General Article

Advances in Methods and
Practices in Psychological Science
2020, Vol. 3(4) 456–465
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920952393
www.psychologicalscience.org/AMPPS

Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them

Jessica Kay Flake[1] **and Eiko I. Fried**[2]
[1]Quantitative Psychology & Modelling, Department of Psychology, McGill University, and
[2]Department of Clinical Psychology, Leiden University

Abstract

In this article, we define questionable measurement practices (QMPs) as decisions researchers make that raise doubts about the validity of the measures, and ultimately the validity of study conclusions. Doubts arise for a host of reasons, including a lack of transparency, ignorance, negligence, or misrepresentation of the evidence. We describe the scope of the problem and focus on how transparency is a part of the solution. A lack of measurement transparency makes it impossible to evaluate potential threats to internal, external, statistical-conclusion, and construct validity. We demonstrate that psychology is plagued by a *measurement schmeasurement* attitude: QMPs are common, hide a stunning source of researcher degrees of freedom, and pose a serious threat to cumulative psychological science, but are largely ignored. We address these challenges by providing a set of questions that researchers and consumers of scientific research can consider to identify and avoid QMPs. Transparent answers to these measurement questions promote rigorous research, allow for thorough evaluations of a study's inferences, and are necessary for meaningful replication studies.

Keywords

construct validity, validity, psychometrics, replication, *p*-hacking

Received 1/17/19; Revision accepted 6/21/20

A foundational part of science is defining and measuring what is being studied. Before one can study the etiology of depression or the efficacy of an intervention to treat it, one must define the construct under investigation and build instruments that measure it accurately. This is difficult because psychological constructs like depression, personality traits, attitudes, or cognitive abilities often do not permit direct observation. For example, you can directly observe the height of people next to you on the bus, but you often have limited insight into their psychological processes, states, and attributes. If measures do not accurately capture depression, clinicians will deny treatment to people who need it, while prescribing medication to others who do not need it, exposing them to potentially serious side effects. Neither rigorous research design, nor advanced statistics, nor large samples can correct such false inferences.

Identifying and defining constructs is central to developing theory in psychology, because developing some initial way to measure constructs must occur before they can be studied empirically. Construct validation—collecting evidence that the instruments scientists build actually measure the constructs scientists claim they measure—is a difficult and necessary part of the research process (Cronbach & Meehl, 1955). It has taken many programs of research, thousands of studies, and decades of work to identify, define, and measure depression. And still, concerns remain as to the validity of depression as a construct and the use of instruments that measure it (Fried, 2015, 2017; Parker, 2005).

Despite the foundational role that measurement plays in the validity of study conclusions, important information regarding measurement is absent in scientific manuscripts

Corresponding Author:
Jessica Kay Flake, Department of Psychology, 2001 McGill College, 7th Floor, Montreal, Quebec H3A 1G1, Canada
E-mail: kayflake@gmail.com

in the social sciences. Measures are used without reference to their source, evidence that they actually measure the purported constructs is lacking, and unjustified measurement flexibility abounds. Such questionable practices largely go on unseen, but the few metascientific studies on measurement practices shine light on a grim state of affairs. Barry, Chaney, Piazza-Gardner, and Chavarria (2014) reported that between 40% and 93% of measures used across seven journals in educational behavior lacked validity evidence, and Weidman, Steckler, and Tracy (2017) reported that among the 356 measurement instances coded in their review of emotion research, 69% included no reference to prior research or a systematic development process. In their review of the relation between technology use and well-being, Orben and Przybylski (2019) reported that researchers pick and choose within and between questionnaires, "making the pre-specified constructs more of an accessory for publication than a guide for analyses" (p. 181).

When scientists lack validity evidence for measures, they lack the necessary information to evaluate the overall validity of a study's conclusions. Further, recent research on commonly used measures in social and personality psychology showed that measures with less published validity evidence were less likely to show strong evidence for construct validity when evaluated in new data (Hussey & Hughes, 2020; Shaw, Cloos, Luong, Elbaz, & Flake, 2020). The lack of information about measures is a critical problem that could stem from underreporting, ignorance, negligence, misrepresentation, or some combination of these factors. But regardless of why the information is missing, a lack of transparency undermines the validity of psychological science. To shine more light on these issues, we introduce and define the term *questionable measurement practice* (QMP). QMPs are decisions researchers make that raise doubts about the validity of measure use in a study, and ultimately the study's final conclusions. We demonstrate that reporting all measurement decisions transparently is a crucial first step toward improving measurement practices. We start with transparency because a lack of it prevents the evaluation of all aspects of a study's validity: its internal, external, statistical-conclusion, and construct validity (Shadish, Cook, & Campbell, 2002). We show that QMPs are ubiquitous, are largely ignored in the literature, provide researchers with ample degrees of freedom that can be exploited to obtain desired results, and thus pose a serious threat to cumulative psychological science. We address these challenges by providing a list of questions that scientists planning, preregistering, conducting, reviewing, and consuming research can use to identify QMPs, avoid a lack of transparency regarding a study's measurements, and confront egregious practices.

