Comment

PERSONALITY Science

Personality Science
Volume 5: 1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/27000710241257413
journals.sagepub.com/home/ppp
Sage

On the role of variation in measures, the worth of underpowered studies, and the need for tolerance among researchers: Some more reflections on Leising et al. from a methodological, statistical, and social-psychological perspective

Steffen Zitzmann[1,2], Wolfgang Wagner[1], Rosa Lavelle-Hill[3], Alexander J. Jung[1], Hayley Jach[1], Lukas Loreth[4], Christoph Lindner[5], Fabian T. C. Schmidt[5], Peter A. Edelsbrunner[6], Christoph D. Schaefer[7], Robert Deutschländer[8], Stefan K. Schauber[9], Georg Krammer[10], Fabian Wolff[11], Bronson Hui[12], Christian Fischer[1], Lisa Bardach[1], Benjamin Nagengast[1,13] and Martin Hecht[14]

Abstract

We point out potential drawbacks of some of Leising et al.'s (2022a) proposed ways how personality science can be improved. We argue that it is ill-advised to use only one measure for a concept. Also, we argue that researchers should not refrain from conducting a study when a high level of statistical power is precluded. Then, we go one step further and formulate additional ideas of how to improve research. Specifically, we argue that it is a good thing to use different methods rather than only one when attempting to generalize across these methods. Moreover, we argue for a more theory-driven strategy for specifying factor analytic models, and we emphasize that high-quality research is often interdisciplinary. Finally, we point to a particular risk associated with any formal reward system.

[1]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany
[2]Department of Psychology, Medical School Hamburg, Hamburg, Germany
[3]Department of Psychology, University of Copenhagen, Copenhagen, Denmark
[4]Department of Psychology, Kiel University, Kiel, Germany
[5]Faculty of Education, University of Hamburg, Hamburg, Germany
[6]Department of Humanities, Social and Political Sciences, ETH Zürich, Zürich, Switzerland
[7]Faculty of Business Management and Social Sciences, University of Applied Sciences Osnabrück, Osnabrück, Germany
[8]Independent Researcher
[9]Department of Behavioural Medicine, University of Oslo, Oslo, Norway
[10]Institute for Practice Teaching and Practice Research, University College of Teacher Education Styria, Graz, Austria
[11]Department of Psychology, University of Koblenz, Koblenz, Germany
[12]School of Languages, Literatures, and Cultures, University of Maryland, College Park, MD, USA
[13]Department of Education, Brain & Motivation Research Institute, Korea University, Seoul, South Korea
[14]Department of Psychology, Helmut Schmidt University, Hamburg, Germany

Corresponding author:
Steffen Zitzmann, Medical School Hamburg, MSH Medical School Hamburg, Am Kaiserkai 1, Hamburg 20457, Germany.
Email: steffen.zitzmann@medicalschool-hamburg.de

Keywords

methodological pluralism, consensus measure, tolerance, power analysis, Bayes

Received 18 October 2023; accepted 6 May 2024

Scholars have pointed out that research conducted in psychology suffers from questionable research practices of individual researchers as well as structural problems that reinforce the use of these practices (Asendorpf et al., 2013), which has led to calls for improvement (e.g., Nosek et al., 2012). Recently, Leising et al. (2022a) suggested ten ways to improve personality science, which may also serve as a blueprint for other subdisciplines in psychology. We truly appreciate their contribution and largely agree with the authors' main goal to improve scientific quality standards and, thereby, to foster good research, as well as with almost all of the authors' recommendations. For example, we too consider the replicability of findings to be a quality indicator and transparency in terms of detailed reporting and code sharing to be a necessary precondition for successful replication attempts by other researchers. However, we hesitate to offer our unreserved support on all presented arguments because we feel that some points deserve more discussion and reflection. Trying to optimize current practices that might be potentially flawed is only useful and warranted when the means taken to this end are themselves advantageous—a prerequisite that might not hold for some of the proposed steps as well as for the resulting proposed reward system.

Our basic motivation for writing this article is the assumption that consensus regarding scientific knowledge is an important aim that should not be confused with consensus regarding the use of methods. Rather than rewarding the use of consensus methods, we should reward attempts to approach research questions from many different angles with different methods. In the following, we will argue that it is generally ill-advised to use only one consensus measure per concept as the reduction of a (potentially) broad spectrum of different methods to a single method violates essential principles in science, such as the pluralism of methods. Moreover, we will argue that it is unwise to refrain from conducting a study when its design seemingly precludes a low type II error (i.e., a sufficiently high power). After presenting additional ideas of how to improve research, we will express concern about any formal system that is proposed as a means to reward researchers.

Why standardization can hamper science

Leising et al. (2022a) argued that the use of variants of an existing measure or completely different measures to assess a concept would increase the risk of running into jingle-jangle fallacies. A jingle fallacy describes the misguided assumption that two measures with the same label assess the same concept. In contrast, a jangle fallacy appears when two similar measures with different labels are mistakenly assumed to assess different concepts. Leising et al. (2022a) suggested that personality science should collectively develop a single standard measure per concept (i.e., a consensus measure), thereby ruling out the possibility that a diverging result is due to the use of a different measure. We acknowledge the many challenges associated with the existence of multiple measures for the same concept, and we have faced these challenges in our own research. The situation becomes even more difficult when different constructs exist for a concept or even different definitions of the concept. We agree with Leising et al. (2022a) that a measure needs to be tested carefully and improved gradually and that this task can be solved more easily if many research groups work together. At the same time, we ask ourselves whether moving towards the use of a consensus measure would lead to significant advancements of a field.

For example, we notice that one key assumption made by the authors is that science essentially represents cumulative knowledge building, a view of science that misses important aspects of scientific progress. As Kuhn (1970) pointed out, science proceeds in multiple phases. One of these phases is dominated by a common understanding or consensus (i.e., a paradigm), whereas another phase results in a shift of paradigm. Thus, according to Kuhn's (1970) view, scientific progress is not solely cumulative. Leising et al. (2022a) seem to have referred mainly to the paradigmatic phase in which researchers are involved in addressing open questions posed by the predominant paradigm. Of course, using consensus measures can help in this phase. However, as with all measures, a consensus measure has only a limited capacity to make new discoveries. In order for a paradigm to shift, there must be a critical mass of new, surprising and potentially contradicting findings that cannot be explained by the current paradigm. It should be noted that Leising et al. (2022a) acknowledged deviation from a consensus measure only for the means of advancing the measurement itself. In this light, using not just one but many different measures for a single concept can help detect such inconsistent findings and thus for opening up for a paradigm shift. See Hogan et al. (2022) for a similar but not identical argument. Hogan et al. (2022) criticized that Leising et al. (2022a) understanding of high-quality research would focus first and foremost on work coming from the "context of verfication", whilst future progress would lie in the "context of discovery".

