PERSONALITY Science

PsychOpen GOLD

Theory [Invited Paper]

Check for updates

Ten Steps Toward a Better Personality Science – A Rejoinder to the Comments

Daniel Leising[1], Isabel Thielmann[2], Andreas Glöckner[3], Anne Gärtner[1], Felix Schönbrodt[4]

[1] *Faculty of Psychology, Technische Universität Dresden, Dresden, Germany.* **[2]** *Department of Psychology, Universität Koblenz-Landau, Landau, Germany.* **[3]** *Department of Psychology, Universität zu Köln, Cologne, Germany.* **[4]** *Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany.*

Personality Science, 2022, Vol. 3, Article e7961, https://doi.org/10.5964/ps.7961

Received: 2021-12-20 • **Accepted:** 2022-01-10 • **Published (VoR):** 2022-05-06

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

Corresponding Author: Daniel Leising, Faculty of Psychology, Technische Universität Dresden, 01062 Dresden, Germany. +49 351 463 33149. E-mail: Daniel.Leising@tu-dresden.de

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).

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

We respond to the comments (https://doi.org/10.5964/ps.9227) on our “Ten Steps” paper (https://doi.org/10.5964/ps.6029), focusing on the most prominent themes: (1) What motivates scientists?, (2) Consensus-building (Is our field ready? May there be adverse side-effects? How shall we do it?), (3) How may institutional change be facilitated?, (4) Diversity (of participants, stimuli, methodology, measures, and among researchers), (5) The reliability of our proposed scoring system, and (6) The real-world relevance of personality research. We stand by our call for more concerted consensus-building and offer a few clarifications in this regard. We also issue four specific calls to action to our colleagues in the field: (a) specify legitimate paths to greater consensus, (b) explicate what constitutes good “qualitative” research, (c) help establish a widely used, public domain item database, and (d) determine what the most important contemporary goals of personality research are.

Keywords

personality, research quality, research evaluation, quality criteria, reward, incentive, research methodology, consensus building

Relevance Statement

This rejoinder reflects our perspective on some of the major themes that emerged in the course of the lively debate over our target article. We think this debate has showcased numerous urgent needs for improvement in our field. The rich, complex, detailed, and sometimes heated discussions that ensued have been very illuminating and constructive, in our view. The pre-liminary outcome is a relatively clear and comprehensive vision of how personality research might be improved. We encourage all of our colleagues to help move the field in the direction of greater credibility and efficiency together. We offer four specific suggestions for issues that we think are important to tackle, but these are by no means exhaustive. All of this will be very hard work, but it needs to be done, and the rewards may in fact be lasting and very substantial. What an exciting time this is to be a (personality) psychologist!

Key Insights

- Consensus-building will be vital for moving our field forward
- Legitimate paths toward consensus will have to be specified
- Relevant issues: Roadmap, inclusiveness, distribution of power, transparency
- We partly revised our reward scheme and keep developing it

We would like to begin by expressing our sincere gratitude to the many people who were part of this intense process, and who devoted their time and energy to it. In our view, their generous involvement and the lively debates that ensued attest to (a) the vitality of our field, (b) the willingness of our colleagues to improve on the ways in which we do science, and (c) the trust that people put into this new journal. We especially thank John Rauthmann, who initiated this project, and Mario Gollwitzer, who oversaw the editorial process. We thank the three official reviewers (selected by the journal) and the 16 colleagues who volunteered additional reviews before the target article was even submitted. They all helped make this article a lot better. We also thank the 19 colleagues who provided post-publication comments on the target article, which we found to be equally valuable. The purpose of this rejoinder is to address the major themes highlighted in those comments.

One of the provisionary outcomes of this process is an adapted Version 2 of our reward scheme, which may be accessed using this link (https://osf.io/mbgq3/). Within the same OSF project, one also finds a dataset and an R script from a pre-registered pilot study in which we checked the inter-rater reliability of the first version of our reward scheme (see below). We decided to keep developing and testing this reward scheme, and plan to present an improved “ready-to-use” version (including a codebook and further information on its application) by fall 2022. The progress of this development may be tracked using said link.

What Motivates Scientists?

