Journal of Business Research 193 (2025) 115352

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Calibration experiments: An alternative to multi-method approaches for measurement validation in consumer research

Dominik R. Bach [a,b], Edward E. Rigdon [c], Marko Sarstedt [d,e,*]

[a] University of Bonn, Transdisciplinary Research Area Life & Health, Centre for Artificial Intelligence and Neuroscience, Bonn, Germany
[b] Department of Imaging Neuroscience, UCL Queen Square Institute of Neurology, University College London, London, UK
[c] Georgia State University, Robinson College of Business, Atlanta, Georgia, USA
[d] Ludwig-Maximilians-University Munich, LMU Munich School of Management, Munich, Germany
[e] Babeș-Bolyai-University, Faculty of Economics and Business Administration, Cluj-Napoca, Romania

ARTICLE INFO

Keywords:
Calibration
Measurement
Metrology
Psychometrics
Uncertainty
Validity

ABSTRACT

Measurement validation in consumer research is ideally performed within the context of a multi-trait multi-method matrix (MTMM). While statistically well developed, this approach has several shortcomings that limit its domain of application: (1) the requirement for sufficiently unrelated latent variables that can be measured with the same methods, (2) the requirement for conceptually different methods to disambiguate trait from methods, and most seriously (3) the difficulty in identifying a more valid over a less valid method. We compare the MTMM approach to experiment-based calibration, an alternative framework for validating those latent variables that can be externally manipulated. We show how calibration lets researchers make distinctions between even closely related measurement methods, dispenses with the need for unrelated latent variables, and enables optimization of the measurement evaluation procedure itself. Calibration can be an important part of an integrative validity argument in consumer research and, more broadly, across the social sciences.

1. Introduction

Most theories in consumer research are formulated in terms of latent, not directly observable, variables, such as brand image, customer satisfaction, or corporate reputation. Validation is the process of objectively evaluating the quality of measurements of these latent variables and usually rests on multiple sources including quantitative metrics (Kane, 2016; Messick, 1987). The validity of measurements addresses perhaps the most fundamental confound to scientific inference—without valid measurements, valid inference is impossible. Over the past seven decades, psychometrics has developed a canonical approach for measurement validation, embodied in a multi-trait multi-method (MTMM) matrix. This approach features prominently in applied research, including in studies published in *Journal of Business Research* (e.g., Coote, 2011; Czakon et al., 2023; Mishra, 2000; Ong et al., 2015; Suoniemi et al., 2021). Core aspects of the MTMM matrix approach go back to Cronbach and Meehl (1955) and Campbell and Fiske (1959). Convergent validity is the degree to which an observed variable, used as an indicator for a latent variable, has strong correlations with other observed variables purportedly associated with the same latent variable. Discriminant validity is the degree to which such an observed variable has weaker correlations with observed variables purportedly associated with a different latent variable. Convergent and discriminant validity can be embodied in an MTMM matrix when several latent variables are quantified using the same set of multiple methods (Franke et al., 2021). Fig. 1 illustrates an MTMM matrix for the latent traits of choice confidence and political extremism. Choice confidence is relevant in the context of consumer research (Olsson, 2014), and political extremism might be the sort of potentially unrelated trait that might be used to establish discriminant validity for indicators of choice confidence.

While the literature on statistical evaluation of MTMM matrices is fairly well developed (Eid & Nussbeck, 2009; Helm, 2022; Höfling et al., 2009; Oort, 2009; Widaman, 1985), its practical application is not trivial (Zumbo & Chan, 2014), and the methodological inferences that can be drawn are not always as precise as would be desirable (Fiske, 1982; Kenny, 2021). Furthermore, MTMM assessment focuses on the study of individual differences, and its application remains challenging when stable between-person variance is low (Bach, 2023b). Indeed, the emphasis on MTMM matrix assessment in applied research may be partly driven by the absence of available alternatives.

* Corresponding author.
E-mail addresses: d.bach@uni-bonn.de (D.R. Bach), erigdon@gsu.edu (E.E. Rigdon), sarstedt@lmu.de (M. Sarstedt).

https://doi.org/10.1016/j.jbusres.2025.115352
Received 7 December 2024; Received in revised form 10 March 2025; Accepted 30 March 2025
Available online 8 April 2025

This paper describes another approach to measurement validation, one based on experimental calibration. Experiment-based calibration can be an effective approach for indicators of latent variables that change over time and that can be externally manipulated, as is commonplace in consumer research as well as in many fields of psychology. We illustrate the distinct advantages of the MTMM matrix approach and the calibration framework by drawing on the example of choice-confidence measurement, for which different measurement methods support divergent models of the role of overconfidence in financial decisions (Moore, 2022; Moore & Healy, 2008; Olsson, 2014).

