CellPress
REVIEWS

Forum

Why Are Self-Report and Behavioral Measures Weakly Correlated?

Junhua Dang [1,*], Kevin M. King,[2] and Michael Inzlicht[3,4]

Accumulating evidence indicates weak correlations between self-report and behavioral measures of the same construct. We suggest that these weak correlations result from the poor reliability of many behavioral measures and the distinct response processes involved in the two measurement types. We also describe how researchers can benefit from appropriate use of these measures.

Introduction

Self-report and behavioral measures are two of the most popular methods of capturing individual differences in psychology. The same psychological construct is often assessed with both types of measures, with researchers using them interchangeably, often conflating findings across measurement type. However, across a series of domains, recent meta-analyses and large-scale investigations have consistently found that self-report and behavioral measures of the same construct were weakly correlated. For example, the average correlation between self-report and behavioral measures of self-control [1,2], emotional intelligence [3], empathy [4], risk preference [5], and creativity [6] ranged from 0 to 0.20, indicating a weak (or nonexistent) association between these two types of measures of the presumed 'same' construct.

This weak association suggests that self-report and behavioral measures might be inherently different and thus cannot be considered interchangeable indicators of a single construct. Our goal here is to: (i) explain why self-report and behavioral measures are bound to be weakly correlated by paying careful attention to the psychometric properties of these measures; and (ii) describe how researchers can benefit from appropriate use of them in research.

The Reasons for Weak Correlations

The reasons for weak correlations are both methodological and conceptual.

Reliability Paradox

In an important paper, Hedge and colleagues raised awareness of what they called the reliability paradox, which is created by the fact that the very features that make a task robust in an experimental sense make them unreliable in a psychometric sense [7]. To appreciate this, note that many behavioral measures were originally developed to produce replicable experimental effects (via within-person contrasts) and the use of behavioral measures in between-person (or correlational) studies came without sufficient psychometric scrutiny. The design of these behavioral measures inherently reduces their reliability because they maximize within-person variance at the expense of between-person variance.

$$Reliability = \frac{Variance\ between\ individuals}{Variance\ between\ individuals + Error\ variance}. \quad [1]$$

Specifically, to produce robust experimental effects, between-person variability (i.e., individual differences on a particular psychological construct) within a condition needs to be as low as possible, which means that most (if not all) people respond to the experimental manipulation. For example, the color-naming Stroop task was designed to maximize the (within-person) difference between congruent (naming the red color of the word RED) and incongruent (naming the red color of the word GREEN) trials, with the result that nearly everyone shows Stroop interference with little (between-person) variability around this interference effect.

Mathematically, variance between individuals is in the denominator in tests of mean difference between conditions such as the *t*-test and ANOVA, meaning that lower between-person variability within conditions produces larger experimental effects. However, because variance between individuals is also in the numerator in measures of reliability, as shown in Equation 1, lower between-person variability hampers the reliability of these behavioral measures:

Moreover, error variance tends to be high in behavioral measures, which can also dramatically attenuate reliability. One source of error variance is trial-by-trial variation in performance, which is especially problematic when the number of trials is limited, as is typical in most behavioral measures. Recent research suggested that the test–retest reliability of both the Stroop task and the Flanker task could be dramatically improved by increasing the number of trials and by using hierarchical models to remove trial-by-trial variation [8], but this is rarely done in practice, resulting in suboptimal test–retest reliability. Error variance in behavioral

tasks is also increased by various situational factors, such as the emotional state of the participant, the noise and illumination of the laboratory, the distance between the participant and the screen, and the presence of other people, to name but a few.

As a result, behavioral measures (typically having low reliability) are likely to be weakly correlated with self-report measures (often with high reliability) because the reliability of two measures limits the correlations that can be observed between them, with lower reliability leading to weaker observed correlations, as shown in Equation 2:

$$Sample\ correlation = 'True' correlation \sqrt{Reliability\ (x) * Reliability\ (y)}. \quad [2]$$

Divergent Response Processes

Some behavioral measures have good reliability yet are still poorly associated with self-report measures [3,4]; furthermore, correcting for low reliability does not always improve correlations between self-report and behavioral measures [1,5]. An alternative explanation, therefore, is that despite sharing the same name (Box 1), self-report and behavioral measures are distinct because they are designed to measure very different response processes.

