PERSONALITY Science

PsychOpen GOLD

Theory [Invited Paper]

Check for updates

Ten Steps Toward a Better Personality Science – How Quality May Be Rewarded More in Research Evaluation

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 e6029, https://doi.org/10.5964/ps.6029

Received: 2021-01-30 • **Accepted:** 2021-07-04 • **Published (VoR):** 2022-05-06

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

Reviewing: *Round 1* - Rainer Bromme; Dirk Hagemann; Manfred Schmitt. Open reviews are available. [see Index of Supplementary Materials]

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).

Abstract

This target article is part of a theme bundle including open peer commentaries (https://doi.org/10.5964/ps.9227) and a rejoinder by the authors (https://doi.org/10.5964/ps.7961). We point out ten steps that we think will go a long way in improving personality science. The first five steps focus on fostering consensus regarding (1) research goals, (2) terminology, (3) measurement practices, (4) data handling, and (5) the current state of theory and evidence. The other five steps focus on improving the credibility of empirical research, through (6) formal modelling, (7) mandatory pre-registration for confirmatory claims, (8) replication as a routine practice, (9) planning for informative studies (e.g., in terms of statistical power), and (10) making data, analysis scripts, and materials openly available. The current, quantity-based incentive structure in academia clearly stands in the way of implementing many of these practices, resulting in a research literature with sometimes questionable utility and/or integrity. As a solution, we propose a more quality-based reward scheme that explicitly weights published research by its Good Science merits. Scientists need to be increasingly rewarded for doing good work, not just lots of work.

Keywords

reproducibility, credibility, research evaluation, assessment, quality, metrics, publication

Relevance Statement

There have been many calls and concrete suggestions for establishing higher scientific standards in psychology. However, better scientific work does require significantly more time, energy and resources, and these additional investments are not appropriately rewarded under the current incentive structure in academia. This may explain why "Open Science" – although widely accepted in principle – is still very much lacking in terms of implementation. To overcome this problem, we recommend the use of a metric that explicitly rewards the use of good scientific practices such as replication, pre-registration, and model formalization in published research. The goal is to stop rewarding researcher behavior that mainly just maximizes one's research output in terms of quantity, and to start rewarding behavior that is likely to actually contribute something solid to the scientific knowledge base.

Key Insights

- Personality psychology needs to improve in terms of scientific rigor
- The current incentive structure in academia impedes such improvement
- We suggest explicitly rewarding scientists for conducting more rigorous research
- We also suggest rewarding them for engaging in consensus-building

The purpose of this paper is to present a number of concrete steps that personality researchers (as well as researchers in other areas of psychology) may take to significantly improve the scientific standards in their field. The selection represents the consensus that was achieved in our group of authors, which was tasked with compiling such recommendations by the Personality Psychology and Psychological Diagnostics section of the German Psychological Society (Deutsche Gesellschaft für Psychologie, DGPs). To foster the implementation of these practices, we propose incentivizing their adoption by weighting an author's publications with quality indicators in performance evaluations.

What is Good Research?

As our goal is to make personality research "better", it is necessary to first outline what we consider "good" research. Good research works toward understanding and predicting important phenomena. To be able to achieve that goal, its theories need to be falsifiable, and its findings need to be reproducible. Notably, *cumulative progress* in science means that, over time, the underlying theoretical ideas should become more valid (i.e., more in line with true relations in the real world), more precise (i.e., "degree of precision", Popper, 1934/2002, p. 105), more comprehensive (i.e., a model accounts for more phenomena; "level of universality", Popper, 1934/2002, p. 105), and/or more parsimonious (i.e., the same phenomena are accounted for by simpler models with fewer assumptions; “Occam’s razor”, see also Popper, 1934/2002, p. 121). This image of the scientific process has been repeatedly likened to biological evolution (Marcum, 2017).

Given the complexity of most of the phenomena that we are trying to understand, this almost always requires intense collaboration among researchers (for further reasons, see below). And because our human capacity to thoroughly understand this complexity is quite limited, there is still a high likelihood of mistakes and dead ends in scientific research. There is no such thing as smooth sailing here. Theories can never be established as true (Popper, 1934/2002). Only a small proportion of our theoretical ideas will turn out to be useful, sometimes (Box & Luceno, 1997). Hence, scientists who are actually interested in learning something new must constantly subject their own ideas to scrutiny, testing how well they fare in explaining and predicting reality. Mismatches between predictions and reality point to the necessity for improvement (in terms of, e.g., theory and/or measurement). Therefore, good quality of research is not evidenced by a shiny facade, but rather by the researchers’ willingness to look its shortcomings - and sometimes blatant failures - in the eye. Shortcomings and failures constitute an opportunity for everyone to learn something, and thus need to be laid open, rather than hidden. Only if shortcomings are laid open, they can lead to new or improved theories (Popper, 1934/2002) or changes in auxiliary assumptions (Lakatos, 1970). Similarly, changes in paradigms (“revolutions”; Kuhn, 1962), or changes in taxonomies (Kuhn, 1990), are usually spurred by an accumulation of evidence (“anomalies”) that cannot be accounted for anymore within the bounds of currently existing ones.

Not all fields of science are collectively learning at the same speed, however. Platt (1964) pointed out that the progress made in various branches of science differs by an order of magnitude, and suggested that the more successful fields apply what he calls "strong inference research": This implies accepting that theories are always wrong or at least preliminary, and should be improved in a stepwise, joint effort. We share that view. Theories or hypotheses should be specified so precisely that they can be falsified, or tested against each other, in decisive empirical studies. Over time, researchers may reach increasing consensus regarding results and their interpretation. The whole procedure is applied in an iterative fashion, always comparing the most promising remaining theories and hypotheses at present with one another.

There have been important updates of a naive falsificationist approach (e.g., Lakatos, 1970; Meehl, 1990a), and fundamentally different views have been articulated, as well (e.g., Kuhn, 1962, 1990). For brevity, we will not cover the full history and discussions in detail here (see Oreskes, 2020, Chapter 1, for a review). Current approaches (e.g., Oreskes, 2020; see also Kuhn, 1962) particularly highlight that scientific belief systems cannot be objectively ranked concerning their fit with empirical evidence or their rationality. Instead, it is assumed that "objectivity" is established as part of a social process. Hence, scientific facts are claims about which agreement has been reached among scientists in the respective field (*consensus*; Oreskes, 2020, p. 127). Systematic consensus analyses of peer reviewed articles are suggested as methods to detect consensus (Oreskes, 2020, p. 130). Oreskes argues that diversity concerning demographics, beliefs, and values is important to balance out bias in this social process of consensus construction. Furthermore, Oreskes argues that multiple methods and even imperfect data should also be used instead of "methodological fetishism" that relies on one particular method only (p. 134) (see also Feyerabend, 1993). Finally, it has been questioned whether single decisive experimental studies (critical experiments; see Platt, 1964) can be conducted in practice since (in many cases) not only a theory but further auxiliary theories/assumptions are being tested at the same time (under-determination, Duhem-Quine thesis, see Oreskes, 2020, p. 32, p. 37).

Many psychologists try to apply variants of a methodological approach that is to enable critical theory testing in line with Platt (1964), or at least a probabilistic variant of it that aims to strengthen or weaken our probabilistic belief as to whether a theory is true or false. Too often, however, the use of said techniques is not yet strict enough, and/or basic pre-conditions for using them (e.g., sufficient specification of theories and constructs) are not met, casting considerable doubt on the validity of many conclusions. A recent analysis of papers published in Psychological Science between 2009 and 2019, for example, showed that only 15.33% of the authors explicitly claimed to test predictions from theories and that many of the theories were only mentioned in one publication (McPhetres et al., 2021). The authors conclude: "We interpret this to suggest that the majority of research published in this flagship journal is not driven by theory, nor can it be contributing to cumulative theory building." (McPhetres et al., 2021, Abstract)

In the current paper, we argue in favor of a much stricter adherence to some core principles of scientific conduct, and we provide some suggestions as to how that adherence may be fostered. For example, we suggest that researchers should engage more actively in structured social processes with the aim of building consensus among them (cf. Oreskes, 2020). This may concern various aspects of the scientific process (e.g., terminology, measurement etc., see below). We also argue in favor of greater theory specification. Taking measures such as these should help solve some of the major problems that currently stand in the way of doing *efficient* psychological science, as outlined by Platt (1964) (cf. Oreskes, 2020). Furthermore, we argue that researchers need to be explicitly rewarded for engaging in such Good Science practices, and we suggest a concrete reward scheme for doing so.

What Aren't Good Indicators of Research Quality?

Good science may result in many publications that get cited by many researchers, and attract large sums of research funding. This, however, does not mean that researchers who are publishing a lot, get cited a lot, and acquire large grants are necessarily doing Good Science. In fact, it is well-documented that such purely quantitative measures of research productivity do reflect all sorts of influences apart from research quality (see below). Furthermore, they are very easily manipulated (Chapman et al., 2019; Kwok, 2005). Both theoretical work (e.g., Dawes, 1980; Engel, 2015) and empirical evidence (Van Noorden, 2013; Young et al., 2008) suggest that the use of such quantity-based evaluation criteria tends to encourage behavior that may ultimately compromise the credibility of research altogether (Chapman et al., 2019).

Therefore, given the parameters of the current incentive structure in academia, we challenge the notion that scientific merit is appropriately captured by how *much* a researcher publishes, how *often* a researcher is being cited, or how *much* grant money a researcher acquires. The extent to which such numbers are contaminated with influences independent of scientific merit, and sometimes even detrimental to scientific merit, is simply too high to keep the current evaluation practices going. We thus call upon our colleagues to join us in an attempt to visibly improve these practices by switching to a more quality-based reward system. Further below in the present paper, we outline such a system in detail. What should count is *what* you publish. In order for that to be achieved, the current incentive structure needs to change. We agree with many of the suggestions made as part of previous, like-minded initiatives (e.g., Asendorpf et al., 2013a, 2013b; Back, 2020; Chambers, 2017; Dougherty et al., 2019; Henrich et al., 2010; Lindsay, 2020; Nosek & Bar-Anan, 2012; Nosek et al., 2012; https://sfdora.org/), but aim to go a step further by proposing a specific, actionable scheme of rewards for actually taking steps toward a better (personality) science.

Ten Steps Toward Improving the Scientific Standards in Personality Research

In the following, we present a number of concrete steps that may be taken - by individual researchers, by groups of researchers, or as a general policy in the field as whole - to improve the scientific standards in personality research. Several of these steps have already been proposed and advocated by other researchers (see above), so we only embrace them here. Others have not been proposed before. Each step by itself already has the potential to contribute significantly to an improvement of scientific standards in personality research, which is our ultimate goal. The more of these steps we manage to implement into our everyday research practices, the better.

However, it needs to be acknowledged that most of these quality improvements come at a price: Almost always that price is a significant amount of extra time and energy that has to be invested. And sometimes implementing a step will actually require the allocation of additional *financial* resources. There is just no way around it: Better research tends to be significantly more difficult, expensive, and time-consuming. The most convenient way of dealing with this fact would be to settle for less, to aim lower, to content oneself with a situation in which much of the published literature is of very limited scientific value, and then try to do likewise and enlarge that kind of literature even more. The alternative, and this is what we advocate here, is to wholeheartedly *embrace* the goal of raising scientific standards in the field, and to explicitly *reward* any attempt at doing so.

Our specific recommendations are organized along two broad themes: The first theme is the promotion of greater *consensus-building* among researchers working in the field. The first five of the concrete steps that we outline below may be subsumed under this theme. The second theme is the improvement of the *credibility of empirical research.* The remaining five steps may be subsumed under this theme.