Several commentators and reviewers (e.g., Bromme, Gollwitzer) seemed to call us out for allegedly suggesting that scientists are primarily or even exclusively motivated by extrinsic rewards such as research productivity metrics, or their more distal consequences such as job security and financial gain. We neither made nor would make such a claim. To the contrary, in our target article we deliberately laid out all the purely intrinsic rewards that working in academia has to offer, in considerable detail.

However, is there any good reason for employing a research evaluation system that is so obviously *at odds* with what science should be about, namely robust, incremental knowledge gain? Would it not be more reasonable to *align* incentives as closely as possible with our scientific ideals? In fact, this was the main purpose of the reward scheme that we proposed. Researchers are human beings, too - why should they be exempt from the proven, strong effects of external rewards? And the situation is even worse for early career researchers, whose ability to even only continue their careers depends most strongly on their willingness to align their behavior with the existing reward structure.

One aspect of this that we clearly did not talk about enough in our paper is *risk*. Pursuing the path of more open and rigorous science does entail significantly greater risks than current mainstream practices do: The risk of publicly being proven wrong when the results one obtains contradict one’s own theory or pre-registration. The risk of not being able to replicate an effect that one previously found and believed in. The risk that someone else finds a mistake in one’s published analysis code. In addition, it will also require considerably greater effort to even *get* to those sobering experiences. There are reasons why most open science practices have still not become mainstream.

Perhaps unexpectedly for some, aiming for more consensus may entail considerable risks, too: The risk of finding out that some or even most others in the field do *not* share one’s own views as much as expected. Or the risk that those others actually have the better arguments or data supporting their views. Our own experiences with this type of work are yet fairly limited (e.g., adversarial collaborations in Glöckner & Pachur, 2012; Marewski, Bröder, & Glöckner, 2018), but these experiences do not at all confirm that consensus-building is easy and friction-free. Continuing to *avoid* this type of work is almost certainly the more comfortable choice. Reports from other fields (Zachar & Kendler, 2012; Zachar, Krueger & Kendler, 2016) clearly support this view.

Consensus

It seems that our suggestion to more actively foster consensus-formation in personality science met with the gravest concerns from the commentators (Asendorpf & Gebauer; Corker; Denissen & Sijtsma; Fernandes & Aharoni; Gollwitzer; Hagemann; Hilbig et al.). We stand by our proposal. In this section, we will highlight some of the concerns that were brought forward, and offer suggestions as to how these concerns may be addressed constructively.

Is our Field Ready for Consensus-Building?

Several commentators (Adler; Beck et al.; Corker; Hilbig et al.; McLean & Syed) questioned the adequacy of the current literature in our field as a basis for consensus-building. We largely share these doubts, and concede that this important point may not have become clear enough in our paper: If one argues that scientific consensus must be based on robust evidence rather than extraneous factors (e.g., power differences between researchers), while at the same time lamenting the questionable value of many empirical contributions in the field, then the conclusion can only be that we first need to improve the quality of the *pieces* of the puzzle (i.e., individual papers) before attempting to assemble them into a more coherent whole.

Hagemann rightly pointed out that not all of our five types of consensus are created equal. Specifically, Criteria 1 to 4 in the first version of our reward scheme pertain to types of consensus that should help *foster* incremental knowledge gain, whereas Criterion 5 is about the type of consensus that would *evidence* incremental knowledge gain. So, there is a certain logical order to this that probably did not become clear enough in our paper, either. The following order of improvements would be most plausible in our view: To even *permit* incremental knowledge gain, researchers first need to better harmonize their goals (Criterion 1), terminology (Criterion 2), measures (Criterion 3), and ways of handling data (Criterion 4). The more this is achieved, the more subsequent empirical research on substantive (i.e., non-methodological) questions may become suitable for being integrated (e.g., by meta-analysis; see Corker), to ultimately yield a more consensual picture of the current state of knowledge. Adhering to criteria for credible empirical contributions (Criteria 6-10) would certainly help achieving that. Although the order of entries in our rewards table is ultimately arbitrary, we re-arranged it in the revised Version 2 to better reflect this more “natural” way in which our field may progress.

May Consensus-Building Have Adverse Side-Effects?