Questioning the Validity of Psychological Science

In the last decade, psychology built a vernacular for describing aspects of research that undermine the conclusions of a study. These include, among others, researcher degrees of freedom, *p*-hacking, hypothesizing after results are known, motivated reasoning, and fishing. These terms, often discussed under the umbrella term *questionable research practices* (QRPs), tend to span three main issues that potentially co-occur: a lack of transparency, ignorance or negligence on the part of the researcher, and the intent to misrepresent the data. The general emphasis of this discussion is that there are attributes of studies that raise questions, concerns, or grave doubts about studies' conclusions.

QRPs have been defined as practices that exploit ambiguities in what is acceptable for the purpose of obtaining a desired result (Banks et al., 2016; John, Loewenstein, & Prelec, 2012). Readers may question various design, analytic, and reporting practices because it is unclear if authors have presented biased evidence to favor a particular claim. *QRP* is a diffuse term, in that it has been used to describe overtly unethical practices carried out by researchers to mislead, but also to describe ambiguous practices. For example, there might be good reasons for excluding participants, but when reasons are not reported transparently, the practice is questionable. Such ambiguous QRPs are sometimes discussed as completely unintentional on the part of the researcher, rather than due to nefarious intent. This conflation of intent and transparency in the discussions about QRPs led some authors to urge for an exclusive focus on transparency (Banks et al., 2016; Fanelli, 2013; Sijtsma, 2016); Wigboldus and Dotsch (2016) argued that the term *questionable reporting practices* should be used in preference to *questionable research practices*.

A focus on transparency is necessary because research entails many decisions at all stages of the study pipeline, such as theory, design, data collection, analysis, and reporting. Even for the scientist with the best of intentions, these decisions can be difficult to navigate. Flexibility is inherent in the research process and exists regardless of the researcher's conscious intent to exploit it. This flexibility has been described as a "garden of forking paths" (Gelman & Loken, 2013), in which each decision takes one down a potentially different path. Critically, the decisions made along the way can lead to fundamentally different conclusions, even when researchers begin with the same raw data and have the same research question (Silberzahn et al., 2018; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Some may intentionally explore every forking path in the garden to find a significant result, whereas others may simply get lost, taking a few wrong turns.

Taken together, questionable research or reporting practices can range from accidental omission of information to fraud, and include justifiable practices, practices arising from ignorance and negligence, practices meant to mislead and misrepresent, downright lying, and many combinations of these factors. Consistent with calls for disclosure (Simmons, Nelson, & Simonsohn, 2011), preregistration (Cybulski, Mayo-Wilson, & Grant, 2016; Nosek et al., 2019), and open science more generally (Aguinis, Ramani, & Alabduljader, 2018; LeBel, Campbell, & Loving, 2017; Nosek & Bar-Anan, 2012), transparency goes a long way toward confirming or allaying doubts about the validity of study conclusions. These calls for transparency, as well as reporting guidelines (e.g., CONSORT; Schulz, Altman, Moher, & the Consort Group, 2010), prescribe the transparent reporting of measurement practices. However, questionable measurement practices persist, and a thorough examination of this problem and how to address it is lacking. Transparency is the first step, and a necessary step, toward getting answers to questions about the validity of scientific research, and the need for transparency pertains to measurement just as much as it pertains to statistical analyses.

Questionable Measurement Practices

Researchers make decisions in all phases of the research process, including when they are considering how to measure the psychological constructs they are studying. Here, we use the term *measurement* to describe any approach that researchers take to create a number to represent a variable under study. On the one hand, this definition of measurement is broad, encompassing the entire process that requires decisions, from theorizing about the nature of a given phenomenon to operationalization to analysis. On the other hand, we do not discuss cases of qualitative measurement that do not result in a number, which undoubtedly can also raise questions about validity. Issues of qualitative measurement and transparency are out of the scope of the current article, but an important area for future work.