Consensus Building (Steps 1 to 5)

Generally speaking, and in no way specific to our field in particular, personality psychology would benefit considerably from closer collaboration among individual researchers (Forscher et al., 2020). But closer collaboration requires the development of a common language, common standards, and common practices. The first five of the individual steps that we propose here all align directly with that broad theme. Specifically, we argue 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. We argue that researchers need to be rewarded both for *contributing to the emergence* of such consensus and for *basing their own research* on such consensus. Needless to say, a consensual viewpoint first has to emerge before new research can be based on it, or even challenge it. Therefore, the development of - ever improving - consensus has to be declared an important research goal of its own (cf. Oreskes, 2020), and the current version of any consensus that was achieved has to be explicated. At present, psychology is often lacking in both of these regards. One of the reasons may be that consensus work tends to be very demanding and tiresome, while not being rewarded much. Thus, this type of work needs to be rewarded *a lot* more, in order to actually get people to take it upon them. We will now address our individual suggestions regarding consensus building in more detail.

Step 1: Common Research Goals

In our view, personality psychology would benefit from defining a shared research agenda outlining some of the pressing issues that the field is currently attempting to understand. In many other fields of science, this has been common practice for decades (e.g., Adolphs, 2015; Millennium Problems in Math; Sustainable Development Goals). In Biomedical and Neuroscience, the Human Connectome Project (HCP) and the UK Biobank are examples of successful collaborative and interdisciplinary endeavours that address explicit, common research goals, are properly funded and produce valuable empirical output.

Of course, a shared research agenda should not preclude off-mainstream research. But it would serve at least two purposes: First, it may not only strengthen the collaborative *spirit* in the field, but it may also directly lead to more concrete collaborative *action* (e.g., multi-center data collection efforts; Psychological Science Accelerator: https://psysciacc.org/). Questions that concern the strongest opposing theoretical or empirical views in the field naturally qualify most easily for such a list of important issues. In fact, important points on the agenda may result from disagreement that is only detected in the process of working toward consensus (e.g., Oreskes, 2020, p. 130).

Second, formulating a common research agenda more explicitly will make the direction in which the field is headed more visible to the public. This definitely involves a risk of being criticized for investigating seemingly irrelevant matters, or for not investigating matters considered to be more relevant. The general freedom of research is not to be undermined, of course. But a more lively and continuous exchange - both within scientific communities and with the broader public - over the work that we as scientists are doing (and why we are doing it) would certainly be beneficial, and may actually help retain or restore some trust in science (Oreskes, 2020, p. 57).

Step 2: Common Terminology

The language that scientists use to address the phenomena that interest them should be superior to the everyday language in terms of precision. In fact, that is the whole point of scientists *having* such a language of their own (Leising & Borgstede, 2020). Ideally, that precision should be perfect (i.e. bijective), meaning that a given term is used for exactly *one* thing, and *no other term* is used for that thing. For example, physicists would not find it acceptable to use a term like “second” for an exactly specified amount of time *and for other, somehow “similar” amounts of time, or mass*. Physicists would also not find it acceptable to label that same amount of time “second” *or “minute” or “kilogram”.*

Unfortunately, personality psychology leaves much to be desired in this regard, as already lamented many years ago by Block (1995). Psychologists have often used the same term to denote different things (e.g., “narcissism”: Ackerman et al., 2019; Cain et al., 2008; “agreeableness”: Thielmann et al., in press; “autonomy”: Hmel & Pincus, 2002), and/or different terms to denote the same or highly overlapping things (e.g., Leising et al., 2013, 2016; Moshagen et al., 2018). The former has been called the jingle-fallacy, whereas the latter has been called the jangle-fallacy (Block, 1995). And the situation has *not* improved much in the course of the past few decades. But change is still possible. What the field desperately needs, in our view, is a concerted effort at streamlining its use of language. There are already a few examples of building domain ontologies in psychology that serve the purpose of formally naming and defining concepts and their potential interrelations (e.g., Gray, 2017; Poldrack & Yarkoni, 2016; Spadaro et al., 2020; West et al., 2019; see also classic work on nomological nets, Cronbach & Meehl, 1955, and lexical taxonomies, Kuhn, 1990).

However, personality psychology does have an inherent problem in this regard, which the “harder” sciences (e.g., physics) do not have to the same extent: For many or even most of the phenomena that we as personality psychologists are interested in, terms already exist in the everyday language (e.g., “aggression”, “intelligence”). There is thus a risk of increasing *confusion* rather than clarity when such terms, which already have a semantic web of their own, are also used to denote a scientific construct (Leising & Borgstede, 2020). That is because the new, scientific use of the term will almost certainly diverge from how the same term is used by non-scientists (and the latter is almost never precisely known). Also, person descriptors taken from the everyday language tend to be highly *evaluative* (Anderson, 1968; Dumas et al., 2002; Leising et al., 2012). By using evaluation-laden terms for denoting their constructs, personality psychologists may convey the message that certain levels on such a dimension are intrinsically more desirable than others, which does not square well with the general goal of scientific objectivity. In terms of how research findings are processed and interpreted (both by laypeople and by scientists), it probably does make a difference whether, for example, the same personality trait is called “Neuroticism” or “Sensitivity”.

A radical solution to this problem would be to abandon natural language labels for psychological constructs altogether, and to replace them with abstract ones (e.g., Cattell & Nesselroade, 1967). Researchers may then still investigate how these abstract dimensions (e.g., “dimension 4”) *relate* to everyday language descriptions of the same target persons, while taking care to avoid the impression that these two things are interchangeable. This solution has the potential drawback that artificial personality dimensions may be harder to grasp for laypeople, which may complicate communication with the public.

If one continues to use everyday language to denote the inter-individual differences we as personality researchers are interested in, there is a significant risk of creating conflicting semantic (nomological) webs for the same sets of terms. The worst-case scenario in this regard would be that the same terms are used in certain, fuzzy ways by laypeople, and in *other fuzzy* ways by scientists. If this were the case - and it may very well be the case in contemporary personality research - it would be better to abandon the “scientific” version altogether. The remaining alternative is to keep using everyday language terms but to *standardize their use* as much as possible. There are numerous examples for such normative processes in other branches of science (e.g., astronomers re-defined what is and what is not to be called a “planet”; Evans, 2008). If this approach is pursued, the terms that *are* to be used should still be as evaluatively neutral as possible, for the reason outlined above.

It cannot be denied that personality research still has to come up with a terminology that is 1) unambiguous, 2) efficient (e.g., minimizing redundancy), 3) consensually adopted, and thus ultimately (4) more *useful* than the language that laypeople use for describing many of the same phenomena. And this superior utility would have to be demonstrated (e.g., in terms of predictive validity), rather than just be assumed. The use of terms does not become more scientific just because it is a scientist who engages in it.

Step 3: Common Measurement Practices

Psychologists often take considerable freedoms in how they measure something (Elson, 2016, 2017, 2019; Elson et al., 2014; Flake & Fried, 2020): Sometimes, they use *variants* of an existing measurement procedure (e.g., only a subset of stimuli, or different presentation times or presentation orders). Sometimes, they use completely different procedures for measuring (allegedly) the same thing (e.g., dozens or even hundreds of different ways of measuring depression or emotions; Flake & Fried, 2020; Santor et al., 2006). And sometimes the content domains captured by different measures are so strongly overlapping that treating them as being separate becomes fairly dubious (e.g., the overlap between depression, self-esteem, dispositional optimism, and neuroticism; Leising et al., 2016).

This ongoing habit is a major obstacle for cumulative knowledge building (Flake & Fried, 2020); for example, meta-analyses are difficult or even futile if measures (which superficially share the same label) do not sufficiently converge (Steel et al., 2008). Obviously, this issue overlaps with the previous one - lack of a binding terminology.

We argue that personality psychologists must find more common ground as to what should be measured how (i.e., concerning standard operationalizations of constructs). The most prominent example that comes to mind concerns the assessment of personality itself. Two scales do not become interchangeable in terms of content just because they are given the same name (the *jingle*-fallacy; see above). And neither do two scales become non-interchangeable just because they are given different names (the *jangle*-fallacy; see above) (Flake & Fried, 2020; Block, 1995). Rather, the measurement (non-)equivalence of different scales has to be demonstrated empirically (Thielmann & Hilbig, 2019). Perhaps the most elegant way of avoiding this whole problem would be to simply *not use different measures of (allegedly) the same thing anymore*. In fact, we do argue that 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. Moreover, we argue that such consensus measures *have to be in the public domain*.

Needless to say, the identification of a consensus measure should be based on the best available evidence to date, and is always to be considered privisionary. Also note that such consensus-building is not meant to impede research aiming to improve our measurement practices themselves. Such research is and will remain necessary, but it clearly has a different objective and thus needs to be set apart from research that just *uses* the best measures that are currently available. Not separating these two strands of research from one another will only prolong the existing uncertainty over how comparable the outcomes of different studies really are.

There may be longer or shorter versions of the same measure, to accommodate situations in which resources are more or less limited. Shorter versions should always be completely included in longer versions, to ensure "upward compatibility" in analyses, or equivalence between a shorter and a longer version of the same measure has to be demonstrated empirically. Because much of personality research claims to address issues of universal relevance, measurement equivalence across different languages and cultures is also an issue that has to be dealt with. Furthermore, sometimes it might make sense to employ independent (but measurement-equivalent) "secondary measures" of a construct alongside the respective "gold-standard", to reduce the risk of a "mono-method" bias (i.e., results being owed - at least in part - to *how* something is measured, as opposed to *what* is being measured).

In our view, a viable solution would be that researchers working on the same issues *join* each other in creating and thoroughly validating a new consensus measure for their construct of interest and then put it in the public domain, with all of them becoming authors on the one research paper that introduces the respective measure. The alternative would be to mandate a group of experts to come up with recommendations regarding one or several of the measures that already exist, to be included in all studies on the respective subject matter. This has just happened, laudably, in the field of clinical psychology (in regard to the measurement of anxiety and depression; Farber et al., 2020). Of course, such a selection would always remain preliminary, and the process would again have to be as transparent and participatory as possible, to avoid being hijacked by lobbyists. Whichever road is taken, papers describing a consensually adopted measurement practice may be referred to in subsequent papers, in such a way that they basically *replace* parts of the traditional "measures" section.

Step 4: Common Practices in Data (Pre-)Processing and Analysis

Psychology researchers also take considerable freedom in how they *analyze* their data once it has been collected. This may result in quite different outcomes, even if the exact same dataset is analyzed in regard to the exact same research question. For example, an effect may become statistically significant according to one type of analysis, but not significant according to another, and/or effect size estimates may differ from one another. Problems of this kind have been well documented across several different branches of research (Botvinik-Nezer et al., 2020; Rohrer et al., 2017; Silberzahn et al., 2018). And even if two researchers use the exact same statistical procedures, they may still obtain different results depending on how the data were *pre-processed*. Due to the sheer level of complexity in the respective data, problems such as these may be particularly pronounced in analyses of biological data (e.g., EEG, fMRI).

We therefore think that we as researchers need to work toward greater consensus in regard to best-practice approaches to pre-processing and analyzing our data. The ultimate goal is to drastically limit the available number of "researcher degrees of freedom" in this regard, by fixing the analytical pipeline to, ideally, a single path. This may first be done locally by individual labs which define and then publish their “standard lab practices” or “standard operating procedures” as a point of reference for others (Lin & Green, 2016). Later, one may attempt to achieve a more far-reaching, and ideally field-wide consensus. Again, even such a field-wide consensus would still remain preliminary, and thus likely to be later replaced by an ever better one. At present, we are not aware of a single example for such a field-wide consensus in personality psychology, but that does not mean that consensus work is futile, or unnecessary. A powerful tool in this kind of work may be collaborative forking path analyses (Wacker, 2017), in which large groups of researchers compare procedures in a systematic and transparent manner, with the ultimate goal of establishing an ideal, and shared, data-analytic pipeline.

Step 5: Common Views on the Current State of Theory and Knowledge

Large parts of the research literature (not only) in psychology are redundant, because the theory sections of hundreds or thousands of papers basically recount the same things over and over again. What is worse, authors often *selectively* report what the current state of the literature supposedly is, in order to frame the “story” that they would like to tell with their current paper in the most effective way possible, and/or because they are simply not aware of some of the existing research (e.g., due to the file drawer effect or jingle/jangle issues; see above). As a remedy, we propose that groups of authors investigating the same issues work on consensus documents outlining the current state of their field regarding 1) what can count as established knowledge, and 2) what is not yet known but needs to be known, and why. This obviously overlaps somewhat with our recommendation to formulate common research goals (Step 1). The difference is that Step 1 is geared more toward the scientific community’s interface with the broader public, whereas Step 5 is more of a science-internal affair, and thus likely to be much more detailed and technical in nature. The two kinds of consensus are not interchangeable, but should of course be made to harmonize with each other as much as possible. The resulting consensus documents (outlining shared views on theory development and evidence base) may be referred to in subsequent publications and basically replace much or all of the traditional “theory”/“background”/“introduction” sections there.