A prominent example that is often used for illustrative purposes stems from the early days of optics (Hacking, 1983). In order to better understand the phenomenon of light, researchers used many different approaches for studying it, among which was also Bartholin's use of calcite (i.e., a transparent and colorless crystal). If you were to place a calcite crystal on this page, you would see this writing twice. Technically speaking, beside the ordinary light beam, this tool allows observers to view a second "extraordinary" beam. For Huygens, this phenomenon was a massive challenge. He wrote that he "was in a sense compelled to make this inquiry, because the refractions in this crystal seemed to overthrow my foregoing explanation of regular refraction." As a consequence, Huygens had to refine his theory by assuming and incorporating the rotational elliptical propagation of light. The use of calcite and many other measures led to findings that could not be explained by the dominant theories at that time and finally resulted in a new understanding of light as partially having wave characteristics and thus to an advancement in physics. Further examples can be found in psychological science. For instance, due to the weak internal structure of the (postulated) unidimensional Brief Self-Control Scale (BSCS; Tangney et al., 2004)—one of the most widely used questionnaires for measuring general self-control—various researchers used item subsets of the original scale to measure two-dimensional concepts of self-control (for an overview, see Lindner et al., 2015). Despite exhibiting a poor fit, the two dimensions of de Ridder et al. (2011) proved to be powerful predictors of relevant outcomes. Thus, departing from the intended original BSCS was beneficial for the development of new concepts differentiating between inhibitory and initiatory self-control (de Boer et al., 2011; Nilsen et al., 2020).

Even if a simple questionnaire is used, it may flexibly be adapted. Because personality is typically measured by asking questions about behavioral dispositions, resulting questions are naturally only relevant given certain assumptions. To illustrate this, the first item measuring extraversion as part of the FFM (DeYoung et al., 2007) reads "I am the life of the party", and persons indicate how much they agree with this statement. This makes the assumption that people go to parties, and they must have at least one, preferably more, relevant reference points of a party in their memory to determine whether they were "the life of it." In this example, reference points of parties might be something that new parents or older people find hard to conjure up and might answer "neither agree nor disagree" despite perhaps actually being extraverted. This example raises the question of whether we need to calibrate scales for different ages and life stages, as has been done in measures of personality disorders (Barendse & Thissen, 2006; van Alphen et al., 2006). The same applies across different cultures. For example, the UPPS-P scale for measuring impulsivity (Cyders et al., 2007) measures sensation seeking with the item "I would enjoy the sensation of skiing very fast down a high mountain slope." People in different cultures and countries might have never seen snow, and thus find it difficult to imagine what the sensation of skiing would be like. The point here is that it is a hard, perhaps impossible task to measure personality using a single measure that suits people across all ages, life stages, and cultures. To better suit the targeted group, the questionnaire may thus be adapted; that is, the items may be reformulated, some items may be replaced, new items may be added, or a completely different, nonoverlapping set of items may be used in a study (see Horstmann & Ziegler, 2022, who made the case for adjusting the wording of items to fit a certain language level). Notably, this does not mean that both (the original and the adapted) measures capture all aspects of a concept equally well.

Sometimes, it might even be a good strategy to choose a measure that emphasizes one aspect more than another when this aspect is more central to the study. If the goal is to predict behavior within a certain domain or context, a more specialized scale or measure might be preferable to a broader instrument (bandwidth-fidelity dilemma; Cronbach & Gleser, 1957; Salgado, 2017). For example, researchers predicting behavior in real world consumer data (rather than in a laboratory setting) may be more inclined to use a "shopping impulsivity" scale as opposed to a "catch all" impulsive disposition measure. Another example stems from self-concept research. When comparison effects are the target of investigation (e.g., in research on the internal/external frame of reference model), researchers should be aware that these effects are usually stronger when the corresponding comparison processes appear in the item formulations (e.g., "I am better at math than my classmates;" see Wolff et al., 2021). Only if the field uses different measures, will we gain insights into which flavor of concept predicts or does not predict different dependent variables in different populations and contexts. This heterogeneity enables the progression in our understanding of hard to measure and sometimes hard to define concepts, such as personality traits. Of course, such a freedom in the choice of measure comes at the price of less standardization, but it also allows measures to be more closely tailored to the requirements of a particular study (see Ziegler, 2014). Admittedly, this does not necessarily contradict Leising et al. (2022a) demand for consensus measures because such measures may be developed for different application contexts.

In a similar vein, we think that using different constructs for the same concept can be helpful. In this article, we adopt the view that constructs should be distinguished from concepts. According to this view, a construct is only a proxy that lies between the concept and its indicators (see Rigdon, 2012; see also Uher, 2021, for a deeper discussion). A construct is created to enable operationalization and validation (e.g., by developing a nomological net, Cronbach & Meehl, 1955). In the simplest case, a somewhat different construct may be obtained by slightly modifying the existing one, for example, by allowing correlated uniquenesses, which can become necessary as a result of an imperfect translation of the items into another language (see Schmidt et al., 2017). In other contexts, a different structure or even a completely different type of construct may be needed (see Lu et al., 2023). Constructs can be formed in different ways, one of which is using factor-based methods. These methods use a common factor to explain the correlations between a construct's indicators. Another, sometimes even more suitable way to form constructs are composite-based methods, which combine the indicators into a composite (Hair et al., 2021). Although factor-based methods are more popular in psychology, composite-based methods can be superior to factor-based methods not only from a theoretical point of view but in practice. An example of such composite-based methods is the partial least squares method, which Wold (1982) and also others recommended because models with structural relations between constructs that are formed by this and similar methods exhibit better small-sample properties (Rigdon et al., 2017; Tenenhaus et al., 2005; Zitzmann & Helm, 2021).