In psychology, the possible approaches to measurement are vast, and the measures in a study introduce a stunning source of flexibility with copious decisions to navigate. We can think of no single psychological construct for which there exists exactly one rigorously validated measure that is universally accepted by the field, with no degrees of freedom regarding how to score it. Measurement in the social sciences usually comes with a garden of forking paths. For example, data collected with a 10-item questionnaire can be summarized in one sum score 1,023 different ways by summing different subsets of items (e.g., using Items 1–3 to calculate a sum score, or using Items 2–10 or Items 6–8, and so on), and even more possibilities exist if multiple subscales or different analytic techniques can be used to obtain a final score. The purpose of the current article is to focus on the need for transparency regarding the measurement in empirical studies. We argue that transparency is a necessary first step if psychological scientists are to confront ignorance, negligence, and misrepresentation of measures' use and reform measurement standards with the goal to improve psychological research.

We define *questionable measurement practices* as decisions researchers make that raise doubts about the validity of the measures used in a study, and ultimately the validity of the final conclusion. We introduce the term to raise awareness of measurement as a foundational aspect of the research process that has largely been obscured in reporting and that threatens all aspects of a study's validity, as we demonstrate in what follows. Like QRPs, QMPs can range from a lack of transparency all the way to the kinds of misleading practices the community generally regards as questionable.

QMPs Threaten All Aspects of Validity

QMPs are not just philosophical threats that researchers specialized in measurement should concern themselves with: They are ubiquitous and highly problematic. A lack of information about the measures in a study introduces uncertainty in all aspects of the study's validity. Thus, QMPs are broader than QRPs, which generally focus on distorting statistical results. Shadish et al. (2002) described four types of validity that contribute to the overall validity of a conclusion: internal validity, external validity, statistical-conclusion validity, and construct validity. Transparency is crucial because studies with more validity evidence can make stronger claims than studies with less validity evidence. Our discussion of validity is circumscribed, but we define each validity type and provide an example of how issues of opaque measurement can threaten it.

Internal validity captures the aspects of a study's design that support causal claims about the relations between variables. Measures can threaten internal validity. For example, internal validity cannot be thoroughly evaluated in the absence of information needed to determine if the measurement properties in a study differ between two treatment conditions or across time. External validity concerns the generalizability of findings across different populations and settings. A lack of transparency regarding measurement can threaten external validity when the necessary information to evaluate if the measures are sample- or population-specific is missing. Statistical-conclusion validity concerns whether conclusions from a statistical analysis are correct. This is difficult to evaluate when undisclosed measurement flexibility generates multiple comparisons in a statistical test, which can bias the test's results. We note that much of the discussion regarding QRPs focuses on statistical-conclusion validity. Finally, construct validity encompasses how the constructs in a study are operationalized and is threatened when the requisite information to determine this is missing.

Every claim from a study has some number of threats to its validity, and with this validity typology, one can categorize those threats. Detailed information about the measures in a study is needed to fully evaluate each type of validity and the study's conclusions. Scientists can agree or disagree about what constitutes the necessary evidence, what the best practices are, or what methodologies and designs are most rigorous. For example, three reviewers may have three different ideas about the most rigorous way to measure depression, but reporting measurement practices transparently enables the scientific community to vet such validity claims in the first place.

QMPs cover a whole range of issues, including lack of transparency, ignorance, negligence, and misrepresentation—all of which betray a *measurement schmeasurement* attitude. If the field took the validity of measure use more seriously, researchers would report more information, provide better training to early-career colleagues, and demand rigorous practices during the review process. Given the breadth of the problem, this article is focused on transparency and only touches on issues of ignorance, negligence, and misrepresentation. A debate about how bad current measurement practices are is not possible if the information to evaluate them is largely missing. Transparency does not automatically make science more rigorous, but it facilitates rigor by allowing thorough and accurate evaluation of the evidence. Eliminating potential questions about how researchers measured the constructs in their study is a first and necessary step to evaluating and ultimately bolstering the validity evidence.

Using Questions That Promote Validity of Measure Use

When information about the measures in a study is lacking, the information needed to evaluate the validity of the study is also lacking. In this section, we provide questions researchers should consider in order to identify QMPs and avoid them (see Table 1). These questions can be used to guide researchers in planning and preregistering their studies. Additionally, they enable reviewers, editors, and consumers of psychological research to identify QMPs and facilitate the critical evaluation of studies' validity. They do not dictate the best or correct way to create or use a measure, how to evaluate the validity of a measure, or which validity theories or psychometric models to use. In other words, this article is not a tutorial on scale development, validity theory, or psychometrics. Resources in those areas are abundant, and we have curated a list of them on OSF (Fried & Flake, 2020). Instead, the questions we discuss here are aimed at promoting validity by encouraging reporting of the information needed to evaluate how the measures in a study were selected and used.