In our view, pursuing this approach would have several positive effects: First, it would spare individual authors the effort of repeating themselves again and again, and their readers the effort of going through many different versions of basically the same (but sometimes strangely different) story. Second, it would foster the emergence of a more common view of the state of the art in a field, which may help avoid some “reinventions of the wheel”, lead to more active collaboration (e.g., in trying to fill a consensually identified knowledge gap by collecting data together), and help clarify in advance what outcomes of future research should be interpreted as speaking for/against a given theory.

The Intricacies of Consensus Building

There are not only great chances, but also considerable difficulties, and even a few risks, involved in working toward greater consensus. We will briefly address some of the most important intricacies of such processes here: First, it needs to be asked who should be *responsible* for organizing greater consensus. In our view, this burden lies first and foremost with individual (groups of) researchers - they need to invest more effort into finding or developing and then articulating common ground between them. The key here is to actively reach out to *other* groups - especially ones that one has not yet collaborated with - and start such a process. The particular research habits of the members of an existing local research group do *not* qualify as consensus as we understand it here. Also, we think it will help if the goal of working toward consensus is openly declared by the participants, and if a structured process aiming for that goal is outlined up front. Also, precautions must be taken against the risk that certain authors will try to hijack the process to primarily advance their *own* idiosyncratic viewpoints instead of attempting to find common ground with others. One promising approach to preventing this from happening would be to make the whole process as transparent (e.g., by publicly documenting workflow) and as participatory (e.g., by making use of polling) as possible from the get-go. External moderation (e.g., by scientists from other, unrelated fields) may also have a role to play in this regard. Note that these latter features distinguish a consensus process as we envision it here quite clearly from the more traditional approach in which a relatively small number of authoritative figures in a field is tasked by a journal editor with outlining their specific views on the current state of affairs (e.g., as review papers or encyclopedia chapters) or submit such papers on their own initiative. With this traditional approach, the extent to which the viewpoint expressed in a contribution is consensual among researchers working in the field will usually remain largely opaque. Even then proper peer-review should be able to strengthen consensuality to some extent, but we consider it likely that aiming for a broad and explicit consensus by means of a fair, structured process may go even further in terms of scientific utility.

Second, it needs to be asked how consensus can be defined: what can count as evidence that a viewpoint is sufficiently shared among the members of a research community? The first thing that comes to mind here is the number of people working in a field that endorse a consensus, either through their co-authorship on the respective consensus document, or by public declaration (e.g., on a website showcasing all of those endorsements). The most official way in which the consensuality of a viewpoint may be documented is through formal endorsement by some academic society (which, in turn, may justify such an endorsement through polling among its members). However, such an endorsement will most likely be the provisional end-point, rather than the starting point, of a consensus-building process. Ultimately, it will always be the responsibility of the researchers who think that a certain viewpoint is consensus-worthy to demonstrate that worthiness, and to garner the support of as many of their colleagues as possible for that viewpoint.

Third, it needs to be asked how the specific *content* of a consensus, once achieved, is best documented. In this regard, we argue in favor of using the *journal article* as the primary method of articulating consensus. This would make it possible to earn the most "valuable" type of citations, and thus make it more attractive for potential authors to actually invest the significant effort that is associated with this kind of work. Needless to say, consensus documents have to be published with open access. To support such processes and to minimize the risk that consensus paths may be overlooked, databases assembling all of the relevant consensus documents could be implemented (e.g., *theory databases*; Glöckner et al., 2018). These should provide links to the various versions of a consensus document, and maybe even links to the empirical data that led to the respective updates. These databases should be in the public domain and would ideally be hosted by state funded institutions (e.g., Leibniz Institute for Psychology [ZPID]).

Finally, it needs to be acknowledged that the *value* of any consensus hinges entirely on how well it is justified, in terms of empirical data and conceptual reasoning. If many people agree that something is the case, but there are actually no good reasons for doing so, this is just as bad as if someone actually has very good reasons to believe something is the case, but fails to convince others to join him or her. Based on the feedback that has reached us regarding previous versions of this article, it seems that many of our colleagues in the field seem to fear the emergence of "false" or "bad" consensus (e.g., one that is politically motivated rather than well-supported by the data). We assume that these fears are rooted in some real, negative experiences, and should thus be taken seriously. For example, it *does* matter how well the diagnostic taxa that we use to classify psychiatric cases align with the actual covariation structure of symptoms. The less they do so, the more often patients will find themselves in diagnostic categories, or at some locations on diagnostic dimensions, that are only real in the collective minds of diagnosticians.

So, what can be done to alleviate these fears of "bad consensus"? For one, it needs to be acknowledged that there is an inherent tension between consensus and innovation, and that science is about *both*: Firmly establishing how things should be approached and what is most reasonable to believe, while at the same time encouraging researchers to go further, to explore, and to *question* the current consensus. However, for the latter to become possible, it first needs to be made clear what the current consensus *is* - provided that it even exists. Sometimes, a group of researchers claiming that some viewpoint deserves to be declared consensual may lead another group of researchers to fiercely disagree and to present strong evidence against that claim (and, hopefully, evidence *for* a more sound consensus as well). This would be in the interest of scientific progress, because it would at least entail a *chance* (but no guarantee, of course) to overcome misconceptions and misunderstandings, to subject competing views to critical tests, and thus ultimately to resolve the existing differences. But all of this will not happen unless the first group of researchers goes to work on their version of a consensus document.

Interestingly, the answer to the question of what a “good” consensus is has some overlap with common conceptions of good theorizing: Leaving aside for a moment the question of how many researchers endorse it, we suggest that a consensus is better 1) the *broader* it is in terms of scope (e.g., a given construct is measured in certain ways most of the time, rather than only under certain circumstances), and 2) the more specific it is in terms of actionability (e.g., it prescribes exactly how to go about measuring something, or analyzing some data, rather than just providing some broad principles).

In our view, personality psychology at present certainly does not suffer from *too much* consensus, but from *too little*, and this lack severely impedes cumulative progress. Anyone who has ever attempted to synthesize the findings from several methodologically diverse research studies (e.g., using meta-analysis) may probably attest to this. Much would be gained if more different groups of researchers took it upon themselves to try and figure out what most of them can agree on. This will most likely result in a *smallest common denominator* type of consensus, which in our view is unlikely to become a menace to scientific progress. And a consensus document may definitely contain a section outlining the issues on which the authors do *not* agree with one another (yet) (e.g., Marewski et al., 2018), and ideally, how the dispute might best be resolved (i.e., a roadmap to what has been termed “adversarial collaboration”). Agreeing to disagree this way is conducive to scientific progress as well.

Often, a consensus process may begin at a more “local” level, with relatively small groups of scientists that favor competing ideas first specifying what *each group* agrees on, and only later engaging in the more ambitious task of finding common ground across groups. It also needs to be understood that consensus is always to be considered preliminary. As soon as there is enough evidence supporting a better consensus, the old one needs to be updated or even discarded. To avoid getting stuck with a suboptimal consensus, emerging critical counter-evidence (Platt, 1964) and dissenting opinions based on such evidence have to be valued, and their publication has to be promoted. Versioning (e.g., in an online theory database) will allow researchers to keep track of such developments, and to refer to a specific instantiation of a document, if needed.

Improving the Credibility of Empirical Research (Steps 6 to 10)

Our second set of recommendations concerns individual, empirical research studies. We assemble what, to us, seem to be the most important and the most viable ways in which such studies can be improved in terms of credibility. All of the measures that we recommend here have been proposed before (see above). We present them all in one place, explain once more why they are important, and point to a few critical details regarding their implementation. Notably, none of these measures will by itself *ensure* that a study will yield a valuable contribution to the scientific literature. For example, a dataset stemming from an invalid experiment does not become more valuable just because it is made openly accessible. But, based on our own experience, we are convinced that incorporating as many of these measures as possible into one's research practice will improve the *chances* of making a valuable contribution to science quite a bit. Furthermore, these steps are necessary to enable a proper evaluation of a line of research (see below).

Step 6: Formalization of Theoretical Ideas

We strongly encourage personality researchers to outline their theoretical models in a more formalized (i.e., mathematical or formal-logical) manner, in line with recent suggestions by many other colleagues (e.g., Borsboom et al., 2020, February 29; Glöckner & Betsch, 2011; Oberauer & Lewandowsky, 2019; Smaldino, 2019, 2020; van Rooij & Blokpoel, 2020, July 6; West et al., 2019). This simply means that all input and output variables of a theory are unambiguously measurable (see Step 3), and connected with one another by mathematical or formal-logical (e.g., if - then, AND, OR) operations (Glöckner & Betsch, 2011; Popper, 1934/2002; van Rooij & Blokpoel, 2020, July 6). That way, theories become "objective" such that their predictions cannot be debated or changed post-hoc, after seeing the data. Hence, theory formalization fosters "strong inference research" as suggested by Platt (1964) - in some cases it may even be required to enable any actual testing of a theory to begin with.

Formalization has undeniable advantages over the "narrative" approach that currently still pre-dominates in psychology, which means phrasing one's theories only in natural language terms: First and foremost, mathematical or formal-logical formulations are capable of capturing the complexity of what is going on with an *exactness* that the natural language simply cannot afford. This difference becomes more and more apparent the greater that complexity is. Their greater exactness also makes mathematical and formal-logical models more easily *falsifiable*. In many cases, formalization thus increases the empirical content ("empirischer Gehalt"; Glöckner & Betsch, 2011; Popper, 1934/2002, p. 96) of a theory, or even ensures that a theory has empirical content at all. Furthermore, formal models make it easier to determine whether a theory is free of tautologies and/or contradictions, and their greater exactness is also an asset in terms of cumulative progress. For example, the very same parameter that is explained by one model (explanandum) may serve as a predictor variable (explanans) in another model, thus providing a seamless, unambiguous fit between the two models. When models are joined or integrated this way, stringent formalization will help ensure that all of the dynamics contained in its parts will be preserved, and automatically generate concrete predictions for the overall model (i.e., predict outcome values for all possible values of all predictor variables from *both* of the previously separate models).

Mathematical and formal-logical formulations are also likely to make *redundancy* in the research literature better visible. Regardless of what *names* two researchers give to the individual parameters in their models, if these differently named components are thought to interact with one another in the same way (as obvious from the respective mathematical formulations) in bringing about whatever phenomenon is to be explained, the models are structurally identical. If the input and output variables are also the same - or have measurement equivalence - the models are redundant. *Partial* redundancy (i.e., overlap) between models will also become easier to spot when models have been properly formalized. In fact, we consider it highly likely that substantial proportions of various “different” theories in personality psychology *are* redundant with one another, which remains opaque as long as they are phrased in natural language terms, but may become all the more obvious once they are translated into an algebraic form (e.g., Leising et al., 2015).

Mathematical and formal-logical theory formulations also make *(non-)compatibility* between models more obvious: They do so by helping identify *in what regard exactly* two models explaining the same phenomenon differ from one another, which may then become the basis for experiments in which the different models are directly pitted against each other in the most informative way possible. Finally, formalization may make *gaps* in one’s modelling efforts visible. When trying to formalize even very basic theories, all of one’s previously implicit assumptions (or lack of assumptions) have to be made explicit, which can be a very enlightening, and often humbling process.

Notably, formalization does *not* necessarily imply high levels of detail or complexity. Rather, especially if a research topic is relatively new and unexplored, models may have to remain relatively simple at first. However, even simple ideas may and should be expressed in a formalized fashion, for the reasons outlined above. This will make it easier to determine later on whether subsequent research has actually led to an improvement, in terms of specification. On a side note, formalization is not necessarily tied to a nomothetic approach either: Rather, we think that qualitative research may benefit substantially from stricter formalization, as well. A prime example would be the formalization of individuals' belief systems.