Perhaps the gravest concerns expressed by many commentators pertained to the possibly detrimental effects of establishing consensus, promoting consensus, or even only striving for consensus. For example, it was argued that seeking and rewarding consensuality may promote the wrong kind of (e.g., conformist, opportunistic, intellectually lazy, riskaverse) attitudes and corresponding behaviors among researchers (Asendorpf & Gebauer; Denissen & Sijtsma; Hilbig et al.). This criticism made us aware that the first version of our reward scheme was in fact lacking rewards for systematic *challenges* to established consensus. In the revised Version 2 of our reward scheme, we therefore included a set of new criteria (1c, 2c, 3c, 4c, and 10c) to explicitly cover this aspect. Given that challenging a consensus will usually be more difficult than just complying with it, we suggest using substantially greater rewards (2.0 points) for the former, as compared to the latter (0.5 points). Note, however, that posing challenges to consensus requires documentation of consensus first, including a sufficient degree of specification to *permit* refutation to begin with (Corker; Scheel, 2022).

Some commentators argued that consensus-building may be detrimental to innovation and creativity (Beck et al; Denissen & Sijtsma), or even shut down further scientific inquiry (Hogan, Harms & Sherman). We disagree. In our view, scientific progress is marked by the constant necessity to *balance* innovation and consolidation. Yes, unmitigated consensus leaving no room for innovation anymore may stall progress; but constant innovation with no discernible consensus ever emerging is not conducive to scientific progress, either. In fact, innovation is *not* even a scientific value in itself - it only is to the degree that it leads to, or at least may lead to, a demonstrable *improvement* in the respective knowledge base. Demonstrating such improvement requires that current assumptions and the evidence supporting them are well-documented. Also, it is much more difficult to convincingly claim that you are contributing something new and important to the field if it has not yet been documented what the current and — in your eyes — unsatisfactory state of the art *is*. In their comment, Beck et al. argue that "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". This risk is real. What Beck et al. fail to mention, however, is that challenges to consensus are *logically impossible* when no-one knows what the consensus is, or whether it even exists. Thus, our call for more intentional consensus-building is by no means an attempt to shut down scientific debates in our field. Rather, it should be viewed as an attempt to better structure those debates, and thus make them more traceable and efficient.

On a side-note, our call for more building and embracing of consensus seems to have reminded some commentators of the still ongoing debate around the factorial and cross-cultural validity of the Big Five personality factors (e.g., Galang & Morales; Hogan et al; Klimstra). Suffice it to say that none of us has any particular loyalty to that model, nor do we have any stakes in its proliferation. One of us (IT) has repeatedly argued that a six-factor structure is more valid (e.g., Thielmann et al., 2021). One of us (DL) is not even convinced that the source of item-covariation in the relevant studies lies in the targets (Letzring et al., 2021; see also Borkenau, 1990). So, our call for greater openness to explicit consensus-formation clearly is *not* a covert appeal to finally give in and accept the Big Five as law of the land.

How Shall We Build Consensus?

We as personality psychologists, and psychologists more generally, have very little experience with consensus-building. Fortunately, the necessity for more explicit consensus-building is now being increasingly recognized in many different branches of science, so we may learn a lot from the experience of colleagues who are doing this kind of work already (e.g., Aczel et al., 2021; Hagger et al., 2016; https://forrt.org/glossary/; http://www.ich.org). Trailblazing in this regard are the ManyLabs-style collaborations, for example. These necessarily require some local consensus amongst the dozens or even hundreds of team members involved, regarding (a) what the most important research questions are, and (b) what standard measures shall be used by everyone. Answers to these questions are jointly developed, building upon diverse viewpoints from different countries, in an inclusive bottom-up process (e.g., Visser et al., 2021).