First, behavioral measures tap responses to uncommon stimuli in a specific and highly structured situation, whereas self-report measures ask participants to reflect on their behaviors across a variety of unstructured real-life situations. Second, behavioral measures are based on performance such as reaction time and accuracy, whereas self-report measures are based on perceptions of performance, which reflects subjective judgments about performance rather than performance itself.

Third, behavioral measures tend to tap people's maximal performance because they encourage people to do their best, while self-report measures tend to tap people's typical performance about how they usually behave. This distinction is similar to the competence–performance discrepancy, in which 'competence' refers to the ability to perform activities whereas 'performance' refers to actual performance of activities, which reflects not only ability but also motivation, effort, and willingness. For example, new data show that dark personality traits (Machiavellianism, psychopathy, and narcissism) are not related to highly reliable behavioral measures of empathy, but are strongly correlated with self-report measures of empathy, suggesting that individuals with dark personalities have the ability but not the disposition to empathize [9].

Appropriate Applications in Research

Understanding these methodological and conceptual differences could lead to a better understanding of how to use these measures in research. Measures with high reliability (i.e., most self-report measures and select behavioral measures such as the working memory span task) can be used to predict individual differences in real-life outcomes, and reliable self-report and behavioral measures may explain incremental variance above each other because they are likely to assess different constructs. Recently, there has been a trend to search various biomarkers (e.g., event-related potentials, fMRI activation patterns, heart rate variability) by relating them to available individual difference measures. We note, with some caution, that attention must be paid to the reliability of these potential biomarkers because recent studies have demonstrated poor reliability of many biological measures themselves, such as measures of fMRI-based functional connectivity [10].

Measures with low reliability, resulting from low between-person variance, are not suitable for measuring individual differences. Mathematically, as shown in Equation 2, it is expected that these behavioral measures will be weakly correlated with any other measures. Not only are such tasks weakly correlated with self-report measures, they are weakly

Box 1. Low Between-Person Variability and the Jingle-Jangle Fallacy

The jingle-jangle fallacy refers to the tempting but often erroneous assumption that two measures with the same name tap the same construct (the jingle fallacy) or two measures with different names tap different constructs (the jangle fallacy). The current Forum mainly focuses on the jingle fallacy and the contribution of low between-person variability to this fallacy. Low between-person variability, however, may also contribute to the jangle fallacy. For example, the Black–White implicit association test (IAT), a behavioral task, is thought to assess implicit racial bias that is not amenable to consciousness and thus distinct from explicit racial bias. One basis of this claim is the lack of any meaningful association between the Black–White IAT and explicit measures of Black–White attitudes. Recent analyses, however, suggest that the weak correlation between the IAT and explicit racial bias results mainly from the IAT's failure to capture between-person variability [13]. This leaves an uncomfortable question: despite being thought to reflect the unique construct of implicit bias, does the IAT mostly reflect explicit bias? Regardless of whether the IAT commits the jangle fallacy, it appears clear that it is a poor measure of individual differences in attitudes (implicit or otherwise) due to low between-person variance.

Box 2. Implications for Executive Functions

Executive functions are measured by various behavioral tasks with low between-person variance [7]. Mathematically, these measures should thus be weakly correlated with each other and real-life outcomes, which should be observed in studies with sufficient statistical power where the false discovery rate is minimized. However, because many studies are conducted with small samples, observed correlations will tend to be heterogeneous. This has played out in at least two ways: confusion about the structure of executive functions and apparent associations between executive functions and real-world outcomes that disappear with high-powered samples. Regarding the structure of executive functions, although three factors (i.e., updating, inhibition, and task-switching) have been identified to represent executive functions by early research, follow-up studies failed to replicate this structure, with the number of identified factors ranging from a single factor to as many as five factors [11]. This inconsistency can be partly explained by low between-person variance that produces heterogeneous results in modeling these tasks in small-sample studies. Similarly, regarding predictive validity, early studies using underpowered designs found mixed results for the association between executive functions and real-world outcomes (e.g., dietary intake). Later studies, however, failed to reveal any underlying association once adequate samples were used [2,14].

correlated with other, supposedly related tasks. This insight may shed light on, for example, the considerable inconsistency regarding the unity and diversity of executive functions [11]. For the same reason, these tasks should also have poor predictive validity for real-life outcomes (Box 2). For this reason, we are concerned with recent changes to psychiatric practices, such as the Research Domain Criteria initiated by National Institute of Mental Health, which have put increasing weight on the diagnostic information provided by behavioral tasks. To be direct: these tasks tend to have low reliability, making them unsuitable for psychiatric diagnoses of individuals.