In order to enable such progress, psychology curricula need to be strengthened in terms of mathematics, logic, and formal modelling. Formal modelling is still not very common today in psychology for the simple reason that most psychologists do not know much about it. But that may be changed, by better training, and by more intensive research collaboration with mathematicians and modelling experts in particular (Meehl, 1990b). Also, a wealth of helpful resources has recently become available to foster theory formalization (Devezer et al., 2021; Gray, 2017; Robinaugh et al., 2020; Smaldino, 2020; van Rooij & Baggio, 2020; van Rooij & Blokpoel, 2020; West et al., 2019).

It has been argued that some psychological phenomena might just be too complex to be described in terms of mathematically-formalized theories (Sanbonmatsu & Johnston, 2019). However, recent work shows that such a formalization *is* possible, at least in principle, for complex psychological theories, as well. In the most comprehensive analysis of this kind to date (Hale et al., 2020; West et al., 2019), 76 prominent (mainly verbal) theories of behavioral change were successfully specified using logical propositions and graphical displays, including definitions of all relevant concepts.

Step 7: Making Pre-Registration Mandatory for Confirmatory Claims

Pre-registration should be made mandatory whenever researchers want to make a "confirmatory" claim, that is, assert that a theory is able to reliably *predict* a certain outcome (Wagenmakers et al., 2012). The main goals of pre-registration are 1) to rule out HARKing (i.e., hypothesizing after the results are known; Kerr, 1998), which is the common but inacceptable practice of incorrectly claiming that what one found is what one had expected in advance, and 2) to rule out data-dependent analysis choices (e.g., outcome switching, strategic dropping of outliers or dependent variables) that are often used to achieve results that are better aligned to the hypotheses or simply more presentable (Hardwicke & Wagenmakers, 2021).

Trustworthy and sustainable platforms should be used for all kinds of pre-registrations. A pre-registration may be embargoed for a limited amount of time (i.e., receive a time stamp but not be made public yet), to offer protection from the possibility of scooping. At some point, however, pre-registrations *must* go public, in order to prevent selective reporting. Thus, the practice may also be expected to help reduce the so-called "file drawer problem" (i.e., non-significant research remaining unpublished, which leads to overestimations of effect sizes in meta-analyses). This effect can be maximized when research reports are required to contain a complete registry of all pre-registrations that the authors have ever performed with regard to a given hypothesis.

At present, pre-registrations still differ significantly from one another in terms of specificity. As a rule of thumb, pre-registrations should always be as specific as possible, because the higher the specificity, the more informative the results will be. An ideal pre-registration covers the participants, experimental conditions, and measures on which the data is or will be based, the exact hypotheses, and the exact statistical tests that will be applied to the data to test those hypotheses, including a mention of how multiple testing will be dealt with, and a priori rules for excluding observations (if applicable). Ideally, the goal of a pre-registration should be to reduce "researcher degrees of freedom" to *zero*.

In cases where a pre-registration uses an existing dataset that has already been analyzed with regard to other research questions, those previous analyses have to be made transparent as well, to provide a window into how "pre" the current pre-registration really is.

All too often, the analyses that are actually performed deviate from those that were pre-registered, without these discrepancies being acknowledged (Claesen et al., 2019). Reviewers and editors have to enforce transparency in this regard, by barring papers with undeclared discrepancies between pre-registered and performed analyses from publication. We recommend making it a routine practice to detail all deviations from a pre-registration in the Methods section of a paper. Ideally, published papers should link to elements of the pre-registration using unique identifiers (“smart preregistrations”, https://osf.io/t8yjb/; see also the consensus pre-registration template from APA, BPS, COS, DGPs, and ZPID, https://prereg-psych.org/index.php/rrp/templates).

Notably, research results may only be labelled “confirmatory” in nature if the specifics of the respective analyses were already anticipated in the pre-registration. *This* is the kind of research result that may appropriately be accompanied by *p*-values. Everything else must be reported as being “exploratory”, and using *p*-values (especially ones not corrected for multiple testing) will usually be inappropriate in this context. We recommend using two distinct headings (e.g., “pre-registered analyses” vs. “other analyses”) in Results sections to explicitly mark this fundamental difference.

Various pre-registration templates are available or under development (Bosnjak et al., 2021; https://help.osf.io/article/229-select-a-registration-template). We are aware that some researchers argue that a theory (or a grant proposal) may by itself be specific enough to serve as a kind of surrogate for a pre-registration. However, theories and grant proposals usually do not specify exact operationalizations, statistical analyses, or rules for outlier exclusions, to name just a few required elements of proper pre-registrations. To maximize the beneficial effect of pre-registration on research credibility, this additional layer of exactness is needed.

In our own experience, strict pre-registrations of the kind that we advocate here can be very humbling experiences for at least two reasons: 1) in the process of writing them, it may become obvious just *how* vague and fuzzy one’s theory actually is, and 2) once the results are in, it may become obvious how lacking or just plain wrong one’s theory actually was. Thus, pre-registrations have the potential to become real game-changers in psychological research, by significantly *lowering* the subjective confidence of psychologists in their own theoretical understanding of the world, while at the same time improving considerably on the validity of those (probably few) theoretical claims that actually survive the process.

The most recommendable but also the most effortful variant of pre-registration is the submission of a “registered report” to a journal (e.g., Chambers, 2019). Here, the whole analysis plan (and - in cases where the data still has to be collected - the data collection plan) is submitted up front for review. The journal may award an “in principle acceptance” to a positively reviewed manuscript, possibly accompanied by a few recommendations for improvement. Such a positive decision means that the *planned* research is deemed important and sound enough for the journal to publish it irrespective of the outcome. This format is the most recommendable for at least two reasons: 1) the preceding review almost invariably helps improve the quality of the research by a large margin, and 2) the whole process will be documented in the most visible manner, which makes it much more difficult to hide “inconvenient” (e.g., theory-incongruent) results.

These benefits come on top of those that all pre-registrations share anyway (see above). In fact, recent research suggests that this new publication format is actually capable of dramatically reducing the problem of false positives in the research literature (Scheel et al., 2021).

All forms of pre-registration, if done properly, require significant amounts of time and effort. It is not at all unusual to spend additional weeks or even months on this phase of the research process, because all the conceptual vagueness and uncertainties that would remain opaque under the “traditional”, non-pre-registered approach now become painfully apparent and have to be dealt with. This more careful preparation, however, not only increases the quality of the work and helps avoid unnecessary glitches and flaws. It may also save a substantial amount of time later in the process, when one is analyzing the data and writing up the results. Nevertheless, research utilizing proper pre-registration invariably requires greater care and involves greater risk, which is why scientists need to be incentivized significantly for undertaking it.

Step 8: Making Replication Attempts a Routine Practice

Effects that have been independently replicated may generally be regarded as being more trustworthy than effects that have not. We assume that the so-called “replication crisis” in psychology (Pashler & Wagenmakers, 2012) is in part a consequence of the field’s systematic undervaluing of replication studies (Asendorpf et al., 2013a; Nosek et al., 2012). Whenever possible, researchers should attempt to replicate their own findings and/or others’ findings, and let the world know about these attempts, regardless of whether they succeeded or not. All such attempts should thus be pre-registered, again to avoid selective reporting, and to enable realistic meta-analytic estimations of effect sizes.

Obviously, just calling for journals to show greater openness to publishing replication attempts will not be enough. Also, creating whole journals that publish nothing but (failed) replication attempts will not make this kind of work more attractive (Nosek et al., 2012). A more promising way of tackling the existing *novelty bias* of academic journals (i.e., preferring novel findings over replications) would be for journals to explicitly reserve a certain quota of pages - per year, or per issue - for such research, and to call on authors to help fill those pages. Journals that do acknowledge the crucial role of replication attempts for scientific progress that way should be considered “better” than journals that do not. Moreover, writing replication papers may become more attractive for authors when journals explicitly encourage an “ultra-brief” format (e.g., 2 printed pages) for this type of contribution. This is possible because replication attempts should share most or all of their background and methods sections with the original work, which would thus not have to be repeated. These parts of a replication paper may instead be replaced by a simple reference to the original work.

Step 9: Planning for Informative Studies

The goal of conducting a scientific study is to *learn* something. If the design of a study implies that the result will most likely be uninformative, one could rightly ask why such a study should be conducted at all, as it may well just be a waste of money, effort, and (researchers' as well as participants') time. Thus, studies should always be planned in ways that ensure a high probability of being informative and providing strong evidence. Frequentist hypothesis tests will provide the most informative results if their expected false-positive and false-negative error rates are low (Bayarri et al., 2016). Thus, the sample size and other features such as research design and planned analyses should be chosen with the goal of minimizing those error rates under the existing resource constraints. If acceptable error rates may not be achieved given those constraints, conducting the study should normally be discouraged (there may be exceptions, e.g., when it is very difficult to recruit participants from the population of interest).

What does "acceptable" mean in this regard? The answer to this question depends on the specific effect under scrutiny, and ideally considers the (potentially asymmetric) costs of wrong decisions in either direction. As a rough rule of thumb, and in line with standards currently advocated by many in the field, we recommend a statistical power of at least 80% to detect small to medium-sized effects, while keeping the expected false-positive rate low (no higher than 5% for two-tailed hypothesis tests). For example, no fewer than 100 (200) target persons should be tested when the expected effect size is comparable to a correlation of $r = .30$ (.20). For studies which claim new discoveries, even more rigorous standards have been advocated (especially a false-positive rate < 0.5%; Benjamin et al., 2018).

The respective deliberations should be explicated, ideally as part of a pre-registration (see Step 7). This necessarily involves the expected effect size and a justification for the same. If such a justification is based on previous studies, it should take into account the likely influence of publication bias which may have led to an inflation of the effect sizes previously reported. Ideally, however, the expectation is justified in terms of the minimum effect that would be seen as theoretically and/or practically relevant. If multiple analyses are planned, the problem of error inflation will have to be accounted for, as well. If the goal is to estimate a parameter, a study will only be informative if the estimate has a sufficiently high precision. In this case, one should plan for the desired width of the confidence interval (Maxwell et al., 2008). If one wants to compare models using Bayes factors, a Bayes factor design analysis provides the necessary means to tune a study's design in a way that it has a high chance of providing compelling evidence (Schönbrodt & Wagenmakers, 2018). However, given that there are no established benchmarks yet regarding these latter conceptualizations of informativeness, we just mention them here and only account for statistical power in our proposed reward scheme (see below).

More broadly, studies are informative to the extent that the participant sample is representative for the population of interest, and to the extent that the experimental stimuli are representative for the "population" of environmental conditions whose influence one aims to understand. Contrary to some misconceptions, the representativeness of a study is largely independent of its sample size. Rather, "representative" means that all members of the population have the same chance of ending up in the study sample, leading to distributions of variables that are similar in the population and the sample. Both kinds of representativeness (especially the second) are yet very rare in personality psychology, despite their crucial importance for the validity of any conclusions that may be drawn from a research study (Henrich et al., 2010). Thus, participant and stimulus representativeness need to be rewarded much more. Notably, a study may have representativeness in one or both of these regards *without* testing any a priori specified hypothesis — so, informative studies do not have to be confirmatory in nature.

Step 10: Sharing Data, Code, and Materials by Default

Transparency should be a core value of any science. In order to enable other researchers to critically evaluate the soundness of a piece of scientific work, and to possibly build on that work, authors should routinely make the materials, data, and analysis code associated with it publicly available. Sometimes, there may be legitimate reasons to restrict the access to research data (e.g. as scientific use files; Gollwitzer et al., 2021) or even not to share some of these things at all (Meyer, 2018), but then these reasons need to be explained in the paper (e.g., some types of biopsychological raw data are difficult or impossible to completely anonymize). As an alternative, "synthetic data sets" (i.e., data sets that mimic real data sets by preserving statistical properties and relationships between variables) may be used (e.g., see Quintana, 2020 for a practical guide and R script).