2. MTMM-based validity assessment

Spearman (1904a, 1904b) launched fundamental innovations that continue to dominate quantitative social science. Spearman argued that very high correlations among disparate tests associated with a wide range of abilities proved the existence of a single, unitary "intelligence" or "intellectual energy" factor. Spearman's work thus gave rise to the notion of "convergent validity": different tests of intelligence should correlate with each other.

Later scholars rejected the theoretical notion of one general intelligence factor in favor of different facets of intelligence, such as word understanding or logical reasoning (e.g., Thurstone, 1935). With that, it was no longer desirable for all tests to correlate strongly. Two tests associated with two conceptually unrelated latent variables should not correlate highly, thus demonstrating "discriminant validity" (Cronbach & Meehl, 1955). If discriminant validity is not established, "researchers cannot be certain results confirming hypothesized structural paths are real or whether they are a result of statistical discrepancies" (Farrell, 2010, p. 324; see also Radomir & Moisescu, 2019). A prominent approach that relies on the MTMM matrix for discriminant validity assessment is Henseler et al.'s (2015) HTMT metric, which is computed as the ratio of average heterotrait-heteromethod and monotrait-heteromethod correlations (Franke & Sarstedt, 2019).[1] In its original presentation, the HTMT metric relied on the geometric mean of the average monotrait-heteromethod correlations, but variants have been proposed that apply different calculation rules (e.g., Ringle et al., 2023).

Further complicating the assessment of test validity, test outcomes or scores can be the result not only of substantive latent variables but also of aspects of the tests themselves (Campbell & Fiske, 1959). The language used in paper-and-pencil tests, for example, can profoundly affect scores on many tests of entirely independent latent variables. Such effects of test features (i.e., language in verbal tests) became known as method effects or common methods bias. Thus, if two independent traits are assessed with similar instruments and the resulting measurements are highly correlated, a method effect is present (Campbell & Fiske, 1959). This complicates the interpretation of convergent validity scores. In order to distinguish the impact of latent variable from method effect,