Finally, we should acknowledge and appreciate that there may also be different definitions of a concept, each coming with particular strengths as well. To give an example, in social psychological science, the concept of tolerance has been defined as involving liking others' beliefs, preferences, and practices, or regarding them as something good. However, according to a recent definition by Bernd Simon, tolerance is defined as the attitude that one accepts others' beliefs, preferences, and practices despite one's disapproval of them (Simon, 2020; Simon et al., 2019). Importantly, unlike the former definition of tolerance, the latter definition includes disapproval as a definitional condition (Gibson et al., 1992). A somewhat broader definition of tolerance was recently put forward by Verkuyten et al. (2022). The fact that different definitions exist testifies that tolerance allows, if not calls, for them. Each definition sheds a different light on the phenomenon and thus helps deepen our understanding of it (see Zitzmann, Loreth, et al., 2022; see also Fernandes & Aharoni, 2022, who emphasized that conceptual differences are legitimate). It does not hamper science as long as researchers are aware of the differences between these definitions and interpret findings in strict accordance with the specific definition used—a strategy that can also help reduce the risk of jingle-jangle fallacies. One way to achieve this is address these fallacies openly as Schmidt et al. (2018) and Keller et al. (2016) did. For example, Keller et al. (2016) found that the enthusiasm that teachers perceive while teaching and the enthusiasm that teachers display were both termed teaching enthusiasm. Ever since, researchers have clearly stated which of both concepts they used.

Leising et al. (2022a) most compelling argument for why consensus measures and possibly also consensus constructs and definitions would be needed in psychology is that efforts to engage in cumulative knowledge building, including meta-analysis, would otherwise be "difficult or even futile" (p. 9). However, such differences are not problematic for meta-analyses, at least when an adequate model is chosen. In meta-analytic research, the mixed-effects model has become the gold standard for two decades now. In this model, effect sizes are expressed in terms of true effect sizes and deviations from these true effect sizes. In addition, the true effect sizes in turn are expressed as an average effect size plus study-specific deviations from this average effect size. How much the deviations from the true effect sizes vary defines the sampling error, and the variance of the deviations from the average effect size describes how the study-specific true effect sizes vary around the average effect size. The latter variability is often referred to as "study heterogeneity" (e.g., Overton, 1998). Because a certain amount of this heterogeneity may be due to differences in measure for a concept, the differences are inherently accounted for by the model. Hence, statistical tools for accounting for the variation in measures are readily available, and thus, from a statistical point of view, there is no need for reducing the spectrum of different measures to a single one. In addition, the use of multiple measures in combination with random effects allows for generalized conclusions regarding effect sizes, whereas using only a single measure would tie conclusions to this measure. As Yarkoni (2022) pointed out, researchers are usually interested in general conclusions, and thus, they agree that multiple samples need to be drawn in order to generalize findings. If Leising et al. (2022a) idea of a consensus measure was adopted to samples, this would mean that a single (consensus) group be studied. It is very clear that this would not allow findings to generalize beyond this group.

Overall, we agree that a consensus measure can help address open questions posed by a paradigm. However, we think that disregarding other measures for a concept may produce artificially more consistent findings. This means that a consensus measure might prevent the emergence of inconsistent findings, and the predominant paradigm persists for longer (because inconsistencies accumulate more slowly). As a consequence, there will be less pressure to innovate and potentially less scientific progress in terms of paradigm shifts. While some researchers may opt for a specific measure, construct, and definition of the concept in their studies, it is important to acknowledge that there are many other (potential) measures, constructs, and definitions for the same concept, which represents something valuable rather than a threat to good research.

We fear that if we were to use only consensus measures, consensus constructs, and consensus definitions, this could be seen as a step back to Popperian thinking. Admittedly, a great deal of research in psychology is devoted to Popper's doctrines, and Leising et al. (2022a) even referred to these ideas in their explanation of good research. However, these ideas have long been superseded by other ideas, such as those by Paul Feyerabend but also by pragmatists (see Albers et al., 2018, for an example of pragmatism in psychology). Similar to us, Feyerabend (2010) argued that the prescription of only one method can even hamper science, and as Hacking (1983) put it, we should not really expect something as colorful as science to be tied to a single method. For example, indices addressing the fit of a model to the data are routinely used to justify a researcher's decision for or against the validity of the model. However, there are other methods that can help researchers assess whether a model is valid. For example, researchers can make use of theory: When X must impact Y more strongly than the other way around for theoretical reasons, a model indicating the opposite is invalid and should therefore be rejected even when this model fits the data (see Stone, 2021).

Our view is in line with Oreskes (2020), who explicitly warns about methodological fetishism and emphasizes methodological pluralism as a central component of a science. She argues that a strong *scientific* consensus will emerge only if researchers arrive to a great part at the same conclusion despite using different methods. Similarly, Zitzmann and Loreth (2021) made the case for an "almost anything goes" attitude toward methods (see also Klimstra, 2022, who even included qualitative methods). While researchers should remain open for other methods, a basic scientific framework of logic and evidence still defines the limits (see Hilbig et al., 2022).

It is interesting to note that Leising et al. (2022a) idea of a consensus has also been criticized by other commentators, such as Corker (2022), Denissen and Sijtsma (2022), Fernandes and Aharoni (2022), and Hagemann (2022). In a nutshell, their criticisms can be summarized as follows. These authors argued that any consensus measure would be biased, because it expresses what the mainstream thinks is the "right measure." Even worse, it has been argued that the choice of the consensus measure would potentially be influenced by a few powerful people, and without a means to protect the choice from being influenced too much by these persons, they would essentially determine the consensus measure (e.g., Adler, 2022; Beck et al., 2022; Fedorenko et al., 2022; Galang & Morales, 2022; McLean & Syed, 2022). In other words, the choice would not be grounded in true consensus. Also, commentators have argued that the dictate of a consensus measure would have the potential to devaluate research that does not obey it, which can negatively influence an otherwise naturally evolving science (e.g., Asendorpf & Gebauer, 2022; Hilbig et al., 2022; Klimstra, 2022). This latter argument bears some similarity with our argument, which is yet different. For example, whereas Asendorpf and Gebauer's (2022) argued from an evolutionary perspective on personality science, we adopted Kuhn's theory of paradigm and an "anarchist view" to argue against consensus, using concrete examples from physics and psychology. We believe that these views add greatly to the discussion by shedding a different light on the issue.