In the course of the last decade, sharing materials, data, and code has become fairly easy and straightforward, as several suitable online platforms are now available for this purpose (e.g., OSF, Zenodo, PsychArchives, OpenfMRI). Although the exact ways in which psychological research data should be made available are still subject to debate, some consensus has begun to emerge. We encourage the use of a standardized data format and a standardized structure of folders and files, to foster transparency, mutual understanding, and collaboration among different labs. Furthermore, data sets should be documented with informative meta-data, including comprehensive codebooks and detailed information on the data collection (e.g., recruitment, sampling procedure, or relevant technical details such as sampling rates in psychophysiological research). Some emerging standards in this regard have recently been published (e.g., Gollwitzer et al., 2021) or are currently under development (e.g., psych-DS: https://psych-ds.github.io/, or PsyCuraDat: Blask et al., 2020).

The additional workload associated with data sharing will vary considerably, depending on 1) how complex and extensive the data is and 2) which approach a researcher adopts: Mere “data dumps” accompanied by little to no explanation will often be of limited or no value at all to others. This practice should thus not be rewarded. The minimum requirement, in our view, is to make available all of the data underlying the analyses reported in a paper, in a way that enables others to easily reproduce these analyses without having to consult with the authors of the original paper again. This means that it has to be at least made clear which terms in the paper refer to which variables in the dataset(s), and how to interpret the individual levels of these variables (e.g., 1 = female, 2 = male). Ideally, research data should be shared in accordance with FAIR criteria (Findable, Accessible, Interoperable, Reusable; Wilkinson et al., 2016). This will give others the opportunity to use the data for investigating *additional* research questions without having to consult with the original authors.

Making data accessible in FAIR format can easily consume weeks of additional work, especially when the data structure is complex. Doing so should thus be rewarded significantly, as it may be of tremendous value to the field, in particular in combination with our proposed consensus steps 1 to 5: If meta-data of shared data sets refer to an accepted ontology and well-defined measurement instruments, search engines may, for example, retrieve all data sets that use a certain measurement method or that refer to certain constructs of interest. Standardized data formats allow best-practice preprocessing pipelines for certain data types to be used across data sets (as already done in the BIDS data format, see http://bids-apps.neuroimaging.io/apps/).

As the same data collection effort may spawn several papers, we propose that all of the data, code, and materials associated with a given project should be stored in a single place online. Duplications should be avoided by any means possible. If not all of the data is supposed to be publicly available from the get-go, access to parts of the data may be embargoed for a limited and defined amount of time, during which the original authors may conduct their planned analyses on those parts of the data (Gollwitzer et al., 2021). The directory containing all the content related to a project should become associated with the first research article based on this content, and we should make it a custom that all subsequent papers based on the same data (by the same or different authors) should cite that first paper. That way, preparing materials, data, and code in ways that make them easy to use for others would become even more attractive for authors. Moreover, such an approach would make it less likely for meta-analyses to treat the same data as if it came from independent samples, which may happen if sample overlap is not explicitly specified in the original research articles. Also, citing that first paper, and thus the associated directory, will make it unnecessary to repeat the “21-word solution” (Simmons et al., 2012) again and again, because a complete list of everything that was done in the course of the original data collection will be permanently available online.

Pathways Toward Implementation

As noted above, most of the steps that we advocate here have been proposed before, in one form or another. Yet there is a very palpable discrepancy between the intuitive plausibility of these steps in the service of Good Science, and the relative reluctance with which they are being adopted in many places. Some of this may be attributable to a mere *lack of information* on the side of potential adopters, whereas some may be attributable to the inherent *inertia* of a system as big and complex as academic psychology. Structures and processes tend to be perceived as unchangeable the longer they have existed unchanged, and often responsibilities are diffuse and communication pathways are inefficient.

However, we assume that some of this reluctance is attributable to the fact that many of these steps are in conflict with vested interests on the side of people and institutions who *profit* from the status quo, and from a social dilemma structure (Dawes, 1980; Nosek & Bar-Anan, 2012): Becoming an author on large numbers of rather weak research articles is not only relatively easy, but also a promising way toward earning all sorts of rewards under the current incentive structure (e.g., jobs, money). “Scientists” who are inclined to behave accordingly will be motivated to keep it that way, and to fight, or at least not support, any attempt at improving scientific rigor. Research institutions interested in improving their standings in rankings will be motivated to reward those same people for doing just that. Moreover, the primary interest of commercial scientific publishers is not to maximize scientific quality, but revenue (Aspesi et al., 2019). Hence, the collectively rational solution (“getting it right”) is not yet sufficiently aligned with incentives for individual researchers (e.g., “getting it published”), resulting in a dilemma structure that entails potentially detrimental effects for scientists that adhere to higher standards (Nosek et al., 2012).

The result is, lamentably, a vast and badly organized research literature incorporating far too many contributions of questionable value. Knowledge development progresses more slowly than it could, and large amounts of public resources are being wasted. This needs to change, and we are optimistic that it will. Our express purpose here is to accelerate this development. In the remainder of this paper, we will outline our vision of how good (i.e., efficient, transparent, and cumulative) science in personality psychology and beyond may become more of the rule rather than the exception.

Changes in Reward Mechanisms

Academia has much to offer to researchers in terms of rewards. These may be roughly grouped into two clusters, *intrinsic* ones and *extrinsic* ones: As for intrinsic rewards, a career in science may provide a person with an opportunity to spend a major part of their life investigating issues that, optimally, are intriguing, exciting, and relevant to society. Such work may be intellectually challenging - which some of us perceive as an attractive quality in itself – and sometimes very rewarding, when hard work finally pays off and an important problem appears to have been solved, or at least brought closer to a solution. Given the complexity of the phenomena of interest, it has now become virtually impossible to do all of this by oneself, so psychological research is — by necessity — becoming more and more collaborative in nature, as well. This in turn may entail constant interactions and stimulating exchanges with other gifted, ambitious, and inspiring people who are interested in the same or similar issues as oneself. What a privilege that is! For these reasons alone, many young idealistic people aspire to a career in academic psychology.

In addition, there are very substantial rewards involved of a more *extrinsic* nature: An unmatched level of job security once tenure has been attained, often coupled with relatively good salaries and an equally unmatched amount of freedom to decide what one wants to work on, how, and with whom. Academic leadership positions such as professorships may also entail quite a bit of power and prestige, both of which may be more appealing to certain personality types than to others.

So, there are many good reasons why a person may want to work in (e.g., psychological) science. A major problem, however, lies in what one has to do in order to *attain* a permanent position in academia. As the number of those jobs is limited, it is inevitable that evaluation criteria have to come into play. We think that the way in which research productivity has been evaluated in the past – and still is being evaluated in many places – is at the core of many of the problems that the research literature in psychology has been shown to have. Most importantly, the sheer *number* of peer-reviewed publications (co-)authored and the *number* of citations to those publications (both of which contribute to the now infamous *h*-index; Hirsch, 2005) are often given a lot of weight when ranking scientists in terms of their potential and/or achievement (Abele-Brehm & Bühner, 2016; Anderson et al., 2007; Chapman et al., 2019; Fong & Wilhite, 2017). Doing so is both practical and convenient, as these numbers may easily be obtained from a database such as Web of Science, Scopus, or Google Scholar. Unfortunately, the *validity* of both of these presumptive measures of scientific merit is unsatisfactory.

Indubitably, good scientific work may lead to impressive publication and citation numbers. The problem is that these numbers are too strongly contaminated with influences that are unrelated, or even in direct opposition, to basic principles of Good Science like the ones we laid out above. For example, the number of papers that is needed to elucidate the degree of (non-)overlap between different personality constructs and/or measures, as well as the number of papers introducing “new” personality constructs or measures will increase the *less* consensus and conceptual clarity there is in a field as to what the relevant dimensions are, and how they should be named (see our Step 2) and measured (see our Step 3). Therefore, personality researchers tend to “reinvent the wheel” again and again (Phaf, 2020). This may happen for self-serving reasons, because they just want their *own* names to be attached to some important topic. It may also happen by accident, because they are simply unable to locate the relevant contributions in the avalanche of publications that already exist, and/or cannot invest the necessary time and energy to thoroughly check for redundancy between their own and previous research. The much-needed antidote to this state of affairs would be a concerted effort to streamline the field (e.g., in terms of constructs and measurement practices), but this would be very hard work and not be sufficiently rewarded at present. We argue that this needs to change.

Often, co-authorships do not so much reflect a person's sizeable scientific contributions, but rather their having power over others, being embedded in a large collaborator network, having access to some kind of research technology or infrastructure, or simply being paid (back) a favor (Anderson et al., 2007; Fong & Wilhite, 2017; Ioannidis, 2008; Kwok, 2005; Reisig et al., 2020). Mechanisms such as these will not increase the number of publications in a *field*, but the number of publications attributed to an individual, which may reap great rewards for that person. In the earlier stages of an academic career, such a reward may consist in (e.g.) being awarded tenure. In the later stages, a person's mere number of co-authorships may even directly translate into personal financial gain (e.g., as a bonus payment from the institution at which he or she is employed).

At present, citation counts are equally questionable measures of scientific merit (Fong & Wilhite, 2017; Thorne, 1977). Obviously, if scientific work is not recognized (cited) at all, it cannot contribute to knowledge development, and there are many examples of scientific papers that get cited a lot for all the right reasons. However, large citation counts may also be the result of an intellectually undemanding treatment of sexy topics. Simple and superficial papers may be read and cited much more readily than difficult and detailed ones. Recently, Serra-Garcia and Gneezy (2021) presented evidence showing that citation rates actually predict the *non-replicability* of findings, and tentatively explained this with the trade-off between scientific rigor and "interestingness" faced by editorial teams.

In personality and clinical psychology, devising a *measure* that is then used by many people almost guarantees high citation numbers. But many of the most popular measures in these fields have actually been developed in a rather "quick and dirty" fashion and are not backed by a lot of conceptual and/or empirical work (e.g., regarding possible redundancy with other measures). Furthermore, the chance to devise an *authoritative* measure (e.g., of some official diagnostic category) that has good chances of being regularly cited often hinges upon one's networking history more than anything else.

Citation counts may also reflect voluntary or coerced efforts to please (potential) reviewers and/or journal editors, personal favors, and the workings of several different types of feedback loops (e.g., papers getting cited just because they were cited before) in which, again, papers get cited more just because they were cited before (Fong & Wilhite, 2017; Teplitskiy et al., 2020). At the same time, there is surprisingly little empirical evidence that citation counts do reflect what most researchers consider to be good *quality* of research (Aksnes et al., 2019; Dougherty & Horne, 2019).

In our view, the problems discussed above reflect an interaction between the current incentive structure (rewarding high publication and citation numbers, largely irrespective of content) and the willingness of individual researchers to take “the path of least resistance”. For example, the number of co-authorships attributed to an author is very easily inflated, at virtually no cost to the people inflating it (Borkenau, 2012), and at very little risk of ever being “caught” or even sanctioned for doing so. Kwok (2005) outlines a strategy that has proven highly successful in securing co-authorships despite close-to-zero involvement with the actual research, while at the same time making it almost impossible to prove that these co-authorships are undeserved. These unfortunate facts put younger researchers in particular in a very difficult situation, as they may come to ask themselves whether doing *good* research instead of just “playing the game” will actually *harm* their own career prospects (cf. Dawes, 1980; Nosek et al., 2012).

An Explicit Scheme for Rewarding Quality in Personality Science

We assume that the disproportionate role played by the mere *numbers* of publications and citations in research evaluation have contributed substantially to the apparent lack of cohesion and integrity that seems to permeate parts of the research literature (not only) in psychology. In our view, this problem needs to be addressed. There has been no shortage of public declarations that, somehow, “quality” needs to play a more prominent role than mere quantity in assessments of academic merit and potential. But calls for greater scientific ambition and rigor will remain ineffective as long as the incentive structure in academia, definitely rewarding quantity more than anything else, continues to stand in the way of change (Chapman et al., 2019). Many — especially young, non-tenured — academics report being faced with the dilemma that the practices that will help them the most in their personal career (or to even have such a career) do not align with those that would be required to help ensure robust scientific progress (Abele-Brehm & Bühner, 2016). Therefore, the incentive structure itself needs to change.