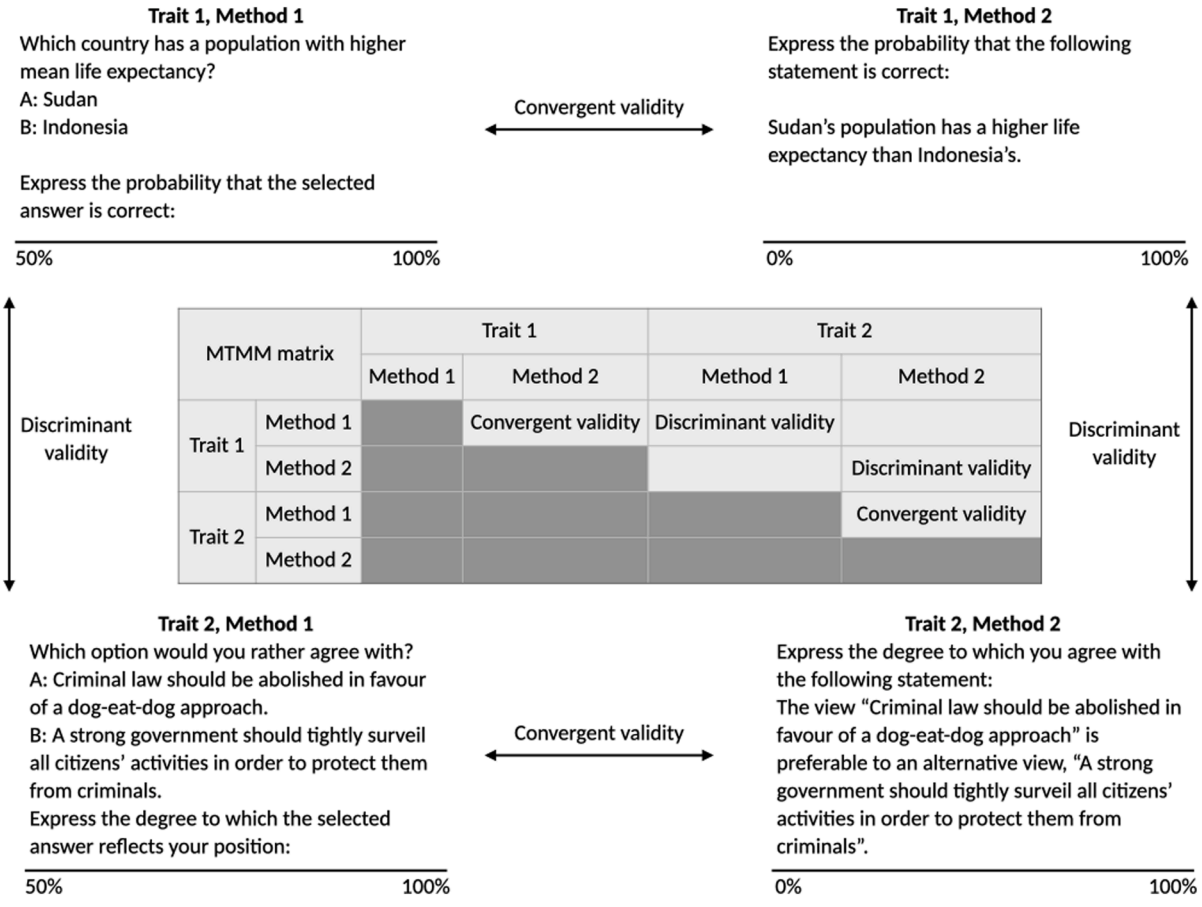


Fig. 1. Example of an MTMM matrix assessment. Notes: Two traits are assessed with two methods; the target variable (Trait 1), taken from Olsson (2014), is "choice confidence" and the unrelated trait required for establishing discriminant validity (Trait 2) is "political extremism." Both of these traits are considered latent variables that can be measured with different methods. The MTMM matrix is the correlation matrix between the 2 × 2 measurements. Dark grey fields are ignored. The figure notes the fields that are associated with convergent validity and with discriminant validity.

[1] For a review of discriminant validity assessment techniques, see Rönkkö and Cho (2022).

"…*more than one trait as well as more than one method* must be employed [italics in original]" (Campbell & Fiske, 1959, p. 81) within the same validation study. Campbell and Fiske's (1959) MTMM approach relied on "eyeballing" a correlation matrix—simply counting the number of correlations that met or failed to meet a criterion, with no inferential basis.

The advent of confirmatory factor analysis created the opportunity for a more formal approach. For example, Jöreskog (1971) suggested comparing the fit of models that either constrain, or do not constrain, correlation between common factors, as a way of testing for discriminant validity of the indicators of those common factors. Fornell and Larcker's (1981) proposal, by contrast, sought to extract information on convergent and discriminant validity from factor model parameter estimates. Their somewhat simplistic approach has been highly popular—Google Scholar records almost *130,000 citations* for this single paper as of April 2025.

Later, more elaborate techniques stipulated complex factor structures such that the model closely and precisely corresponded with the relationships believed to underlie the data, including representations of both traits and methods as common factors. Such an approach can yield incremental goodness of fit tests for convergent and discriminant validity, within the constraints of the common factor model (Widaman, 1985).

More recently, multilevel modeling approaches have been applied to MTMM data, taking numerical outcomes across multiple latent variables and using multiple methods to be clustered by respondent (e.g., Eid et al., 2008). Still, such models are somewhat rare, perhaps partly because they are inclined to problems involving nonconvergence and improper solutions (Kline, 2011) and limitations in model complexity (Maas et al., 2009).

3. Complications of the MTMM matrix approach

While MTMM-type validity assessments are commonly applied in consumer research and across the social sciences (Kenny, 2021), this approach comes with several challenges. These relate to the identification of (1) a sufficiently unrelated latent variable that can be assessed with the same method, (2) unrelated methods to independently capture the same latent variable, and (3) determining which method better captures the latent variable under consideration. To illustrate these problems, consider Olsson's (2014) *Journal of Business Research* study of methodological diversity in the measurement of (over)confidence—see also Pillai (2014). We selected this particular example for two reasons. First, choice confidence is regarded as a stable trait with sufficient interindividual variability to enable MTMM matrix measurement validity assessments—but at the same time, it also depends on the type of choice being made, and thus it can be manipulated externally. Second, the measurement of choice confidence poses a substantive problem that we believe has not typically been highlighted in methodological MTMM discussions: different measurement methods, even though purporting to measure the same latent variable and generally converging, support different prevalence of overconfidence in the population, with implications for models of financial trading (e.g., Deaves et al., 2009). This highlights the importance of determining which of two convergent measurement methods is the "more valid" one.