Why a (seemingly) underpowered study can be worth conducting

We strongly agree with Leising et al. (2022a) suggestion to always plan studies in such a way that the type II error rate will be as small (i.e., a high level of power). To ensure a high level of power, the authors suggested that power analyses should be performed in advance. However, we deliberate the extent to which researchers should categorically refrain from conducting an underpowered study, especially when constraints exist that seemingly preclude a sufficiently high level of power (e.g., a limited budget). We think that this suggestion deserves some qualification for two reasons: First, a power analysis can be biased in either direction and thus be unreliable (i.e., it can underestimate the actual power of a study). Second, even a truly underpowered study can be worth conducting, because it can still serve as input for related meta-analyses. Regarding the first point, besides other possible reasons (e.g., false assumptions about the data-generating mechanism), this bias may be due to the use of a different estimator to analyze the data although the model is the same. For example, it is well known that Bayesian estimators with a regularizing effect on estimates can be less variable (e.g., they provide smaller standard errors; Greenland, 2000; Zitzmann et al., 2021) and thus also more highly powered than the estimators typically used in power analyses tools such as G*Power (Faul et al., 2007) or *PowerUp!* (Dong & Maynard, 2013).

To illustrate the effect of choosing a Bayesian estimator on a study's power, we pick out a specific example from organizational psychological science, but we want to emphasize that downward bias of power analysis is by no means limited to this example or to this field of research. A person may be assessed by eliciting ratings from a group of others to rate these persons, for example, employees rating their team leaders' leadership skills (e.g., Croon & van Veldhoven, 2007). The assessed leadership variable can then be related to other variables, such as the employees' achievement, in order to study the relationship between leadership and employee achievement. As a study's budget is often given and limited, and the power critically depends on the sample size, researchers might wish to find the optimal numbers of team leaders and employees to maximize the power to detect the relationship of interest under the given budget (e.g., van Breukelen, 2013; Zitzmann, Wagner, et al., 2022). The dotted black line in Figure 1 shows this maximized power based on a power analysis conducted with usual software (e.g., *PowerUp!*) as a function of the size of the slope in the model. In addition, the figure shows the true power when the data are analyzed with a mildly regularized estimator (i.e., a Bayes estimator with a weak, not necessarily accurate prior distribution). As Zitzmann, Wagner, et al. (2022) argued, with such an estimator, the level of power can be increased. For example, using such an estimator may lead to an acceptably high power of .80, although the initial power analysis suggested a power of only 74%. This increase is not very large, but it indicates that studies can still be sufficiently powered even though conventional power analyses did not suggest this. Of course, researchers could decide their estimator prior to the power analysis and taylor the power analysis to this estimator, but this is a difficult task for most of them because it requires advanced statistical knowledge, especially when they use anything other than the simplest models. Although our argument is not per se an argument against Leising et al. (2022a) demand for sufficiently powered studies, it highlights an issue with the judgment of whether a particular study is sufficiently powered, rendering an a priori power analysis an unreliable indicator of high-quality research.

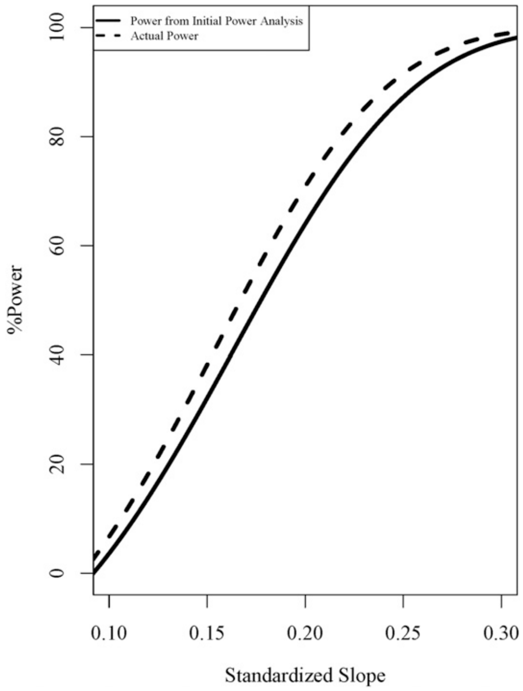

Figure 1. Power from initial power analysis versus actual power.

The second, perhaps more convincing reason why we question Leising et al. (2022a, 2022b) suggestion is that such a study can still be informative when its results are used in subsequent meta-analyses. Even when an underpowered study failed to detect an existing effect, it can still contribute by adding data to a meta-analysis and can thus help reduce uncertainty (i.e., by reducing the standard error).

An often-neglected feature of meta-analyses is that beside yes-or-no questions (e.g., whether personality affects certain outcomes), these analyses allow to investigate how effects vary across features of the study. This allows researchers to understand under which conditions effects are weaker or stronger. Hence, meta-analyses can also generate new findings. Relevant study features may include the specific measure, construct, and definition of the concept. For instance, to investigate the role of the measure more explicitly, meta-analytic models can be extended by adding a discrete variable with as many categories as there are measures as a moderator for the effect sizes using dummy coding (see Möller et al., 2020, for an example). However, besides such methodological variables, the effect size may also depend on more substantive variables, such as the studied population. In order to identify moderating variables and to quantify their role by using so-called meta-regressions, these analyses should ideally be based on a large body of studies to allow for a robust understanding of the moderations. Needless to say, to make use of underpowered studies and studies that failed to detect an existing effect, these studies need to be published together with their necessary characteristics and statistics to be included in meta-analyses, and it needs to be ensured that any form of bias in publication is minimized because otherwise, this practice may lead to overestimation of effect sizes (Nujiten et al., 2015). We would like to add that instead of running meta-analyses on (partly) underpowered studies, researchers who are faced with insufficient resources could combine their resources with those of other departments and conduct a multicenter study with sufficient power such as in clinical psychology, where such a study is conducted when one department alone cannot raise enough resources for a sufficiently powered study (see the manylabs/many babies projects or the psychological science accelerator).