We will not say much about citations, as that is a topic of its own and research illuminating the proper and improper ways in which papers are (not) being cited seems to have just begun. We do embrace the view that citation is, in principle, a valid way of acknowledging the quality and relevance of a scientific contribution. However, there also is a lot of room for improvement in that regard. For example, one may add *qualifiers* to citations, telling the reader *why* a given paper is being cited in a particular context.

In the remainder of the present paper, we will focus only on *publications*, however. Researchers must be rewarded not for publishing *a lot*, but for conducting and publishing *good* research. Fortunately, the question of what good research is does not lie completely in the eye of the beholder. There are some standards in this regard that seem to be widely acceptable (see above), even though their implementation is lacking so far. We suggest explicitly weighting the research that a person is involved in by its adherence to such standards. Table 1 shows a tentative reward matrix for published papers that might be used to this effect. The underlying mechanism is basically a multi-attributive utility analysis, in which reward points reflect the respective weight of each attribute. A similar approach has recently been suggested (Dougherty et al., 2019) but remained at the level of person ratings and used only two holistic attributes (i.e., Transparency & Reproducibility efforts; Quality and Scope of Publications; https://osf.io/gp5qt/), whereas we suggest a more detailed rating system for single papers.

The idea to take some measure of quality into account when assessing research productivity is not new, of course. Sometimes, the impact factor of the journal in which an article has appeared is used for that purpose. However, this practice has long been denounced for a number of valid reasons (e.g., in the San Francisco Declaration on Research Assessment; https://sfdora.org/read; The PLoS Medicine Editors, 2006). For example, even if one does accept the number of citations to a paper as a measure of that paper's scientific merit, the number of citations attracted by papers in the same journal varies dramatically, making the impact factor an extremely imprecise measure of the (likely) impact of individual contributions. Also, given the potential rewards associated with publishing in "high impact" journals, authors may actually be willing to cut a few more corners in order to "make it" into one of those journals, even at the cost of possibly endangering the integrity of their research (Brembs et al., 2013; Dougherty & Horne, 2019). We therefore argue in favor of assessing the quality of individual papers directly, by explicitly weighting them with their Good Science merits, and to pay much less attention to the journals in which they were published.

The column named "Reward Points" in Table 1 contains our suggestions as to how much *additional* value a paper with a given desirable property should be assigned. Under the current incentive structure, ignoring impact factors, any paper that is published in a peer-reviewed journal would receive one point (see first data row of the table). The other rewards listed in Table 1 are supposed to go *on top of that*, to explicitly acknowledge the greater value of papers that have certain desirable properties. For example, a paper that "includes an algebraic or formal-logic formulation of the theory being tested, and how it relates to measured variables" (6a) should not receive just one point, but (1.0 + 2.0 =) three points (e.g., in an evaluation of an applicant's publication record). Notably, these rewards are also meant to be *additive*. The more desirable properties a paper has, the more it should count. For example, a paper that not only meets our criterion 6a, but also has been submitted as a registered report (7b) should receive (1.0 + 2.0 + 2.0 =) five points. Sometimes, paper properties are not independent of one another. For example, data can be made openly available (10a) without being documented in FAIR format (10b), but the reverse is not true. In such cases, the reward values in Table 1 are also supposed to *combine*, in order to avoid hierarchies among individual entries: Authors should be rewarded for making their data publicly available (+ 0.5) and additionally for using FAIR

Table 1

Proposed Reward Scheme

Step	Paper feature	Reward points
0	Paper gets published in a peer-reviewed outlet	**1.0**
1a	Presents broad consensus regarding important research goals	**5.0**
1b	Addresses important research goals that were outlined in consensus document	**0.5**
2a	Presents broad consensus regarding terminology	**5.0**
2b	Uses terminology from consensus document	**0.5**
3a	Presents broad consensus regarding measurement practices	**5.0**
3b	Uses measurement practices from consensus document	**0.5**
4a	Presents broad consensus regarding data pre-processing and/or analysis	**5.0**
4b	Uses consensus practices regarding data pre-processing and/or analysis	**0.5**
5a	Presents broad consensus regarding state of knowledge and/or theory development	**5.0**
5b	Builds directly on consensus document regarding state of knowledge and/or theory development	**0.5**
6a	Includes algebraic or formal-logic formulation of theory being tested, and how it relates to measured variables	**2.0**
6b	Includes account of how the tested formal theory relates to previous formulations of the same or related theories	**1.0**
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**
7b	Is a registered report	**2.0**
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**
9a	Includes pre-registered a priori power analysis / sample size planning based on specific and realistic expected effect size estimates	**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**
9c	Demonstrates representativeness of participant samples(s) in regard to the population of interest	**3.0**
9d	Demonstrates representativeness of stimuli in regard to the environmental conditions of interest	**3.0**
10a	Data is made open	**0.5**
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**
10c	Code is made open (and well documented)	**0.5**
10d	Materials are made open (and well documented)	**0.5**
10e	All data, materials and code from a project are found in a single directory online	**0.5**

format in doing so (+ 1.0), resulting in an overall value of (1.0 + 0.5 + 1.0 =) 2.5 points for a paper that has no other desirable properties apart from this one.

For example: Burt spends three years working with a group of colleagues from multiple other labs to establish consensus among them as to how their favorite construct shall be measured in the future (3a). The consensus paper documenting the outcome of that massive undertaking will count as (1.0 + 5.0 =) six points in the CVs of each of its authors. Furthermore, any future paper actually using those agreed-on measurement practices (3b) will count as (1.0 + 0.5 =) one-and-a-half points in the CVs of each of that paper's authors. Another example: Lisa submits a registered report (7b) about a study she plans that includes a well justified sample size calculation (9a) resulting in an expected Alpha (type I error rate) of .01 and a Beta (type II error rate) of .10 (9b) based on a realistic expected effect size estimate. The paper gets an in principle-acceptance. Then Lisa conducts her study, and reports her findings strictly distinguishing between confirmatory and exploratory analyses (7a). In response to a request by an anonymous reviewer, Lisa also conducts an additional study, attempting to replicate the same effect with even greater statistical power (8). When Lisa's performance as an assistant professor is evaluated by her tenure committee, that paper receives (1.0 + 2.0 + 0.5 + 2.0 + 1.0 + 1.0 =) 7.5 points. Note that this reward value is independent of whether Lisa's replication attempt succeeded or not, as long as it is published as a part of the paper.

The use of an explicit reward scheme like this (e.g., in making hiring decisions) may have a number of desirable effects: First, it would help establish transparency (for applicants and committee members alike) as to what evaluation criteria will be used in regard to scientific quality, and thus also improve on fairness and - probably - inter-rater agreement. Second, it would make it less likely that committee members, despite being initially committed to prioritizing quality, switch back to mainly quantitative assessment later in the process. Third, it may make visible a potential discrepancy between the evaluation criteria that *should* be applied in the service of acquiring robust scientific knowledge, and the criteria that typically *are* being applied in research evaluations (e.g., in university rankings). This in turn could become the starting point for an important discussion.

Especially to readers who have little experience yet with the Good Science practices listed in Table 1, some of the reward values we suggest may seem a bit outrageous at first. By meeting a number of these criteria, a single paper may easily acquire ten times the value of an "ordinary paper" not meeting those criteria. However, we think that the values we propose here may actually be deemed relatively *modest.* They represent our consensual, but still preliminary view of what would be fair, based on our own practical experience as researchers. Doing Good Science *is* much more demanding, which is why researchers so often shy away from it, which in turn is why it is so necessary to reward it more. Still, these are just our suggestions. If an institution wishes to adopt our basic premise and reward the Good Science practices that we advocate here, but just not as much as we propose (or even more!) or in a different way, it is very easy to change these reward values, either individually, or by multiplying all of them with the same number (other than 1). At present, most academic institutions effectively use a factor value of zero for that multiplication. An assessment in terms of fewer, more holistic attributes (cf. Dougherty et al., 2019) would also be viable, as long as it still aligns with the same broad ideas of what Good Science is about (see below).

It may also be argued that such a system of weighting each of the individual papers that a researcher has published with a long list of potential scientific merits is too complicated to be practical. We are not that skeptical. For example, applicants for positions in academia will probably not mind compiling such a list for themselves. Once they have it, it can be sent to any potential employer, and be amended any time with additional, more recent publications. We assume that most applicants who have been made aware that their quality-weighted list of publications may be subjected to checks by the hiring committee at any time will prefer to stick to the truth in what they report. To ease the burden on applicants or candidates and to draw the focus of assessment even more on the most important contributions in terms of content (instead of quantity), one may ask them for an assessment of their (e.g., five) most important publications only.

It may also be argued that the criteria listed in Table 1 are too “technical” in nature, and that we neglect the importance of originality, innovation and relevance in good research. This is inevitably so because originality and innovation, although definitely important, are much harder to assess in a sufficiently objective manner (Starbuck, 2005). Anyone who has ever received – or was involved in providing – “split reviews” may probably attest to that. Thus, the criteria we propose should be viewed as relatively broad, abstract indicators whose presence will make it more likely for research to lead to credible, incremental knowledge growth. In line with many previous calls for reform (see above), we are convinced that these criteria will already go a long way in improving the overall quality of our research. Any desirable qualities of research that go *beyond* these (e.g., originality and relevance) 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. If the goal is to combine criteria for methodological quality (like ours) *and* originality/innovation/relevance in a single score, the latter could also be rated on a global scale by reviewers and then multiplied with the former.

Finally, it should be noted that there are some qualitative differences between the consensus-building criteria (1a, 2a, 3a, 4a, and 5a) and all of the other criteria listed in Table 1. Whereas the latter may basically appear in any combination in any empirical paper, it is most likely that a paper may only be characterized by *one* of the former. The reason is, simply, that consensus-building is such an enormous task. At present, only very few papers in our field present any systematic attempts at consensus-building at all, and each of these covers only one of the five domains we introduced above. Therefore, it should be clear that no paper will ever be able to score high on all of our criteria at once. Rather, most papers will either present a new consensus (5 points) or be empirical in nature and then incorporate about a handful of desirable properties that such papers may have.

Of course, one typical risk of all static reward systems is that people might try to outsmart them. We assume that the reward system we propose here is much harder to outsmart than one in which authors only need to somehow acquire as many authorships as possible. However, one will still have to remain wary as to whether authors - through their work - are true to the *spirit* of such a system, or whether they are just trying to somehow maximize reward points. Additional measures such as requesting explicit research philosophy statements and/or annotated CVs (Dougherty et al., 2019) might be used to reduce the likelihood that such attempts go unnoticed.

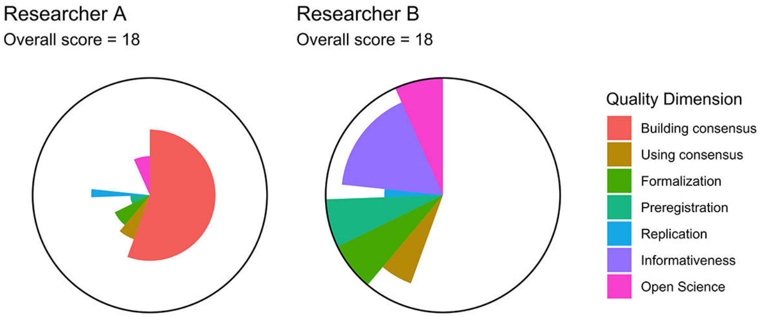
Multidimensional Rating vs. Total Score

Up to here, we treated our proposed rating scheme as being largely unidimensional: The more points a paper gains overall, the better. In many instances, however, it may be useful to engage in more detailed analyses: For example, the 24 desirable paper properties listed in Table 1 may be clustered into 7 broad categories: Building consensus (1a, 2a, 3a, 4a, and 5a), Using consensus (1b, 2b, 3b, 4b, and 5b), Formalization (6a, 6b), Preregistration (7a, 7b), Replication (8), Informativeness (9a–9d), and Open Science (10a–10e). Aggregating across a researcher's (e.g., most recent) publications and then visualizing scores in these seven categories separately may be helpful for various purposes: For example, it may be used to identify *areas of possible improvement* in that researcher's habitual research practices (e.g., Peter never cared much about replication so far, but he should). When combined with an annotated CV (Dougherty et al., 2019), such an analysis may also be used to make transparent certain unavoidable *impediments* to implementing desirable research practices that are owed to a particular researcher's field of study. For example, if Trudy studies patients with very rare conditions, that will make it much harder for her to obtain large samples, or to run many replications. Still, Trudy could aim for robustness of her research in other areas, such as pre-registration.