Table 1. Six Questions to Promote Transparent Reporting of Measurement Practices

Question	Information to report
1. What is your construct?	Define the construct Describe theories and research supporting the construct
2. Why and how did you select your measure?	Justify the measure selection Report existing validity evidence
3. What measure did you use to operationalize the construct?	Describe the measure and administration procedure Match the measure to the construct
4. How did you quantify your measure?	Describe response coding and transformation Report the items or stimuli included in each score Describe the calculation of scores Describe all conducted (e.g., psychometric) analyses
5. Did you modify the scale? And if so, how and why?	Describe any modifications Indicate if modifications occurred before or after data collection Provide justification for modifications
6. Did you create a measure on the fly?	Justify why you did not use an existing measure Report all measurement details for the new measure Describe all available validity evidence; if there is no evidence, report that

How to use these questions

Here we introduce a list of questions that, when answered, can help you avoid QMPs by increasing transparency (Table 1). If these questions are unanswered, readers may doubt the validity of the measure use and final conclusions of the study. They may also suspect ignorance, negligence, or misrepresentation of the data, though identifying those causes is not the focus of these questions. Answering all the questions thoroughly does not automatically make study inferences valid, but it does give the scientific community information needed to critically vet scientific claims and support the authors in calibrating them.

In confronting the lack of transparency and validity of measurement practices, there are two levels to consider: decisions made by individual researchers while conducting individual studies and systematic questionable practices that appear again and again in the literature. In discussing these questions, we first address how individual researchers can bolster their transparency and dispel doubts about the validity of measure use in a given study. Then by presenting findings from the metascientific measurement literature, we consider how these questions, left unanswered, raise grave concerns about the validity of results in psychology. Metaresearch is often required to shine light on pervasive questionable practices. For example, a meta-analysis of clinical trials can reveal severe publication bias, indicating that many individual researchers may engage in questionable practices. In the following sections, we outline the information needed to answer the six questions, provide advice for how to obtain that information, discuss ramifications of leaving the questions unanswered, and provide examples from the metascience literature that indicate pervasive questionable practices in psychology.

1. What is your construct?

The first question to answer is, what construct are you trying to measure? What *is* it? Answering this question requires substantive expertise, involves reviewing the literature, and requires reading about the theories of the phenomenon under study. Depending on the existing literature, it may require further development of a theoretical model or paradigm. There are likely different facets of the construct of interest, and they could be largely orthogonal or have correlated dimensions.

Reporting clearly what the construct is and how it is defined in relation to the theories that support it allows readers to agree or disagree about its theoretical underpinnings. For example, there are numerous definitions of and theories about psychological constructs such as emotions, attitudes, mental disorders, and personality traits. If a target construct is not defined clearly at the outset of planning a study, the ambiguity in what is being measured will make navigating all the future measurement decisions difficult, and many opportunities to exploit this ambiguity, knowingly or otherwise, will present themselves. With no construct definition as a guide, it is easy to get lost in the garden of forking measurement paths.

After theoretically defining your construct, you need to operationalize it. This builds on the work to theoretically define the construct by specifically defining it in such a way that it is measurable. For example, Robinaugh et al. (2019) provided a formalized theory of panic disorder, which defined all variables and processes relevant to the construct of panic. Such theorizing helps researchers to operationalize the construct and processes of panic disorder and, subsequently, to select a particular measurement that is consistent with the stated theory and definition.

Transparently reporting what construct was measured and how it was operationalized goes a long way toward promoting rigor by encouraging thorough evaluation of the theory underling the construct. Reviews on a number of constructs, including giftedness (Dai & Chen, 2013), charisma (Antonakis, Bastardoz, Jacquart, & Shamir, 2016), and control (Skinner, 1996), have demonstrated that constructs are often ill defined, which leads to confusion and invalid inferences in the literature. A lack of conceptual clarity for a construct is the first step toward measurement heterogeneity, measurement flexibility, and the profusion of untested measures. For example, in their review of the literature on mental-health literacy, Mansfield, Patalay, and Humphrey (2020) identified “conceptual confusion, methodological inconsistency and a lack of measures developed and psychometrically tested” (p. 11).

2. Why and how did you select your measure?

Explaining why you selected a specific measure to assess a construct is important because there are usually multiple methods and potential instruments to choose from. For instance, there are at least 280 scales for measuring depression (Santor, Gregus, & Welch, 2006) and at least 65 different scales to measure emotions, 19 of which are devoted specifically to anger (Weidman et al., 2017). Researchers should use theoretical, empirical, and practical evidence to transparently justify their scale selection, ideally prior to data analysis.