Some might criticize our "defense" of underpowered studies by arguing that these studies would result in an inflated study heterogeneity and an invitation to do science poorly. Regarding the first point, it is important to note that study heterogeneity is defined as the variation of true effect sizes τ^2 (van Hippel, 2015), and thus, it cannot be affected by power. Only the sampling error σ^2 will be affected. As a consequence, whereas the estimates will be more scattered due to the larger sampling error, the study heterogeneity will remain unaltered. Alternatively, study heterogeneity may be defined as a relative quantity that compares the variation of true effect sizes to the sampling error, $\tau^2/(\tau^2 + \sigma^2)$ (i.e., the idea behind the prominent I^2 measure). However, even when study heterogeneity is defined this way will study heterogeneity not inflate if studies are underpowered. Rather, I^2 will *de*crease in this case because the denominator will be increased through σ^2. Of course, this only holds true when τ^2 is held unchanged. Note that with small studies, the statistical power of the commonly used estimate of I^2 to detect significant study heterogeneity can however be low (e.g., Huedo-Medina et al., 2006).

The second point that we would invite researchers to do science poorly is valid only when one accepts the premise that conducting underpowered studies *is* poor science, showing that this point is a circular argument.

Further suggestions to improve personality science

We have argued that variation of methods is generally a good rather than a bad thing. However, in practice, this type of variation usually occurs between studies, not within studies. As a consequence, in a given study, method effects cannot be separated from the effect of interest. However, the authors of that study still want to gain insights that are rather independent of the concrete method used, meaning that they want to generalize across methods, although this aim is hardly ever stated explicitly. Specifically, they want to generalize to the size of the effect of interest that would be obtained if all measures from the population of measures were administered to assess the outcome variable. Of course, administering all measures in one study is not practical. Alternatively, researchers could use a subset of different measures (e.g., different questionnaires) and subject the resulting data to a mixed-effects model with a random effect parameter describing the specific contributions of the measures. The effect size obtained from such a model generalizes to the study's true effect size that would be yielded if all (potentially) available measures had been employed. In other words, generalization across different measures is wanted and certainly possible even in a concrete study (see also Yarkoni, 2022, who suggested that models including random effects be used more routinely in psychology). We think that using different measures or methods in a study in combination with an appropriate statistical approach and thereby allowing for generalized insights can be considered an indicator of high-quality research.

Regarding our second suggestion, it is instructive to note that a great deal of research is concerned with studying relations between concepts. To this end, factor analytic techniques have extensively been applied, particularly in personality science. In the measurement of a concept, several items are typically used, which may in theory all be equal indicators—an assumption that we believe most researchers implicitly make. However, factor analysis tends to find different loadings for the items, because freely estimating the loadings comes with a better model fit. As factor-analysis leads to some items defining the meaning of the latent variable more than other items (i.e., they correlate stronger with the latent variable than the other items), this produces a misfit between our understanding of the concept reflected by the (equally weighted) contents of the items and the actual meaning of the latent variable in a given study (see Robitzsch & Lüdtke, 2022; see also Steger et al., 2022, for a very similar argument). When the concept does not match the latent variable, we may make faulty conclusions from the data about the concept's putative mechanisms, correlates, and theories of change. Moreover, this could also be a problem for assessment "in the real world." For example, questionnaires are frequently used to diagnose persons with clinical disorders or assess whether incarcerated persons may be at risk of recommitting a crime. In these instances, scale scores are typically used in which items are equally weighted. However, if this type of score does not match the latent variable because factor analysis found different loadings, then diagnostic decisions based on scale scores will not necessarily be backed up by factor analytic evidence of validity and predictive capacity. Thus, when studying relations between concepts and the (implicit) assumption is that all items reflect the concept equally well, a model with equal loadings should be selected rather than the locally best-fitting one. It is interesting to note that this practice is also in line with a particular reading of the classical test theory according to which all items are equally associated with the latent variable, which corresponds to equal loadings in factor analysis (Bollen, 1989; McNeish & Wolf, 2020). Admittedly, theory might prescribe a more nuanced pattern of loadings. In this case, a model with this specific pattern should be specified. As working with sound constructs is desirable, using such a more theory-driven strategy for specifying factor analytic models is good research.

Personality scientists are specialists in their discipline. However, we argue that a group of similar-minded specialists alone will not be able to answer big questions in a satisfactory manner. Significant advances happen at the borders between fields. They can only be achieved by combining the strength of many different disciplines or even different sciences. Each field has its own theories and its own methods for obtaining and interpreting data, and combining these ideas may stimulate new theories and research that provide powerful means to address questions. As an example, consider once more the work of Bernd Simon and colleagues. They have laid important foundations for future research on tolerance, benefitting greatly from political science and philosophy (e.g., Brown, 2006; Forst, 2013; Marcuse, 1970; Scanlon, 2003). Consulting other fields of research "payed dividends" in strengthening their own research. Indeed, interdisciplinary research has become increasingly central to academic interest, and Okamura (2019) found that interdisciplinarity increased impact significantly. Thus, in our view, interdisciplinarity is another important quality indictor. However, interdisciplinary research calls for a system that allows to validly assess also the quality of the contributions from other fields. Instead of trying to create a "one size fits all framework," which risks being better suited to a certain field, one suggestion is that the contributions be explicitly rated in various categories of scientific rigor. However, this would require other fields to develop own perspectives on what constitutes good research and their researchers to act as reviewers in order to assess the interdisciplinary work of others.

Why a formal reward system bears the risk of intolerance

Based on Leising et al. (2022a) ten steps to improve personality science, they proposed a formal reward system. The system is described in their article, and it is essentially an array of features that will be rewarded (e.g., a publication will get five reward points if it presents broad consensus regarding measurement practices). Although we agree that there is a dire need to improve the current system, we believe that Leising et al. (2022a) system can also be viewed as controversial, especially with regard to what the reward system would imply for researchers. Leising et al. (2022a) themselves mentioned that good research requires more time, effort, and financial resources. As a consequence, the reward system would automatically favor those who are already favored (e.g., researchers at good/established universities with financial resources and more support), thus contributing to inequalities.

Moreover, the proposed reward system points to a preference towards pre-registration and confirmatory work. With many researchers moving toward analyzing existing datasets using also methods that are exploratory by nature, the reward system may threaten to disadvantage these researchers. For example, will the use of bottom-up approaches, such as machine learning, in the analysis of passively collected big data "lose points" and perceived rigor for laying less emphasis on theoretical considerations and well-defined hypotheses? Moreover, the reward system refers to power analyses and sample size planning. When dealing with noisy large datasets and using an algorithmic modeling approach, these steps are not appropriate. Similar penalties are suggested for not providing open data access, which in principle is the ideal scenario but is not often possible in practice, particularly when working on proprietary datasets.