Despite these notable drawbacks, these measures still have utility. For example, because they are sensitive to within-person experimental manipulations, they can be important for studying the processes that underlie task performance or the contexts that enhance or detract from task performance. They may also be useful for predicting the short-term waxing and waning of an attribute for the same individual. For example, an experience sampling study found that momentary changes in performance on a Go/Nogo task predicted snack consumption in the following hour (i.e., within-person effect), although individual differences in performance on the same task did not predict snack consumption (i.e., between-person effect) [12].

Concluding Remarks

The weak correlations between self-report and behavioral measures of the presumed same construct result from the poor reliability of many behavioral measures and the distinct response processes involved in these two measurement types. We suggest that only measures with high reliability be used for individual difference research, whereas measure with low reliability may help to predict the short-term waxing and waning of an attribute for the same individual.

Acknowledgments

J.D. was supported by the Swedish Research Council (2018-06664). K.M.K. was supported by a grant from the National Institute on Drug Abuse (DA047247). M.I. was supported by grants from the Social Sciences and Humanities Research Council of Canada (#435-2019-0144) and from Canada's Natural Sciences and Engineering Research Council (RGPIN-2019-05280).

[1]Department of Neuroscience, Uppsala University, Uppsala, Sweden
[2]Department of Psychology, University of Washington, Seattle, WA, USA
[3]Department of Psychology, University of Toronto, Toronto, ON, Canada
[4]Rotman School of Management, University of Toronto, Toronto, ON, Canada

*Correspondence:
dangjunhua@gmail.com (J. Dang).
https://doi.org/10.1016/j.tics.2020.01.007

References

1. Saunders, B. *et al.* (2018) Reported self-control is not meaningfully associated with inhibition-related executive function: a Bayesian analysis. *Collabra Psychol.* 4, 39
2. Eisenberg, I.W. *et al.* (2019) Uncovering the structure of self-regulation through data-driven ontology discovery. *Nat. Commun.* 10, 2319
3. Joseph, D.L. and Newman, D.A. (2010) Emotional intelligence: an integrative meta-analysis and cascading model. *J. Appl. Psychol.* 95, 54–78
4. Murphy, B.A. and Lilienfeld, S.O. (2019) Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychol. Assess.* 31, 1062–1072
5. Frey, R. *et al.* (2017) Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* 3, e1701381
6. Park, N.K. *et al.* (2016) Revisiting individual creativity assessment: triangulation in subjective and objective assessment methods. *Creat. Res. J.* 28, 1–10
7. Hedge, C. *et al.* (2018) The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* 50, 1166–1186
8. Rouder, J.N. and Haaf, J.M. (2019) A psychometrics of individual differences in experimental tasks. *Psychon. Bull. Rev.* 26, 452–467
9. Kajonius, P.J. and Björkman, T. (2020) Individuals with dark traits have the ability but not the disposition to empathize. *Pers. Individ. Differ.* 155, 109716
10. Noble, S. *et al.* (2019) A decade of test–retest reliability of functional connectivity: a systematic review and meta-analysis. *Neuroimage* 203, 116157
11. Karr, J.E. *et al.* (2018) The unity and diversity of executive functions: a systematic review and re-analysis of latent variable studies. *Psychol. Bull.* 144, 1147–1185
12. Powell, D.J.H. *et al.* (2017) Does real time variability in inhibitory control drive snacking behavior? An intensive longitudinal study. *Health Psychol.* 36, 356–364
13. Schimmack, U. (2019) The implicit association test: a method in search of a construct. *Perspect. Psychol. Sci.* Published online October 24, 2019. https://doi.org/10.1177/1745691619863798
14. Egbert, A.H. *et al.* (2019) Executive function and dietary intake in youth: a systematic review of the literature. *Appetite* 139, 197–212