We suggest that you answer this question by considering how your theoretical definition of the construct aligns with potential measures and the existing validity evidence for those measures. First, consider the content and face validity of an item or stimulus: Does the instrument appear, at face, to measure what you are studying? Your answer to Question 1 (“What is your construct?”)

provides necessary guidance when you select measures. Second, after identifying measures that appear to capture the construct as predefined, consider the measures' published validity evidence and if that validity evidence can extend to the current population and context. If there are prior construct-validity and instrument-development studies, they should be cited and summarized in the report. If construct-validity evidence has been gathered from a new or unpublished study, it should be described in detail. Reviewers of the report may agree or disagree about the measure's content validity, psychometric quality, or predictive validity, but citing and describing all validity evidence, previous or current, enables reviewers to do so.

Transparency regarding the (theoretical and operational) definition of the construct—as well as regarding the motivation for selecting a measure—helps prevent thinking that two instruments measure the same construct because they have similar names (i.e., the *jingle fallacy*) or assuming that two measures assess different constructs because they have different names (i.e., the *jangle fallacy*). The profusion of instruments and lack of transparency regarding them contributes to widespread jingle-jangle and makes avoiding it difficult. Transparency helps readers, reviewers, and consumers of research spot jingle-jangle, but the metascientific research focusing on it shines light on systemic measurement practices that introduce serious validity concerns.

There is clear evidence of jingle in the literature, highlighting the need to address questions of what constructs are and how they are measured at the individual-study level. Instruments and tasks that measure different content and phenomena under the same construct name are found in the fields of depression (Fried, 2017), emotion (Weidman et al., 2017), self-regulation (behavioral tasks vs. self-reports; Eisenberg et al., 2019), theory of mind (Warnell & Redcay, 2019), and fear extinction retention (Lonsdorf, Merz, & Fullana, 2019), to list just a few. The ongoing debate as to whether measures of grit (perseverance and passion for long-term goals) and conscientiousness capture the same construct, given their correlation of .84 in a recent large meta-analysis (Credé, Tynan, & Harms, 2017), is a case of potential jangle. These fallacies are not without consequence: The relationship between loneliness and extraversion, for example, was shown to be moderated by the particular scales used; correlation coefficients ranged from −.44 to −.19 (Buecker, Maes, Denissen, & Luhmann, 2020). A lack of clarity in what scales measure, despite their names, muddies the interpretation of the relationship between constructs. When such measurement questions are neglected, evaluating validity can be impossible.

3. What measure did you use to operationalize the construct?

Once constructs are defined and measures selected, the details of each measure used must be reported transparently. This requires reporting where the instrument or task comes from, the exact number of items or stimuli, the wording of items, the response format, which version was used (e.g., short or long, English or Russian), and how the instrument or task was presented to participants. For some types of measures (e.g., reaction time or neuroimaging), specification of hardware and software employed is needed.

The Hamilton Rating Scale for Depression, for instance, exists in versions with 6, 17, 21, 24, and 29 items, and has been translated into numerous languages (Bagby, Ryder, Schuller, & Marshall, 2004). Without information on what instrument was used and how it was presented to participants, the validity of the study cannot be evaluated, and conducting meaningful replications will be a struggle. Many of the questions about the details of a study's measures can be answered if authors make their materials publicly available, and some journals incentivize such transparency by awarding Open Materials badges (e.g., Eich, 2014). These details about the measures used in a study matter because differences in how measures are presented and administered introduce methodological variability that can influence results across a literature. Steinberg (1994) noted that the generalizability of personality assessment is limited by variability in administration context and order of items. Dawes (2008) found that the number of response options can systematically influence the variance of responses, confounding the results of statistical tests.

It is common practice for measurement details to go unreported in the published psychology literature. Of 433 scales reviewed from a random sample of studies published in the *Journal of Personality and Social Psychology* in 2014, 40% were reported without information regarding their source, 19% were reported without indicating the number of items, and 9% were reported without the response scale (Flake, Pek, & Hehman, 2017). This ambiguity contributes to a lack of continuity in the published literature, because studies that appear to have used the same scale could differ in item wording, response format, or number of items, which raises questions about the validity of the studies.

4. How did you quantify your measure?

Researchers usually score an instrument to produce one or more final numbers per person. There are at least three avenues that require transparent decisions: transformations of the responses, selection of which items or stimuli form which scores, and calculation of scores.

Transformations of response scales can be appropriate for various reasons. For example, reverse scoring may be appropriate for a negatively worded item, and collapsing categories may be appropriate if responding is sparse. But researchers need to disclose all such transformations and justify them in their report, so that they could be reproduced and evaluated by readers.