In other words, besides the potentially positive aspects of the reward system (e.g., improving the transparency of research), there is the risk that driven by categorizations of researchers into "good" or "bad" researchers, the system can lead to conflicts in the research community. Proponents of a reward system that penalizes researchers who, for the reasons outlined above, do not meet these criteria not only disapprove of these researchers (because they disapprove of these researcher's work), but they might also disrespect these researchers, because they do not consider them as equal fellow researchers—a clear case of intolerance among researchers. Once researchers who are met with intolerance are discouraged to publish in the same respected journals, other researcher might not become aware and influenced by these researchers' work. Whether researchers must be tolerant at all can be debated. However, we think that without tolerance, a pluralism of methods will not be possible, and without a certain degree of pluralism, science will not flourish and thus not proceed.

Other commentators also focused on the proposed reward system (e.g., Beck et al., 2022; Friedman, 2022; Schmitt, 2022). Similar to our argument, they argued that the system would disadvantage researchers who adhere to other approaches to personality (e.g., Klimstra, 2022; McLean & Syed, 2022). However, unlike these authors and based on a well-established theory of tolerance, we discussed potentially disastrous social-psychological consequences for the community and the resulting adverse effects on the progress of science.

Conclusion

This article is essentially another comment on Leising et al. (2022a). However, we presented arguments and suggestions that go beyond the 20 already published commentaries. While some of the drawbacks of Leising et al. (2022a) that we pointed out had been the subject of discussion before, our lines of reasoning differed from these discussions. For example, in our critique of the idea of a consensus measure, we referred to the history of science and illustrated our point with concrete examples. Although Hogan et al.'s (2022) criticism pointed into the same direction by emphasizing the importance of the "context of discovery," their argument felt somewhat short and remained vague. Moreover, we formulated further ideas how to improve research: generate findings that generalize across measures and other kinds of methods, specify factor models in stricter accordance with the theory, and conduct interdisciplinary research; and we argued against a reward system that is too strict.

It should be noted that Leising et al. (2022b) themselves took the opportunity to respond to the already published commentaries. In their response, the authors clarified some of their original arguments and even qualified them. For example, they qualified their argument that researchers should use consensus measures by stating that they "advocated the inclusion, not the exclusive use, of such measures" (p. 10 f.). Put differently, they suggested that a consensus measure should be included alongside other measures of the same concept. By doing this, they acknowledged not only the existence of other measures but also their added value. In a way, their appreciation of yet other measures is in line with our plea for a methodological pluralism. However, they still seem to think that a consensus measure would come with certain benefits, with the most important one being that this measure is the privileged way to separate substantively relevant from irrelevant influences on effect sizes. As we have argued, the question to which extent differences in effect sizes between studies are due to substantively irrelevant factors, such as different measures, can be addressed through meta-analytic methodology as well (but see Gollwitzer & Schwabe, 2022, who preferred replication projects). Furthermore, this methodology employs random effects, thereby allowing researchers to generalize across different measures—that is, it allows them to generalize to the effect size that would be obtained if all (potential) measures were administered (Yarkoni, 2022). While consensus measures may have some minor merits, we still doubt that these merits fully compensate for their drawbacks.

We thank Leising et al. (2022a) for taking the initiative to make psychology a better science and their inspiring ideas, which we used as a springboard to generate further discussion. In line with Leising et al. (2022a), we value cooperation and see the improvement of our science as a collaborative effort. We truly appreciate their initiative, which we view as a first proposal that needs to be evaluated continuously and improved based on the outcome of these evaluations. If we perceived the spirit of their article correctly, discussions and possible future adaptations are welcome. For example, one could debate Leising et al. (2022a) main message that the responsibility for doing science well lies with the researchers. In our view, decisive changes must also be made at other levels (Krammer & Svecnik, 2020).

To conclude, rather than recommending Leising et al. (2022a) suggestions in all respects, we wish to encourage researchers, reviewers, editors, lecturers, and appointment committees to commit themselves more to a science that is inspiring (sometimes even surprising!), moral, and grounded in creative thinking.

Key insights

- Standardization can hamper science.
- A (seemingly) underpowered study can be worth conducting.
- A formal reward system bears the risk of intolerance.

Relevance statement

In a recent article, which was published in Personality Science, Leising et al. (2022a) proposed ten ways how personality science can be improved, which may also be applicable to psychological science in general. We are a diverse group of 19 researchers with backgrounds in very different subdisciplines of psychology. What unites us is that we all remain skeptical that Leising et al. (2022a) suggestions may sufficiently improve the field. In our article, we point out potential drawbacks of some of the proposed steps, formulate additional ideas of how to improve research, and point to a particular risk associated with any formal reward system, thereby contributing to the ongoing debate.

Author note

John F. Rauthmann was the handling editor.

Acknowledgements

Not applicable.

Author contributions

Alexander Jung: Writing – original draft.
Benjamin Nagengast: Writing – original draft.
Bronson Hui: Writing – original draft.
Christian Fischer: Writing – original draft.
Christoph D. Schaefer: Writing – original draft.
Christoph Lindner: Writing – original draft.
Fabian T. C. Schmidt: Writing – original draft.
Fabian Wolff: Writing – original draft.
Georg Krammer: Writing – original draft.
Haley Jach: Writing – original draft.
Lisa Bardach: Writing – original draft.
Lukas Loreth: Writing – original draft.

Martin Hecht: Supervision, Writing – original draft, and Writing – review & editing.
Peter A. Edelsbrunner: Writing – original draft.
Robert Deutschländer: Writing – original draft.
Rosa Lavelle-Hill: Writing – original draft.
Stefan Schauber: Writing – original draft.
Steffen Zitzmann: Conceptualization, Writing – original draft, and Writing – review & editing.
Wolfgang Wagner: Writing – original draft.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Not applicable.

Data accessibility statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Supplemental material

Supplemental material for this article is available online. Depending on the article type, these usually include a Transparency Checklist, a Transparent Peer Review File, and optional materials from the authors.

Notes

Not applicable.