Such a more differentiated analysis of individual researchers' Good Science profiles is also well in line with a “compensatory” philosophy in which not everyone can be equally good at everything. For example, researchers who invest the great effort that consensus-building requires will almost by necessity have less time and energy available to invest into empirical studies, and thus be unable to obtain many points on criteria related to such studies. Both types of Good Science should be rewarded, however. This is showcased by the two profiles displayed in Figure 1. Both of the two hypothetical researchers (A and B) whose profiles are displayed here attain the same overall score (18), but by different means. Needless to say, such analyses may also be conducted at even higher levels of aggregation (e.g., whole departments).

Figure 1

Visualizing the Research Quality Profiles of two Researchers (A and B) who Promote Good Science in Different Ways, Through Their Respective Activities



Note. The width of each wedge is proportional to the maximum number of points that may be obtained in each category.

Consequences for Overall Publication Numbers

The widespread implementation of such a more quality-oriented reward system might result in a significant decline of the number of articles being published, both per researcher and in the field overall. According to our view, that would be a very desirable outcome, for several reasons: First, it would mean that the number of manuscripts that need to be *reviewed* also declines, which leaves reviewers more time and energy to give constructive feedback on the fewer manuscripts that *are* being submitted (e.g., including checks of pre-registrations, materials, data, and code). This, in turn, would help improve the overall quality of publications even more. A smaller number of manuscripts under review may also help shorten reviewing times considerably.

Second, lower overall publication numbers would enable researchers to keep better track of what is happening in their field, because it would become possible to *actually read and digest* a larger proportion of what is being published (Phaf, 2020). This in turn may become a key factor in fostering more cumulative — and thus more efficient — research, as all of us would simply be better informed about each other's work, which in turn might spark all sorts of more synergetic projects (Forscher et al., 2020). Third, and most importantly, the signal-to-noise ratio in the psychology literature would improve: Published research may be trusted more, and producing trustworthy research results is what scientists are ultimately being paid for.

Credit Sharing

Our focus in the present paper is on explicating a few core properties of "good" research, and on providing suggestions as to how researchers may be encouraged to do more such research. We will not devote much attention to the question of how the credit for such work may appropriately be shared, as this is clearly an issue of its own (e.g., Ioannidis, 2008). We do, however, explicitly endorse the currently evolving practice to require detailed statements from authors regarding their individual contributions. This is considerably more informative than the traditional approach of indiscriminately assigning greater value to the first and last positions in an order of authors, irrespective of how someone managed to attain one of these positions. Also, the threshold for actually *lying* about these things is probably higher than the threshold for just having someone's name added to a list of authors for some unspecified reason.

Notably, contemporary recommendations regarding the specifics of such a "contributorship model" explicitly *abandon* the requirement that a person making a significant contribution to a research paper must always have been involved with the *writing* of that paper (https://casrai.org/credit; Holcombe, 2019; McNutt et al., 2018). For consensus papers, we recommend naming a task force as author, whose members are listed alphabetically in an appendix. Such a task force should comprise all the individual researchers who not only worked on the consensus paper, but who also explicitly endorse and declare themselves bound to the consensus, at the time of publication.

Top-Down vs. Bottom-Up Implementation

Organizational change is rarely easy. When asking who would be responsible for implementing the changes we propose here, several possible agents come to mind. Individual researchers may do a lot to improve the quality of their own research all by themselves. For example, they may pre-register their research designs and analysis plans, and make their materials, data, and code publicly available. They may also start co-ordinating better with other researchers in terms of terminology and measurement practices, possibly resulting in some sort of preliminary, local consensus among them. This is the bottom-up or "grass roots" approach to change, and we are happy to see more and more personality scientists go down that road already, even *against* the current incentive structure.

However, *institutions* need to embrace these new evaluation practices as well (cf. Dougherty et al., 2019). For example, every psychology department has the responsibility to determine how much weight it will assign to indicators of research quality in making decisions as to who gets a job interview, an award, a bonus, or tenure. The current paper contains a concrete proposal for how a better incentive structure might look like. If Good Science indicators play too little of a role in making these decisions, it is the responsibility of the people within a department to demand and ensure changes to the current incentive structure. Notably, the bulk of this responsibility falls on the *senior* department members (Chapman et al., 2019), because they have the greatest power to actually change the rules. To facilitate an increased use of quality-weighting, it may also be helpful to rethink and improve rewards for committee work.

It is not unheard of, however, that people within a department pass the responsibility for implementing the necessary changes on to the people at the next higher level (i.e., their university leadership) who then externalize that responsibility completely (e.g., to the institutions compiling university rankings, whose evaluation criteria "we will never be able to change"). Also, there is undeniably a multi-level social dilemma involved here (Nosek & Bar-Anan, 2012) in that individuals and institutions who actually dare to move away from behaviors that accord with the current incentive structure will necessarily be disadvantaged for some time, as long as this incentive structure persists.

Given this social dilemma, some relatively technical counter-measures such as "critical mass building" have been proposed (Nosek & Bar-Anan, 2012). Under this approach, more and more researchers would commit to different behavioral standards, but but these commitments would only become effective once a sufficiently large number of their colleagues has done so, as well. Despite being theoretically sound, however, we are fairly skeptical that such an approach will ever be actually be implemented. The much more promising approach would be for individual researchers to change their ways and starting doing their research differently, explicitly accepting the disadvantages that this will earn them under the current incentive structure, in the service of better science. Again, the main responsibility for promoting change this way lies with tenured faculty because for them it *does* take courage to do so (e.g., go against the expectations of their institution's leadership), but it will *not* cost them their careers and/or livelihoods.

In addition, it would certainly be very helpful if academic societies also endorsed the idea of explicitly rewarding people for engaging in Good Science practices. In our case (personality psychology), this would concern societies such as ARP, EAPA, EAPP, DGPs-DPPD, ISSID and/or SPSP. Furthermore, substantial enforcement power also lies with academic journals and funding agencies, whose policies should also reflect Good Science principles as much as possible.

Coda

Talking about "good" science as we do in the present paper logically necessitates the existence of a set of *values* with which the current realities in academic research and publishing may be compared. In writing this paper, we came to realize that the steps toward improvement that we propose here do reflect a few core values of ours. First, transparency: A credible science should have nothing to hide. Transparency is achieved (e.g.) when making one's materials, data, and code openly accessible, but also when admitting (e.g., via pre-registration) that the outcome of a study is not in line with what one expected. Second, collaborativeness: Anyone who has ever been involved in a difficult, effortful, long-term research project knows that credible psychological research is just not possible anymore in a “lone genius” and/or “quick and dirty” fashion. Rather, the work involved (both in terms of energy and mere intellectual difficulty) is often enormous and now requires larger and larger teams of highly qualified and dedicated researchers. Collaborativeness becomes especially evident when groups of scientists take it upon themselves to formulate the current consensus in their field, in order to gain more solid, common ground that their own and others’ future research may then build on. Third, efficiency: It goes without saying that the vast amount of effort that we as psychologists often invest into our scientific work should “pay off” in terms of actual knowledge gains. Efficiency is improved when (e.g.) scientists boil their often relatively vague and fuzzy theoretical ideas down to their essence, by using logical/mathematical formulations, which then allows them to check more carefully for redundancy and compatibility between theories. Efficiency is also gained when researchers allow everyone to re-use the materials, data, and code that they have accumulated in the course of their own research projects. Fourth, accountability: We scientists are predominantly paid by society, for using our intellectual capacity to generate new knowledge that will ultimately be of use to said society. Accountability is improved when (e.g.) researchers start engaging more openly with the public that pays them, regarding the work they do.

We believe that psychologists will be able and willing to align their work more closely with these values when they are explicitly encouraged to, and rewarded for embracing higher quality standards in their research. There is a good chance that the outcome will be a type of psychological science that is more trustworthy and can be better relied upon.

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 Methodological Consultant members to 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 | Visualization (data presentation, figures, etc.) | Writing | Feedback, revisions.

Related Versions: The manuscript was published as a preprint on PsyArXiv. The preprint can be accessed via the following link https://psyarxiv.com/6btc3/

Supplementary Materials

For this article, an Open Peer-Review is available via PsychArchives (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Personality Science. (Ed.). (2022). *Supplementary materials to "Ten steps toward a better personality science – how quality may be rewarded more in research evaluation"* [Open peer-review]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5652

References

Abele-Brehm, A. E., & Bühner, M. (2016). Wer soll die Professur bekommen? Eine Untersuchung zur Bewertung von Auswahlkriterien in Berufungsverfahren der Psychologie [Who should receive the professorship? A research on the evaluation of different hiring criteria for appointments in academic psychology]. *Psychologische Rundschau, 67*(4), 250–261. https://doi.org/10.1026/0033-3042/a000335

Ackerman, R. A., Donnellan, M. B., & Wright, A. G. C. (2019). Current conceptualizations of narcissism. *Current Opinion in Psychiatry, 32*(1), 32–37. https://doi.org/10.1097/YCO.0000000000000463

Adolphs, R. (2015). The unresolved problems of neuroscience. *Trends in Cognitive Sciences, 19*(4), 173–175. https://doi.org/10.1016/j.tics.2015.01.007

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open, 9*(1), 1–17. https://doi.org/10.1177/2158244019829575

Anderson, N. H. (1968). Likableness ratings of 555 personality trait-words. *Journal of Personality and Social Psychology, 9*(3), 272–279. https://doi.org/10.1037/h0025907

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics, 13*(4), 437–461. https://doi.org/10.1007/s11948-007-9042-5

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . .Wicherts, J. M. (2013a). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108–119. https://doi.org/10.1002/per.1919

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . .Wicherts, J. M. (2013b). Replication is more than hitting the lottery twice. *European Journal of Personality, 27*(2), 138–144.

Aspesi, C., Allen, N., Crow, R., Daugherty, S., Joseph, H., McArthur, J., & Shockey, N. (2019). SPARC landscape analysis: The changing academic publishing industry - implications for academic institutions. https://doi.org/10.31229/osf.io/58yhb

Back, M. D. (2020). Editorial: A brief wish list for personality research. *European Journal of Personality, 34*(1), 3–7. https://doi.org/10.1002/per.2236

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology, 72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . .Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Blask, K., Latz, M., Müller, M.-L., Kraffert, S., & Arnold, V. (2020). PsyCuraDat: Development of user-oriented curation criteria for psychological research data. ZPID (Leibniz Institute for Psychology). https://doi.org/10.23668/PSYCHARCHIVES.4477

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*(2), 187–215. https://doi.org/10.1037/0033-2909.117.2.187

Borkenau, P. (2012). Not all authorships are created equal. *Measurement: Interdisciplinary Research and Perspectives, 10*(3), 147–148. https://doi.org/10.1080/15366367.2012.720192

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020, February 29). *Theory Construction Methodology: A practical framework for theory formation in psychology.* PsyArXiv. https://doi.org/10.31234/osf.io/w5tp8

Bosnjak, M., Fiebach, C., Mellor, D., Mueller, S., O'Connor, D. B., Oswald, F. L., & Sokol-Chang, R. I. (2021). *A template for preregistration of quantitative research in psychology: Report of the Joint Psychological Societies Preregistration Task Force.* PsyArXiv. https://doi.org/10.31234/osf.io/d7m5r

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., . . .Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582*, 84–88. https://doi.org/10.1038/s41586-020-2314-9

Box, G. E. P., & Luceno, A. (1997). *Statistical control by monitoring and feedback adjustment.* Wiley.