A next common set of questions is whether certain items or stimuli will be removed and not go into the final score, whether items or stimuli should be grouped, and if they are grouped, how this will be done. Developing a priori decision rules for removing items or stimuli and/or grouping them is ideal. Reporting those decision rules facilitates other people being able to reproduce and evaluate the work.

Finally, there are numerous ways to go from responses to scores in a data set: averaging items, calculating a standardized score, estimating a factor or component score, or grouping study participants into categories. Consumers of research need to know how scores were calculated (transparency) and why they were calculated in that manner (justification). Note that inherent flexibility in quantifying raw data is not isolated to surveys and abounds in others approaches to measurement, such as neuroimaging (Carp, 2012).

In the absence of a clear set of a priori rules (which can be documented in a preregistration), researchers might be tempted to wander down the multiverse of different forking paths presented during scoring (item average vs. factor score, two vs. three subscales), which can lead to different outcomes. But if researchers report ample evidence that supports the decisions made regarding scoring, consumers of the research can rule out threats to validity and be more confident that measurement flexibility was not exploited to obtain the final results. Examples of transparency in quantifying results are preregistering all analyses and sharing the analytic code, preferably using free open-source software, such as R (R Core Team, 2019). Examples of justifications are citing a validation study that shows empirical evidence for a certain scoring approach and conducting a psychometric evaluation from the data collected in the current study. The psychometric models that can be used to make these and related decisions are beyond the scope of this article, but we have provided a resource list including psychometric models elsewhere (see Fried & Flake, 2020).

These examples of measurement flexibility are not hypothetical scenarios that lack consequence, but are at the core of questionable practice and *p*-hacking. Simmons et al. (2011) demonstrated that when researchers have two items instead of one at their disposal, the measurement flexibility can increase the Type I error rate considerably. In analyses of clinical trials published in psychology journals, about 1 in 3 trial registrations did not contain specific information about measurement (Cybulski et al., 2016). This lack of information, including information on the quantification of measures, may therefore threaten the statistical-conclusion validity of clinical interventions. Further, reviews of preregistered protocols in clinical trials have revealed that outcome switching is pervasive (Chan, Hrobjartsson, Haahr, Gøtzsche, & Altman, 2004), and Ramagopalan et al. (2018) found that nearly a third of 89,204 registered studies had their primary outcome changed after the registration was completed. The methodological and metascience literature is clear: Measurement flexibility can be used and is being used to misrepresent and mislead.

5. Did you modify the scale? And if so, how and why?

Using an existing measure reduces researcher degrees of freedom because it is likely that many decisions, such as scoring rules, have already been made. In the best case, such decisions have been justified in previous validation studies that also support the validity of the measure and thus overall study conclusions. For example, the report on a scale-development study likely outlines the response scale and a mapping of items to subscales. However, the use of an existing measure does not eliminate flexibility and threats to validity, as it is common practice to modify measures (Flake et al., 2017; Weidman et al., 2017). Transparency regarding such modification is not only an issue for survey measures: Results from the competitive reaction time task, for instance, have been quantified in over 150 different ways across 130 publications (Elson, 2019; Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014).

Measures can be modified in a myriad of ways, before and after data collection, and these modifications challenge continuity in the published literature. Before data collection, potential degrees of freedom are (a) changing the response type (e.g., from a 7-point Likert scale to a 100-point visual analogue scale), (b) changing the response style or options (e.g., changing a Likert scale ranging from 1 to 7 to a Likert scale ranging from 1 to 5), (c) including additional items, (d) dropping items, and (e) changing item wording or content (in the case of tasks: changing stimuli, presentation time, etc.). After data collection, in addition to recoding items and dropping items, researchers can (f) adapt the scoring (e.g., change which items map to which factors, conduct single-item analyses instead of averaging items, change the type of score calculated). Again, the main point is not that such modifications are never called for, but that if they remain undeclared and unjustified, they threaten the validity of the inferences that can be drawn from a given study.

Modifications offer large sources of flexibility that introduce uncertainty about construct validity and hence interpretation of scores. The metascience literature shows evidence for questionable measurement modification. For example, in a review of 500 measures used in articles published in the *Journal of Personality and Social Psychology*, Flake et al. (2017) found that 18 measures were the result of combining two separate measures into one on the basis of the reliability coefficient α. All modifications, either before or after data collection, need to be reported transparently, justified with all evidence available to the researcher, and ideally decided on in the study planning process. These steps are necessary for the validity of the study to be evaluated comprehensively and for alternative explanations for the final result to be ruled out.