References

Adler, J. M. (2022). Efforts to improve personality psychology must prioritize the what, who, and why, not only the how. *Personality Science*, *3*, 30–32. https://doi.org/10.5964/ps.9227

Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 1–8. https://doi.org/10.1525/collabra.149

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Asendorpf, J. B., & Gebauer, J. E. (2022). Good intentions – unfortunate side effects. *Personality Science*, *3*(2), 5–7. https://doi.org/10.5964/ps.9227

Barendse, H. P. J., & Thissen, A. J. C. (2006). *Hetero-Anamnestische Persoonlijkheidsvragenlijst (de HAP): handleiding [Informant Personality questionnaire (the HAP): Manual]*. Barendse & Thissen.

Beck, E. D., Workman, C. I., & Christensen, A. P. (2022). CRediT where credit is due: A comment on Leising et al. *Personality Science*, *3*, 33–38. https://doi.org/10.5964/ps.9227

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Brown, W. (2006). *Regulating aversion: Tolerance in the age of identity and empire*. Princeton University Press.

Corker, K. S. (2022). There is no viable path to consensus based on the current research literature. *Personality Science*, *3*, 27–30. https://doi.org/10.5964/ps.9227

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. University of Illinois Press.

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

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Cyders, M. A., Smith, G. T., Spillane, N. S., Fischer, S., Annus, A. M., & Peterson, C. (2007). Integration of impulsivity and positive mood to predict risky behavior: Development and validation of a measure of positive urgency. *Psychological Assessment*, *19*(1), 107–118. https://doi.org/10.1037/1040-3590.19.1.107

de Boer, B. J., van Hooft, E. A. J., & Bakker, A. B. (2011). Stop and start control: A distinction within self-control. *European Journal of Personality*, *25*(5), 349–362. https://doi.org/10.1002/per.796

Denissen, J. J. A., & Sijtsma, K. (2022). A new academic incentive structure: Does it fit the psychology of human motives? *Personality Science*, *3*, 18–21. https://doi.org/10.5964/ps.9227

de Ridder, D. T. D., De Boer, B. J., Lugtig, P., Bakker, A. B., & van Hooft, E. a. J. (2011). Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control. *Personality and Individual Differences*, *50*(7), 1006–1011. https://doi.org/10.1016/j.paid.2011.01.015

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Dong, N., & Maynard, R. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on*

Educational Effectiveness, *6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fedorenko, E. J., Barnwell, P. V., & Contrada, R. J. (2022). Consensus in context: Clear disagreement can be the first step toward agreement. *Personality Science*, *3*, 41–44. https://doi.org/10.5964/ps.9227

Fernandes, S., & Aharoni, E. (2022). In unity there is strength but in divergence, unexpected leaps. *Personality Science*, *3*, 21–24. https://doi.org/10.5964/ps.9227

Feyerabend, P. (2010). *Against method*. Verso.

Forst, R. (2013). *Toleration in conflict: Past and present*. Cambridge University Press.

Friedman, H. S. (2022). Three more steps toward a better quality personality science. *Personality Science*, *3*, 44–46. https://doi.org/10.5964/ps.9227

Galang, A. J. R., & Morales, M. R. H. (2022). Consensus-finding and legitimacy: Commentary on Leising et al. *Personality Science*, *3*, 24–27. https://doi.org/10.5964/ps.9227

Gibson, J. L., Duch, R. M., & Tedin, K. L. (1992). Democratic values and the transformation of the soviet union. *The Journal of Politics*, *54*(2), 329–371. https://doi.org/10.2307/2132030

Gollwitzer, M., & Schwabe, J. (2022). Context dependency as a predictor of replicability. *Review of General Psychology*, *26*(2), 241–249. https://doi.org/10.1177/10892680211015635

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, *29*(1), 158–167. https://doi.org/10.1093/ije/29.1.158

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hagemann, D. (2022). Consensus and Diversity – A Comment on Leising et al. *Personality Science*, *3*, 7–10. https://doi.org/10.5964/ps.9227

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2021). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage.

Hilbig, B. E., Moshagen, M., & Zettler, I. (2022). Consensus is not the cure; it's part of the disease. *Personality Science*, *3*, 10–12. https://doi.org/10.5964/ps.9227

Hogan, R., Harms, P., & Sherman, R. A. (2022). Six reactions to 10 steps toward a better personality science. *Personality Science*, *3*, 13–15. https://doi.org/10.5964/ps.9227

Horstmann, K. T., & Ziegler, M. (2022). One measure to rule them all? Commentary on "ten steps toward a better personality science. *Personality Science*, *3*, 56–58. https://doi.org/10.5964/ps.9227

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Keller, M. M., Hoy, A. W., Goetz, T., & Frenzel, A. C. (2016). Teacher enthusiasm: Reviewing and redefining a complex construct. *Educational Psychology Review*, *28*(4), 743–769. https://doi.org/10.1007/s10648-015-9354-y

Klimstra, T. A. (2022). The importance of acknowledging multiple research paradigms and Diversity, Equity, and Inclusion (DEI) for improving personality science. *Personality Science*, *3*, 50–53. https://doi.org/10.5964/ps.9227

Krammer, G., & Svecnik, E. (2020). Open science als Beitrag zur Qualität in der Bildungsforschung [open science as a contribution to quality in educational research]. *Zeitschrift für Bildungsforschung*, *10*(3), 263–278. https://doi.org/10.1007/s35834-020-00286-z

Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022a). Ten steps toward a better personality science – how quality may be rewarded more in research evaluation. *Personality Science*, *3*, 1–44. https://doi.org/10.5964/ps.6029

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. (2022b). Ten steps toward a better personality science – a rejoinder to the comments. *Personality Science*, *3*, 1–15. https://doi.org/10.5964/ps.7961

Lindner, C., Nagy, G., & Retelsdorf, J. (2015). The dimensionality of the brief self-control scale—an evaluation of unidimensional and multidimensional applications. *Personality and Individual Differences*, *86*, 465–473. https://doi.org/10.1016/j.paid.2015.07.006

Lu, J. G., Benet-Martínez, V., & Wang, L. C. (2023). A socioecological-genetic framework of culture and personality: Their roots, trends, and interplay. *Annual Review of Psychology*, *74*, 363–390. https://doi.org/10.1146/annurev-psych-032420-032631

Marcuse, H. (1970). Repressive tolerance. In R. P. Wolff, B. Moore, & H. Marcuse (Eds.), *A critique of pure tolerance* (5th ed., pp. 81–123). Beacon Press.