To avoid the “bad” (i.e., premature, ill-founded, shallow, oppressive) type of consensus that several of the commentators warned against, consensus-building processes will have to be *professionalized.* We do not have a “perfect recipe” for how to achieve this. Still, we would like to offer a few plausible suggestions and clarifications:

First, Beck et al. expressed their concern that consensus-building may ultimately reward “well-known, eminent, and productive individuals (in terms of publication numbers)” and that early career researchers (ECRs) and researchers from underrepresented backgrounds (RUBs) may have too little say in it. McLean and Syed also recommended being wary of how power is distributed in the field. All of this touches on the *diversity* issues that we will highlight further below. Like most other scientific activities, consensus-building undeniably has a social, and even a political component to it (Zachar & Kendler, 2012), because it has to be accomplished by groups of people. Strong guidelines and mechanisms must therefore be established to minimize the influence of extraneous factors in such processes.

Second, only the group of authors who explicitly sign off on a consensus document may be viewed as embracing it. Thus, no group of authors may ever speak for anyone but its members, and no consensus may ever be viewed as all-encompassing. Note that, by virtue of this approach, explicit consensus-building efforts may actually help delineate the *limits* of consensus that currently exist. Third, a group of scientists that sets out to find consensus amongst themselves may actually end up with not one but several competing versions of consensus documents. Ideally, such a group would then offer some insight as to how the best of these competing versions may ultimately be identified. Fourth, even if a group’s effort results in a single consensus document, this may still contain some points that are embraced by more group members than other points, and this can be made explicit.

Fifth, listing the individual propositions constituting the consensus point by point in a very fine-grained manner (and even numbering them) is highly recommendable (e.g., Letzring et al., 2021), because it makes it easier to call out the particular elements of a consensus that one deems questionable. Sixth, any consensus should always be regarded as preliminary and versioning should probably be the norm. In line with this approach, we publish a revised version of our proposed reward scheme along with this rejoinder, as an online Supplement.

Seventh, to repeat a few points from our Ten Steps paper, a proper consensus-building process will require a clear roadmap (e.g., as to how members of the group of authors will be recruited, and what decisions will be made by whom at what points in time), explicit mechanisms to resolve disputes, ways of ensuring fair representation of different viewpoints at the outset (Oreskes, 2019; Fedorenko et al.), as well as transparent documentation (e.g., Aczel et al., 2021; Hagger et al., 2016), and, ideally, independent and impartial oversight. Additional measures may be taken to limit the possible influence of groupthink and conformism (Lane et al., 2021). We believe that we as psychologists should be particularly well-qualified to outline the basic parameters of credible consensus-building processes. Therefore, we encourage the readers of this journal to explicitly embrace this as an important meta-scientific research topic (e.g., by writing another target article on the issue for this journal). This may be of great interest to colleagues from other fields, as well.

How to Facilitate Institutional Change

Several commentators (Fedorenko et al.; Friedman; Schmitt) highlighted the fact that calls for reform similar to ours have repeatedly been made in the past, sometimes decades ago. Obviously, there are powerful factors at work that slow down or prevent the implementation of improvement measures that have long been recognized as desirable. For example, as long as the current academic reward structure remains in place, individual researchers and institutions that deviate from mainstream research and publication practices (prioritizing quantity and expediency over quality) will clearly incur disadvantages for themselves. This social dilemma is indeed very real, has been described several times before (e.g., Nosek & Bar-Anan, 2012), and is also addressed in some of the comments on our paper (e.g., by Schmitt). It does have to be dealt with. As we briefly said in our Ten Steps paper, two main routes to achieving meaningful change can be identified. They are not mutually exclusive, but can be pursued simultaneously.

The first route may be called the “top-down” approach. Under this approach, central governing bodies within the science system (e.g., the American Psychological Association, the National Science Foundation, the German Psychological Society [DGPs] and its divisions) would officially declare that certain system changes will now take place and are binding for everyone. For example, it may be declared that grant proposals must include planned replication attempts, or that the data and materials from preliminary studies on which a grant proposal is based must be openly accessible and routinely checked. Such changes will be made more likely by (a) proactively and consistently communicating their necessity to the public, and (b) figuring out and describing the path forward in as much detail as possible. The less additional effort institutions have to invest into determining what needs to be done, the likelier they will be to take action. Also, the legitimacy of such decisions is certainly strengthened if they are mandated (e.g., through voting) by a large proportion of the respective membership base. This is common practice in many fields of science (Zachar & Kendler, 2012).