6. Did you create a measure on the fly?

Sometimes there are no scales available to study a construct of interest, or there may be good reasons to not use existing scales. When existing scales are used, many decisions have already been made, for example, regarding the response format, the number of items or stimuli, item wording or stimulus presentation, scoring instructions, and definition of subscales. Using a measure exactly as described elsewhere, and providing a citation to the original source, decreases ambiguity and the potential for flexibility.

When researchers create a new measure, they need to address the five main questions we have just detailed and justify their decisions with as much evidence as possible. Without transparent justification for how a scale was created and used, many questions regarding the validity of the study remain unanswered. Further, without guidance from the published literature as to how to answer those five questions, researchers could be tempted to wander down many paths in the measurement garden, which is unlikely to result in a valid interpretation of the score meant to measure the construct and, therefore, unlikely to lead to a valid study conclusion. Of course, the more validity evidence and reasoned justification there is for creating and using a scale, the more threats to validity readers can rule out. But, at a minimum, researchers should disclose why they created the measure, and justify why they used an ad hoc measure rather than an existing measure if such a measure exists. Further, if there is no or little validity evidence for the new measure, researchers should discuss it as a limitation of the study.

A clear pattern is emerging from the metascientific measurement literature: It is common practice to create and use measures with no evidence of systematic development. This has been documented in literature on emotions (69% of scales sampled; Weidman et al., 2017), education and behavior (40–90% of articles sampled; Barry et al., 2014), and social psychology and personality (40% of scales sampled; Flake et al., 2017). Most recently, Shaw et al. (2020) reviewed all of the measures used in the Many Labs 2 studies; 34 of the 43 (79%) item-based scales appeared to be ad hoc, having been created by the study authors and used without supporting validity information. If a study uses undeveloped measures with no validity evidence, many questions remain, and the validity of the study conclusions are cast in serious doubt.

Summary

In this article, we have defined QMPs as decisions researchers make that raise doubts about the validity of measures in a study, and ultimately the final conclusion. We have explained that QMPs can threaten all aspects of a study's validity (internal, external, statistical, and construct) and focused on transparency of measurement as a first, necessary step to improving measurement practices. A lack of transparency makes it impossible for the scientific community to identify potential threats to the validity of a study's conclusions. We have provided examples from the published literature that demonstrate the ubiquity of QMPs and shown that QMPs promote researcher degrees of freedom and threaten the validity and replicability of psychological science. We have listed a set of questions that researchers, reviewers, and readers of scientific work can consider when planning, preregistering, conducting, or consuming research.

Answering these questions transparently facilitates the rigorous evaluation of the validity of research and enables meaningful replication studies. When such transparency is absent, consumers of research are left wondering what a significant or nonsignificant effect or what a replication or nonreplication of an effect means. After all, measurement is the foundation on which all empirical science rests, and if important questions about the measurement are left unanswered, there is little that can be concluded from a study. Researchers may also find that engaging with these questions seriously throughout the research process inspires new lines of inquiry and results in research that strengthens the validity of their measures, ultimately enhancing the quality of their work.

The increased awareness of and emphasis on QRPs, such as *p*-hacking, have been an important contribution to improving psychological science. We echo those concerns, but also see a grave need for broadening scrutiny of current practices to include measurement (Fried & Flake, 2018). From the example of depression we discussed in our introduction, even if the sample size of depression trials is increased, studies are adequately powered, analytic strategies are preregistered, and *p*-hacking stops, researchers can still be left wondering if they were ever measuring depression at all.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions
J. K. Flake and E. I. Fried wrote the manuscript and contributed equally.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Open Practices
Open Data: not applicable
Open Materials: not applicable
Preregistration: not applicable

ORCID iDs

Jessica Kay Flake https://orcid.org/0000-0002-3498-615X
Eiko I. Fried https://orcid.org/0000-0001-7469-594X

Acknowledgments

The content of this article was originally developed for a workshop presented at the annual meeting of the Society for the Improvement of Psychological Science in June 2018 (Grand Rapids, MI) and during a symposium at the annual meeting of the Association for Psychological Science in May 2018 (Boston, MA). Eric Hehman, Yoel Inbar, Kim Luijken, and Donald Robinaugh provided thoughtful feedback that improved the manuscript.