McLean, K. C., & Syed, M. (2022). A different road towards a better personality science. *Personality Science*, *3*, 39–41. https://doi.org/10.5964/ps.9227

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research*, *90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Nilsen, F. A., Bang, H., Boe, O., Martinsen, Ø. L., Lang-Ree, O. C., & Røysamb, E. (2020). The multidimensional self-control scale (MSCS): Development and validation.

Psychological Assessment, *32*(11), 1057–1074. https://doi.org/10.1037/pas0000950

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: A Journal of the Association for Psychological Science*, *7*(6), 615–631. https://doi.org/10.1177/1745691612459058

Nuijten, M. B., Van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172–182. https://doi.org/10.1037/gpr0000034

Okamura, K. (2019). Interdisciplinarity revisited: Evidence for research impact and dynamism. *Palgrave Communications*, *5*, 141–149. https://doi.org/10.1057/s41599-019-0352-4

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

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods*, *3*(3), 354–379. https://doi.org/10.1037/1082-989X.3.3.354

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5-6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Rigdon, E. E., Sarstedt, M., & Ringle, C. M. (2017). On comparing results from CB-SEM and PLS-SEM. Five perspectives and five recommendations. *Marketing Zfp*, *39*(3), 4–16. https://doi.org/10.15358/0344-1369-2017-3-4

Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences*, *4*(9), 9–20. https://doi.org/10.1186/s42409-022-00039-w

Salgado, J. F. (2017). Bandwidth-fidelity dilemma. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–4). Springer.

Scanlon, T. M. (2003). *The difficulty of tolerance: Essays in political philosophy*. Cambridge University Press.

Schmidt, F. T. C., Fleckenstein, J., Retelsdorf, J., Eskreis-Winkler, L., & Möller, J. (2017). Measuring grit: A German validation and a domain-specific approach to grit. *European Journal of Psychological Assessment*, *35*(3), 436–447. https://doi.org/10.1027/1015-5759/a000407

Schmidt, F. T. C., Nagy, G., Fleckenstein, J., Möller, J., & Retelsdorf, J. A. N. (2018). Same same, but different? Relations between facets of conscientiousness and grit. *European Journal of Personality*, *32*(6), 705–720. https://doi.org/10.1002/per.2171

Schmitt, M. (2022). Improving research quality: The roles of the timing and scope of changes in the incentive structure and the quality of committee work. *Personality Science*, *3*(3), 16–18. https://doi.org/10.5964/ps.9227

Simon, B. (2020). A new perspective on intergroup conflict: The social psychology of politicized struggles for recognition. *Theory & Psychology*, *30*(2), 147–163. https://doi.org/10.1177/0959354319887227

Simon, B., Eschert, S., Schaefer, C. D., Reininger, K. M., Zitzmann, S., & Smith, H. J. (2019). Disapproved, but tolerated: The role of respect in outgroup tolerance. *Personality and Social Psychology Bulletin*, *45*(3), 406–415. https://doi.org/10.1177/0146167218787810

Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2022). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *Assessment*, *30*(6), 1811–1824. Advance online publication. https://doi.org/10.1177/10731911221124846

Stone, B. M. (2021). The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Frontiers in Psychology*, *12*, 783226–783234. https://doi.org/10.3389/fpsyg.2021.783226

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, *48*(1), 159–205. https://doi.org/10.1016/j.csda.2004.03.005

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical & Philosophical Psychology*, *41*(1), 58–84. https://doi.org/10.1037/teo0000176

van Alphen, S. P. J., Engelen, G. J. J. A., Kuin, Y., Hoijtink, H. J. A., & Derksen, J. J. L. (2006). A preliminary study of the diagnostic accuracy of the gerontological personality disorders scale (GPS). *International Journal of Geriatric Psychiatry*, *21*(9), 862–868. https://doi.org/10.1002/gps.1572

van Breukelen, G. J. P. (2013). Optimal experimental design with nesting of persons in organizations. *Zeitschrift für Psychologie*, *221*(3), 145–159. https://doi.org/10.1027/2151-2604/a000143

Verkuyten, M., Yogeeswaran, K., & Adelman, L. (2022). The social psychology of intergroup tolerance and intolerance. *European Review of Social Psychology*, *34*(1), 1–43. Advance online publication. https://doi.org/10.1080/10463283.2022.2091326

von, H. (2015). The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*, 1–8. https://doi.org/10.1186/s12874-015-0024-z

Wold, H. O. A. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog, & H. O. A. Wold (Eds.), *Systems under indirect observations. Causality - structure - prediction: Part II* (pp. 1–54). North-Holland.

Wolff, F., Sticca, F., Niepel, C., Götz, T., Van Damme, J., & Möller, J. (2021). The reciprocal 2I/E model: An investigation of mutual relations between achievement and self-concept levels and changes in the math and verbal domain across three countries. *Journal of Educational Psychology*, *113*(8), 1529–1549. https://doi.org/10.1037/edu0000632

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, 1–78. https://doi.org/10.1017/S0140525X20001685

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

Zitzmann, S., & Helm, C. (2021). Multilevel analysis of mediation, moderation, and nonlinear effects in small samples, using expected a posteriori estimates of factor scores. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 529–546. https://doi.org/10.1080/10705511.2020.1855076

Zitzmann, S., & Loreth, L. (2021). Regarding an "almost anything goes" attitude toward methods in psychology. *Frontiers in Psychology*, *12*, 612570–612574. https://doi.org/10.3389/fpsyg.2021.612570

Zitzmann, S., Loreth, L., Reininger, K. M., & Simon, B. (2022). Does respect foster tolerance? (Re)analyzing and synthesizing data from a large research project using meta-analytic techniques. *Personality and Social Psychology Bulletin*, *48*(6), 823–843. https://doi.org/10.1177/01461672211024422

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of Bayesian approaches in small samples: A comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 40–50. https://doi.org/10.1080/10705511.2020.1752216

Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., & Göllner, R. (2022). How many classes and students should ideally be sampled when assessing the role of classroom climate via student ratings on a limited budget? An optimal design perspective. *Educational Psychology Review*, *34*(2), 511–536. https://doi.org/10.1007/s10648-021-09635-4