The work of (hiring, award or promotion) *committees* is located at a lower level within the hierarchy of science institutions. In his comment, Schmitt argued that committee members need to receive better training and be better rewarded for their work. Otherwise, for pragmatic reasons, they would continue to rely mainly on simple-to-use, quantitative productivity metrics, instead of giving appropriate weight to research quality in their decisions. We agree that better rewards for committee work would be highly desirable, but implementing this change would require the respective political will on the side of institutional leadership. Given that the decisions made by committees may affect an institution's ranking results, and such rankings are largely based on quantitative indicators too, the ways in which academic *institutions* are evaluated must also change. Fortunately, a reward point system like the one we proposed in our Ten Steps paper can also be implemented at this level.

The second route to promoting change could be called the "bottom-up" or "grassroots" approach. Here is what *individual researchers* already do, or may do, to help accelerate the shift toward a more credible science: First, they actually implement Good Science practices. By doing so, they do not only strengthen their own scientific contributions but also help raise the standards in their field for everyone else. Second, as reviewers they devote their time and energy to journals that explicitly embrace high methodological standards, and they are not afraid to reject research lacking in methodological rigor or transparency, even if the same research may have been considered acceptable not too long ago. Third, they volunteer to serve on recruiting committees and, once they are in, promote the use of Good Science criteria in evaluating applicants (see Sassenberg et al., 2020, for recommendations).

Diversity

Several commentators highlighted current lacks of diversity in the field, and a corresponding need to increase diversity, while others focused on the risk of losing healthy diversity that already exists (e.g., in the course of consensus-building) and a corresponding need to preserve that diversity (Fernandes & Aharoni; Friedman; Galang & Morales;

Klimstra; McLean & Syed). This is a very broad topic concerning various aspects of research, which we will address separately below.

Diversity of Participants and Stimuli

Most personality research to date uses highly selective samples of participants and stimuli, casting serious doubts on the validity and generalizability of many conclusions (Henrich, Heine & Norenzayan, 2010). There is no question that it would be desirable to use more representative samples of participants and stimuli. Step 9 in our Ten Steps paper covered this aspect, and we believe we are on the same page as most commentators in this regard. Note that aiming for consensus does not necessarily imply endorsement of a universalistic theory (i.e., assuming that all humans have basically the same psychological properties). To the contrary, a consensus may explicitly be limited in scope regarding culture, race, or other relevant dimensions of diversity, or it may include an explicit account of existing group differences along these dimensions (Syed, 2021).

Lesko and Miller's proposal to pay research participants more adequately for their valuable contributions to our research might indeed be a promising path toward that goal: Paying participants adequately is an ethical imperative in itself. But more representative data might be a welcome side-effect when members of under-represented communities are better incentivized for getting involved with research. Notably, if we conducted fewer poorly designed studies and focused instead on those that actually have a chance of yielding robust new knowledge (again: prioritizing quality over quantity), we might have the resources to pay the average research participant more appropriately.

Diversity of Theories

To some extent, a diversity of theories is a good thing, because the current state of knowledge and theory development may only be *improved* if alternatives are considered. Moreover, if some consensus does emerge, it is likely to be of higher quality – and will certainly have greater legitimacy – if a broad range of views was initially considered (Oreskes, 2019). However, theoretical diversity is not a goal of science in itself: A field of science does not become "better" just by handling more theories. For example, alternative theories may (a) persist despite having long been disproven, (b) have *lower* explanatory power when compared to the theoretical mainstream, or (c) be partly or completely redundant with one another, or with current mainstream theories. Thus, when arguing in favor of explaining some phenomenon with a different theory, it is necessary to articulate what exactly the improvement in doing so would be, in terms of precision, scope, and/or parsimony. This will become much easier when the challenged theory has been properly formalized.

Diversity of Methodology

Several commentators argued in favor of *methodological* diversity and expressed concern that our proposed reward scheme would unfairly disadvantage (i.e., not reward) researchers pursuing other, "non-quantitative" types of research (Dunlop; Hagemann; Klimstra; McLean & Syed). We agree that our reward scheme focuses on the typical, "mainstream" empirical study in which data is collected from a group of participants and then analyzed using quantitative statistical methods. This is because (1) it is the most common approach in our field, (2) it is the approach that we ourselves are most familiar with, and (3) current debates over how research quality may be improved have overwhelmingly focused on this type of research. This, however, does not imply that other types of research are less useful or important. To express this sentiment, the revised version of our reward scheme is now preceded by a "disclaimer" stating that it is most appropriately applied to the type of quantitative study that currently constitutes the mainstream approach in personality research.