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience, 7*, Article 291. https://doi.org/10.3389/fnhum.2013.00291

Cain, N. M., Pincus, A. L., & Ansell, E. B. (2008). Narcissism at the crossroads: Phenotypic description of pathological narcissism across clinical theory, social/personality psychology, and psychiatric diagnosis. *Clinical Psychology Review, 28*(4), 638–656. https://doi.org/10.1016/j.cpr.2007.09.006

Cattell, R. B., & Nesselroade, J. R. (1967). Likeness and completeness theories examined by sixteen personality factor measures on stably and unstably married couples. *Journal of Personality and Social Psychology, 7*(4, Pt. 1), 351–361. https://doi.org/10.1037/h0025248

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton University Press.

Chambers, C. (2019). What's next for registered reports? *Nature, 573*(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Chapman, C. A., Bicca-Marquez, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., . . . Stenseth, N. C. (2019). Games academics play and their consequences: How authorship, h-index and journal impact factor are shaping the future of academia. *Proceedings of the Royal Society B, 286*(1916), Article 20192047. https://doi.org/10.1098/rspb.2019.2047

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2019). *Preregistration: Comparing dream to reality*. PsyArXiv. https://doi.org/10.31234/osf.io/d8wex

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science, 8*(3), Article 200805. https://doi.org/10.1098/rsos.200805

Dougherty, M. R., & Horne, Z. (2019). *Citation counts and journal impact factors do not capture research quality in the behavioral and brain sciences*. PsyArXiv. https://doi.org/10.31234/osf.io/9g5wk

Dougherty, M. R., Slevc, L. R., & Grand, J. A. (2019). Making research evaluation more transparent: Aligning research philosophy, institutional values, and reporting. *Perspectives on Psychological Science, 14*(3), 361–375. https://doi.org/10.1177/1745691618810693

Dumas, J. E., Johnson, M., & Lynch, A. M. (2002). Likableness, familiarity, and frequency of 844 person-descriptive words. *Personality and Individual Differences, 32*(3), 523–531. https://doi.org/10.1016/S0191-8869(01)00054-X

Elson, M. (2016). *FlexibleMeasures: Competitive reaction time task*. OSF. https://osf.io/4g7fv/

Elson, M. (2017). *FlexibleMeasures: Go/No-Go task*. OSF. https://osf.io/gsx52/

Elson, M. (2019). Examining psychological science through systematic meta-method analysis: A call for research. *Advances in Methods and Practices in Psychological Science, 2*(4), 350–363. https://doi.org/10.1177/2515245919863296

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment, 26*(2), 419–432. https://doi.org/10.1037/a0035569

Engel, C. (2015). Scientific disintegrity as a public bad. *Perspectives on Psychological Science, 10*(3), 361–379. https://doi.org/10.1177/1745691615577865

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Farber, G., Wolpert, M., & Kemmer, D. (2020, June 20). *Common measures for mental health science: Laying the foundations*. wellcome. https://wellcome.org/sites/default/files/CMB-and-CMA-July-2020-pdf.pdf

Feyerabend, P. (1993). *Against method*. Verso.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS One, 12*(12), Article e0187394. https://doi.org/10.1371/journal.pone.0187394

Forscher, P. S., Wagenmakers, E. J., Coles, N. A., Silan, M. A., Dutra, N. B., Basnight-Brown, D., & IJzerman, H. (2020). *The Benefits, Barriers, and Risks of Big Team Science.* PsyArXiv. https://doi.org/10.31234/osf.io/2mdxh

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

Gollwitzer, M., Abele-Brehm, A., Fiebach, C. J., Ramthun, R., Scheel, A., Schönbrodt, F. & Steinberg, U. (2021). Management und Bereitstellung von Forschungsdaten in der Psychologie: Überarbeitung der DGPs-Empfehlungen. *Psychologische Rundschau, 72*, 132–146. [English version] https://doi.org/10.31234/osf.io/24ncs

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

Hale, J., Hastings, J., West, R., Lefevre, C., Direito, A., Bohlen, L. C., . . ., & Michie, S. (2020). An ontology-based modelling system (OBMS) for representing behaviour change theories applied to 76 theories [Version 1; Peer review: Awaiting peer review]. *Wellcome Open Research, 5*(177). https://doi.org/10.12688/wellcomeopenres.16121.1

Hardwicke, T. E., & Wagenmakers, E.-J. (2021). *Preregistration: A pragmatic tool to reduce bias and calibrate confidence in scientific research.* MetaArXiv. https://osf.io/preprints/metaarxiv/d7bcu/

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

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hmel, B. A., & Pincus, A. L. (2002). The meaning of autonomy: On and beyond the interpersonal circumplex. *Journal of Personality, 70*, 277–310. https://doi.org/10.1111/1467-6494.05006

Holcombe, A. O. (2019). Contributorship, not authorship: Use CRediT to indicate who did what. *Publications, 7*(3), Article 48. https://doi.org/10.3390/publications7030048

Ioannidis, J. P. A. (2008). Measuring co-authorship and networking-adjusted scientific impact. *PLOS ONE, 3*(7), Article e2778. https://doi.org/10.1371/journal.pone.0002778

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

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

Kuhn, T. S. (1990). The road since structure. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1990*, 3-13. http://www.jstor.org/stable/193054

Kwok, L. S. (2005). The White Bull effect: Abusive co-authorship and publication parasitism. *Journal of Medical Ethics, 31*(9), 554–556. https://doi.org/10.1136/jme.2004.010553

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91-196). Cambridge University Press.

Leising, D., & Borgstede, M. (2020) Hypothetical constructs. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer. https://doi.org/10.1007/978-3-319-28099-8_679-1

Leising, D., Borkenau, P., Zimmermann, J., Roski, C., Leonhardt, A., & Schütz, A. (2013). Positive self-regard and claim to leadership: Two fundamental forms of self-evaluation. *European Journal of Personality, 27*(6), 565–579. https://doi.org/10.1002/per.1924

Leising, D., Locke, K., Kurzius, E., & Zimmermann, J. (2016). Quantifying the association of self-enhancement bias with self-ratings of personality and life satisfaction. *Assessment, 23*, 588–602. https://doi.org/10.1177/1073191115590852

Leising, D., Ostrovski, O., & Borkenau, P. (2012). Vocabulary for describing disliked persons is more differentiated than vocabulary for describing liked persons. *Journal of Research in Personality, 46*(4), 393–396. https://doi.org/10.1016/j.jrp.2012.03.006

Leising, D., Scherbaum, S., Locke, K., & Zimmermann, J. (2015). A model of “substance” and “evaluation” in person judgments. *Journal of Research in Personality, 57*, 61–71. https://doi.org/10.1016/j.jrp.2015.04.002

Lin, W., & Green, D. P. (2016). Standard operating procedures: A safety net for pre-analysis plans. *PS: Political Science & Politics, 49*(3), 495–500. https://doi.org/10.1017/S1049096516000810

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology*. Advance Online Publication. https://doi.org/10.1037/cap0000222

Marcum, J. A. (2017). Evolutionary philosophy of science: A new image of science and stance towards general philosophy of science. *Philosophies, 2*(4), 25–35. https://doi.org/10.3390/philosophies2040025

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

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McPhetres, J., Albayrak-Aydemir, N., Barbosa Mendes, A., Chow, E. C., Gonzalez-Marquez, P., Loukras, E., Maus, A., O’Mahony, A., Pomareda, C., Primbs, M. A., Sackman, S. L., Smithson, C. J. R., & Volodko, K. (2021). A decade of theory as reflected in Psychological Science (2009–2019). *PLOS ONE, 16*(3), Article e0247986. https://doi.org/10.1371/journal.pone.0247986

McNutt, M. K., Bradford, M., Drazen, J., Hanson, B., Howard, B., Jamieson, K. H., . . ., & Verma, I. M. (2018). Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2557–2560. https://doi.org/10.1073/pnas.1715374115

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1990b). 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

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review, 125*(5), 656–688. https://doi.org/10.1037/rev0000111

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

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. https://doi.org/10.1177/1745691612459058

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Phaf, R. H. (2020). Publish less, read more. *Theory & Psychology, 30*(2), 263–285. https://doi.org/10.1177/0959354319898250

Platt, J. R. (1964). Strong inference. *Science, 146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: Informatics and the search for mental structure. *Annual Review of Psychology, 67*(1), 587–612. https://doi.org/10.1146/annurev-psych-122414-033729

Popper, K. R. (1934/2002). *The logic of scientific discovery* [Original published as: Logik der Forschung]. Routledge Classics.

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife, 9*, Article e53275. https://doi.org/10.7554/eLife.53275

Reisig, M. D., Holtfreter, K., & Berzofsky, M. E. (2020). Assessing the perceived prevalence of research fraud among faculty at research-intensive universities in the USA. *Accountability in Research, 27*(7), 457–475. https://doi.org/10.1080/08989621.2020.1772060

Robinaugh, D., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. (2020). *Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction*. PsyArXiv. https://doi.org/10.31234/osf.io/ugz7y

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science, 28*(12), 1821–1832. https://doi.org/10.1177/0956797617723726

Sanbonmatsu, D. M., & Johnston, W. A. (2019). Redefining science: The impact of complexity on theory development in social and behavioral research. *Perspectives on Psychological Science, 14*(4), 672–690. https://doi.org/10.1177/1745691619848688

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight decades of measurement in depression. *Measurement: Interdisciplinary Research and Perspectives, 4*(3), 135–155. https://doi.org/10.1207/s15366359mea0403_1

Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Research, 4*(2), 1–12. https://doi.org/10.31234/osf.io/p6e9c

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128–142. https://doi.org/10.3758/s13423-017-1230-y

Serra-Garcia, M., & Gneezy, U. (2021, May). Nonreplicable publications are cited more than replicable ones. *Science Advances, 7*(21), Article eabd1705. https://doi.org/10.1126/sciadv.abd1705

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., & Ullrich, J. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012, October 14). A 21-word solution. *SSRN*, https://doi.org/10.2139/ssrn.2160588

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

Smaldino, P. E. (2020). How to translate a verbal theory into a formal model. *Social Psychology, 51*(4), 207–218. https://doi.org/10.1027/1864-9335/a000425

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., ten Teije, A., & Balliet, D. (2020). *The cooperation databank*. PsyArXiv. https://doi.org/10.31234/osf.io/rveh3

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science, 16*(2), 180–200. https://doi.org/10.1287/orsc.1040.0107

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138–161. https://doi.org/10.1037/0033-2909.134.1.138

Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2020). *Status drives how we cite: Evidence from thousands of authors*. arXiv. https://arxiv.org/abs/2002.10033

Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. *Journal of Personality, 87*(3), 715–730. https://doi.org/10.1111/jopy.12428

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

The PLoS Medicine Editors. (2006). The impact factor game. *PLoS Medicine, 3*(6), Article e291. https://doi.org/10.1371/journal.pmed.0030291

Thorne, F. C. (1977). The citation index: Another case of spurious validity. *Journal of Clinical Psychology, 33*(4), 1157–1161. https://doi.org/10.1002/1097-4679(197710)33:4<1157::AID-JCLP2270330453>3.0.CO;2-B

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature, 495*(7442), 426–429. https://doi.org/10.1038/495426a

van Rooij, I., & Baggio, G. (2020). *Theory before the test: How to build high-verisimilitude explanatory theories in psychological science*. PsyArXiv. https://doi.org/10.31234/osf.io/7qbpr

van Rooij, I., & Blokpoel, M. (2020). *Formalizing verbal theories: A tutorial by dialogue*. PsyArXiv. https://doi.org/10.31234/osf.io/r2zqy

Wacker, J. (2017). Increasing the reproducibility of science through close cooperation and forking path analysis. *Frontiers in Psychology, 8*, Article 1332. https://doi.org/10.3389/fpsyg.2017.01332

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

West, R., Godinho, C. A., Bohlen, L. C., Carey, R. N., Hastings, J., Lefevre, C. E., & Michie, S. (2019). Development of a formal system for representing behaviour-change theories. *Nature Human Behaviour, 3*(5), 526–536. https://doi.org/10.1038/s41562-019-0561-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine, 5*(10), Article e201. https://doi.org/10.1371/journal.pmed.0050201

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.