References

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, *12*, 83–110. doi:10.5465/annals.2016.0011

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 293–319. doi:10.1146/annurev-orgpsych-041015-062305

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, *161*, 2163–2177. doi:10.1176/appi.ajp.161.12.2163

Banks, G. C., Boyle, E. H. O., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., . . . Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*, 5–20. doi:10.1177/0149206315619011

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, *41*, 12–18. doi:10.1177/1090198113483139

Buecker, S., Maes, M., Denissen, J. J. A., & Luhmann, M. (2020). Loneliness and the Big Five personality traits: A meta-analysis. *European Journal of Personality*, *34*, 8–28. doi:10.1002/per.2229

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6*, Article 149. doi:10.3389/fnins.2012.00149

Chan, A.-W., Hrobjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting. *Journal of the American Medical Association*, *291*, 2457–2465. doi:10.1001/jama.291.20.2457

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*, 492–511.

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

Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *Journal of Consulting and Clinical Psychology*, *84*, 753–767.

Dai, D. Y., & Chen, F. (2013). Three paradigms of gifted education: In search of conceptual clarity in research and practice. *Gifted Child Quarterly*, *57*, 151–168. doi:10.1177/0016986213490020

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*, 61–104. doi:10.1177/147078530805000106

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. doi:10.1177/0956797613512465

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, *10*(1), Article 2319. doi:10.1038/s41467-019-10301-1

Elson, M. (2019). *Competitive reaction time task*. doi:10.17605/osf.io/4g7fv

Elson, M., Mohseni, R. M., 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*, 419–432. doi:10.1037/a0035569

Fanelli, D. (2013). Redefine misconduct as distorted reporting. *Nature*, *494*, 149. doi:10.1038/494149a

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science*, *8*, 370–378. doi:10.1177/1948550617693063

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*, Article 309. doi:10.3389/fpsyg.2015.00309

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. doi:10.1016/j.jad.2016.10.019

Fried, E. I., & Flake, J. K. (2018, March). Measurement matters. *Observer*. Retrieved from https://www.psychologicalscience.org/observer/measurement-matters

Fried, E. I., & Flake, J. K. (2020). *Measurement matters*. Retrieved from https://osf.io/zrkd4/

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 166–184. doi:10.1177/2515245919882903

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*, 230–243. doi:10.1037/pspi0000049

Lonsdorf, T. B., Merz, C. J., & Fullana, M. A. (2019). Fear extinction retention: Is it what we think it is? *Biological Psychiatry*, *85*, 1074–1082. doi:10.1016/j.biopsych.2019.02.011

Mansfield, R., Patalay, P., & Humphrey, N. (2020). A systematic literature review of existing conceptualisation and measurement of mental health literacy in adolescent research: Current challenges and inconsistencies. *BMC Public Health*, *20*, Article 607. doi:10.1186/s12889-020-08734-1

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

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., . . . Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*, 815–818. doi:10.1016/j.tics.2019.07.009

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, 173–182. doi:10.1038/s41562-018-0506-1

Parker, G. (2005). Beyond major depression. *Psychological Medicine*, *35*, 467–474.

Ramagopalan, S. V, Skingsley, A. P., Handunnetthi, L., Klingel, M., Magnus, D., Pakpoor, J., & Goldacre, B. (2018). Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: A cross-sectional study. *F1000Research*, *4*, Article 80. doi:10.12688/f1000research.3784.1

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna: Austria: R Foundation for Statistical Computing.

Robinaugh, D. J., Haslbeck, J. M. B., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., . . . Borsboom, D. (2019). Advancing the network theory of mental disorders: A computational model of panic disorder. *PsyArXiv*. doi:10.31234/osf.io/km37w

Santor, D. A., Gregus, M., Welch, A. (2006). Eight decades of measurement in depression. *Measurement*, *4*, 135–155. doi:10.1207/s15366359mea0403_1

Schulz, K. F., Altman, D. G., Moher, D., & the Consort Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials*, *11*, Article 32. doi:10.1186/1745-6215-11-32

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston, MA: Cengage Learning.

Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology/Psychologie canadienne*. Advance online publication. doi:10.1037/cap0000220

Sijtsma, K. (2016). Playing with data—Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*, 1–15. doi:10.1007/s11336-015-9446-0

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Bannard, C. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356. doi:10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, *71*, 549–570. doi:10.1037//0022-3514.71.3.549

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712. doi:10.1177/1745691616658637

Steinberg, L. (1994). Contexts and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*, 341–349.

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, Article 103997. doi:10.1016/j.cognition.2019.06.009

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*, 267–295. doi:10.1037/emo0000226

Wigboldus, D. H. J., & Dotsch, R. (2016). Encourage playing with data and discourage questionable reporting practices. *Psychometrika*, *81*, 27–32. doi:10.1007/s11336-015-9445-1