In our own view, good qualitative research is primarily about *concept formation*. It focuses on the concepts needed to capture the complexity of some phenomenon of interest. It asks how these concepts relate to each other, and to other concepts. This type of scientific activity is undoubtedly essential for scientific progress, because "with no ideas to verify, there can be no research" (Hogan et al.), and it has been noted many times before that psychology has deficits in terms of specifying its theories (e.g., Smaldino, 2019; Oberauer & Lewandowsky, 2019; Glöckner & Betsch, 2011; Glöckner et al., 2018). Along the same lines, Beck et al. also argue in favor of better training in theory building for students. We fully agree.

Unfortunately, qualitative research has taken sort of a backseat in our field in the past few decades, so many of us do not know enough about it, including its quality criteria. Therefore, we propose that those of our colleagues who have expertise in this area start working with the editorial team of this journal and prepare a special issue addressing just that: How does qualitative personality research work? What distinguishes good from bad research of this type (e.g., how important is generalizability, and how is it checked)? How does qualitative personality research relate to the more mainstream approach that our Ten Steps paper mainly talks about? How do we know whether a qualitative research project has yielded some robust scientific insight – one that may possibly even become the subject of a consensus statement? And what are the success stories in this field – what insights were gained that may rightfully be attributed to the use of this specific methodology?

Diversity of Measures

We argued in our Ten Steps paper that personality researchers should start developing and using standard measures for key traits of interest. Note that we advocated the *inclusion*, not the *exclusive use*, of such measures. This means that, for each construct to be measured in a study, a standard instrument should be included, potentially alongside other measures for the same construct. However, in studies using multiple measures for the same construct, a pre-registration should define the primary outcome up front, to avoid cherry-picking later on. Ideally, if one has good scientific reasons for *not* using a standard measure, these should be explicated so others may consider them as well. Mazei, Mertes, Torka and Hüffmeier, as well as Horstmann and Ziegler list several such reasons in their comments.

While writing this rejoinder, we conferred with several colleagues interested in this matter (Ziegler, Horstmann, Mottus, Rauthmann), and we all seemed to agree that it is high time for such a set of public domain standard measures to be developed. What is needed now is a group of people who accept the responsibility for this development. To promote this endeavour, we suggest that this journal issues a call to action. This work would not have to start from scratch because large sets of public domain items are already available (e.g., in the International Personality Item Pool, https://ipip.ori.org/; or at the Leibniz Institute for Psychology ZPID; https://www.testarchiv.eu/de/test/9006151). This means that the effort would require coordinating with several like-minded initiatives (e.g., Condon et al., 2020), to avoid redundancies or unnecessary competition, and instead maximize synergy.

Diversity Among Researchers

There can be no doubt that the group of people who do personality research is sorely lacking in diversity and representativeness as compared to the world's overall population (Adler; Galang & Morales; Klimstra). In a fair global society, one's national, cultural, ethnic and sexual identity would not matter in terms of who gets to do what job. Obviously, the current situation is a different one, which has to change. It has also been argued that "diversity serves epistemic goals" (Oreskes, 2019, p.59) because (e.g.) overly homogenous groups of researchers may be unaware of certain sub-cultural biases that influence the ways in which they plan their studies and interpret their data, which may ultimately harm the validity of their conclusions. We fully agree with the goal of improving fairness and representativeness in this regard, and we believe that both individual scientists and scientific institutions do have a role to play in getting there. Moreover, the fundamental problem of unequal access to high-quality education (which we consider a necessary requirement for a scientific career) will also have to be addressed at a broader, superordinate level (i.e., political government).

Reliability of the Scoring System

Beck et al. argued that it might be difficult to apply our proposed reward scheme in a manner that yields a reliable outcome, based on their own rating of Bem's (2011) paper. We recently conducted a pilot study in which three (out of eight) different raters (students with little to no research experience of their own) judged each of 37 published psychology papers by means of our proposed reward scheme (Version 1). Before engaging with this task, the raters received some limited training in which they rated a handful of other papers and then discussed their ratings with us.

