0033-2909.97.2.251

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science, 2*, Article e6029. https://doi.org/10.5964/ps.6029

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . .Yarkoni, T. (2015).

Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S., Walleczek, J., & Schooler, J. W. (2020). *High replicability of newly-discovered social-behavioral findings is achievable.* PsyArXiv. https://doi.org/10.39227/osf.io/n2a9x

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90–100. https://doi.org/10.1037/a0015108