A table with results from this pilot study can be found in the Online Supplement (https://osf.io/mbgq3/). The reliability of ratings for the ten consensus criteria (numbered 1a to 5b in Version 1) was mostly low to moderate. This was at least partly due to (a) few papers actually presenting consensus of any kind, (b) even fewer doing so explicitly (e.g., by using the word "consensus"), and (c) yet even fewer *building* on some previously documented consensus (see the quartiles in the table, indicating heavily skewed distributions). In fact, in preparing the paper sample for the study, it had been difficult to find any such papers. Obviously, consensus remains quite an exotic topic in psychology to this day. However, the reliability of ratings on our criteria pertaining to the credibility of empirical studies (numbered 6a to 10e in Version 1) was mostly good to excellent: The overall credibility score (mean of Criteria 6a to 10e) achieved an ICC (1,1) of .81 (individual criteria: mean ICC = .58, min = .23, max = .74) and an ICC (1,3) of .93 (individual criteria: mean ICC = .79, min = .47, max = .89).

The Real-World Relevance of Personality Research

Adler as well as Hogan et al. highlighted the sometimes neglected real-world relevance ("application") of personality research in their comments. We could not agree more: Of course, personality research should matter, in terms of being able to predict important outcomes like the ones listed by Hogan et al. However, addressing research questions that "actually matter" (Adler) and pursuing knowledge to "improve the lives of our fellow humans" (Lesko & Miller) will only be worthwhile to the extent that our studies are able to actually yield robust insights. At present, this is too rarely the case.

We personality psychologists could certainly have been more vocal regarding pressing contemporary issues like the re-emergence of authoritarian leadership around the world, or the many well-documented and dramatic cases of unethical behaviors by members of powerful organizations. One of the factors contributing to our relative silence on these matters may have been our doubts regarding the robustness of our own knowledge base, and, accordingly, doubts regarding our ability to actually give well-founded advice. We think this calls for higher ambitions in terms of methodological rigor, topical breadth, and probably timeliness. In order to *strengthen* the real-world relevance of personality research, we encourage our colleagues in the field to write a type 1a consensus paper (possibly for this journal), listing the most pressing contemporary issues and unresolved research questions from a personality standpoint. The 17 Sustainable Development Goals (SDGs; https://sdgs.un.org/goals) proposed by the United Nations might be a helpful starting point for this. In our view, individual differences are indeed highly relevant to many or even most of them.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: Daniel Leising and Felix Schönbrodt are both methodological consultant members of the journal. Isabel Thielmann is a member of the editorial board of the journal.

Author Contributions: *Daniel Leising*—Idea, conceptualization | Writing | Feedback, revisions. *Isabel Thielmann*—Idea, conceptualization | Writing | Feedback, revisions. *Andreas Glöckner*—Idea, conceptualization | Writing | Feedback, revisions. *Anne Gärtner*—Idea, conceptualization | Writing | Feedback, revisions. *Felix Schönbrodt*—Idea, conceptualization | Writing | Feedback, revisions.

Supplementary Materials

For this article the following Supplementary Materials are available via OSF (for access see Index of Supplementary Materials below):

- Data (ratings of research papers)
- Codebook for this data (i.e., meanings of variables)
- R script producing ICCs and quartiles
- Output: ICCs and quartiles
- The original and the revised version of our rating scheme

Index of Supplementary Materials

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2021). *Supplemental materials for: Ten steps toward a better personality science - how quality may be rewarded more in research evaluation* [Data, codebook, script, additional materials]. OSF. https://osf.io/mbgq3

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A. L. M., Bastiaansen, J. A., Benjamin, D., Boehm, U., Botvinik-Nezer, R., Bringmann, L. F., Busch, N. A., Caruyer, E., Cataldo, A. M., Cowan, N., Delios, A., van Dongen, N. N. N., Donkin, C., van

Doorn, J. B., . . .Wagenmakers, E. J. (2021). Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife, 10*, Article e72185. https://doi.org/10.7554/eLife.72185

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

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology, 58*(3), 381–396. https://doi.org/10.1037/0022-3514.58.3.381

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Ziegler, M., & Zimmermann, J. (2020). Bottom up construction of a personality taxonomy. *European Journal of Psychological Assessment, 36*(6), 923–934. https://doi.org/10.1027/1015-5759/a000626

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making, 6*(8), 711–721.

Glöckner, A., Fiedler, S., & Renkewitz, F. (2018). Belastbare und effiziente Wissenschaft: Strategische Ausrichtung von Forschungsprozessen als Weg aus der Replikationskrise [Sound and efficient science: a strategic alignment of research processes as way out of the replication crisis]. *Psychologische Rundschau, 69*(1), 22–36. https://doi.org/10.1026/0033-3042/a000384

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of Prospect Theory. *Cognition, 123*(1), 21–32. https://doi.org/10.1016/j.cognition.2011.12.002

Hagger, M. S., Luszczynska, A., de Wit, J., Benyamini, Y., Burkert, S., Chamberland, P. E., Chater, A., Dombrowski, S. U., van Dongen, A., French, D. P., Gauchet, A., Hankonen, N., Karekla, M., Kinney, A. Y., Kwasnicka, D., Hing Lo, S., López-Roig, S., Meslot, C., Marques, M. M., . . .Gollwitzer, P. M. (2016). Implementation intention and planning interventions in health psychology: Recommendations from the Synergy Expert Group for research and practice. *Psychology & Health, 31*(7), 814–839. https://doi.org/10.1080/08870446.2016.1146719

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., & Lakhani, K. R. (2021). Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Management Science.* Advance online publication. https://doi.org/10.1287/mnsc.2021.4107

Letzring, T. D., Murphy, N. A., Allik, J., Beer, A., Zimmermann, J., & Leising, D. (2021). The judgment of personality: An overview of current empirical research findings. *Personality Science, 2*, Article e6043. https://doi.org/10.5964/ps.6043

Marewski, J. N., Bröder, A., & Glöckner, A. (2018). Some metatheoretical reflections on adaptive decision making and the strategy selection problem. *Journal of Behavioral Decision Making, 31*(2), 181–198. https://doi.org/10.1002/bdm.2075

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

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

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

Sassenberg, K., Glöckner, A., Gollwitzer, M., Kaiser, F., & Lange, J. (2020). Stellungnahme der Fachgruppe Sozialpsychologie zur Qualität, Replizierbarkeit und Transparenz sozialpsychologischer Forschung. https://www.dgps.de/fachgruppen/sozialpsychologie/aktivitaeten/aktuelle-mitteilungen/news-details/stellungnahme-zur-qualitaetssicherung-in-der-forschung/

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development,* Article e2295. Advance online publication. https://doi.org/10.31234/osf.io/8w2sd

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature, 575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Syed, M. (2021, December 3). *Reproducibility, diversity, and the crisis of inference in psychology.* PsyArXiv. https://doi.org/10.31234/osf.io/89buj

Thielmann, I., Moshagen, M., Hilbig, B. E., & Zettler, I. (2021). On the comparability of basic personality models: Meta-analytic correspondence, scope, and orthogonality of the Big Five and HEXACO dimensions. *European Journal of Personality.* Advance online publication. https://doi.org/10.1177/08902070211026793

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S. H., Franchin, L., Frank, M. C., Geraci, A., Hamlin, K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., . . .Zettersten, M. (2021). *Improving the generalizability of infant psychological research: The ManyBabies model* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8vwbf

Zachar, P., & Kendler, K. S. (2012). The removal of Pluto from the class of planets and homosexuality from the class of psychiatric disorders: A comparison. *Philosophy, Ethics, and Humanities in Medicine; PEHM, 7,* Article 4. https://doi.org/10.1186/1747-5341-7-4

Zachar, P., Krueger, R. F., & Kendler, K. S. (2016). Personality disorder in DSM-5: An oral history. *Psychological Medicine, 46*(1), 1–10. https://doi.org/10.1017/S0033291715001543

eapp

Personality Science (PS) is an official journal of the European Association of Personality Psychology (EAPP).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.