Olsson (2014) presented two measures of confidence: the half-range method and the full-range method. In the half-range method, participants answer a binary knowledge question ("Which country has a population with higher mean life expectancy? A: Sudan, B: Indonesia"; Fig. 1) and also report confidence in their choice on a scale from 50 % (no basis to prefer one answer or the other) to 100 % (entirely confident in their choice). In the full-range method, participants indicate their answer to the same knowledge question on a scale from 0 % (certainly option A) to 100 % (certainly option B); the distance from the opposite (unfavored) answer constitutes the confidence metric (i.e., it can range from 50 % to 100 %). On the face of it, the two methods appear rather similar (i.e., both can be said to have content validity), and there is no a priori reason to favor one method over the other.

The first challenge in validating such instruments with the MTMM approach is to find an unrelated latent variable that is conceptually sufficiently different from choice confidence, but that can be measured with the same method. Both methods in the example link a factual statement with a confidence statement. It is at least challenging to disentangle these two aspects and combine the response format with a different question that is unrelated to confidence. The suggestion we make here is meant to illustrate these difficulties. One example of an unrelated latent variable might be "political extremism" in a political attitude survey (Pecot et al., 2021). Thus, one might ask persons whether they agree more with two conflicting and politically extreme statements A or B and then how strongly they agree with the chosen answer, on the full or half-range scale. Of course, the assumption that political extremism is unrelated to the propensity to be confident about one's judgements might be incorrect. This illustrates the general problem in finding suitable unrelated latent variables. Indeed, in their survey of validation practice, Zumbo and Chan (2014, p. 322) observed a "high frequency of convergent evidence […] but relatively low inclusion rates of discriminant evidence."

The second challenge is to find unrelated measurement methods for the focal latent variable; that is, confidence. In our example, the half-range method and the full range method are relatively similar rather than being entirely unrelated. Both require a verbal understanding of the factual statement, both use a visual analogue scale, and both require a fair amount of motivation and sustained attention. This illustrates a general difficulty with the use of the MTMM approach: the methods represented in an MTMM matrix should be distinct, but in practice they are often very similar. The need to use several methods to measure the same latent variable is also a serious practical constraint for all fields in which the range of established methods is limited, when measurement is time-consuming or expensive, or when validating the first method for a novel latent variable.

The third difficulty is that if two measurement methods show convergent validity, the MTMM approach is not designed to tell which of the two is "more valid." It can tell us which of the methods correlates more strongly with an extracted common factor—but because the common factor depends on the set of methods included, it is not the same as the unobserved latent variable (Rigdon et al., 2019).

In the following sections, we present an approach that dispenses with the need for an unrelated latent variable and additional measurement methods. Its results can also give a clear answer as to whether and which observed variable is more closely related to the latent variable in question.

4. Metrology and calibration as the standard approach in physical sciences

Up to here, we have conformed to the standard psychometric perspective on "measurement" as prevalent in consumer research, and generally in the social sciences. According to this perspective, "measurement" means the act of establishing an empirical basis for an unobserved variable, such as obtaining a questionnaire response. This questionnaire response is then transformed into an estimate of the unknown latent variable (e.g., by estimation of a factor model or by forming a sum score). By contrast, "measurement" in metrology—measurement science in the physical sciences—means obtaining a quantitative value for an unknown attribute of an object, such as its mass. This process may already include a data transformation or inference procedure, such as inferring the mass of an object from the observed compression of a spring (Estler, 1999). This difference complicates the comparison of "measurement" concepts across this disciplinary divide, which researchers have started addressing only recently (e.g., Rigdon et al., 2019, 2020, 2023). In the following, we seek to bridge this gap and apply selected concepts from metrology to the social

sciences.

In metrology, researchers routinely compare the performance of their instrument to that of a reference that is well-established—a practice known as calibration (Phillips et al., 2001). This reference may itself be calibrated to another reference, and so forth, creating a chain of calibrations which connects to a so-called measurement standard. In the past, such standards were often based on prototypes such as the prototype kilogram (a specific block of metal) or prototype meter (a specific metal stick) stored in secured vaults under controlled conditions. Such material standards are obviously not usable in the social sciences (Krantz et al., 1971), but they are also not mandatory for calibration. In fact, metrology in recent years has moved away from material standards and toward experiment-based calibration. Since 2019, all of the seven main international standard or "SI" units used by the physical sciences are calibrated against the outcomes of carefully designed experiments that prescribe particular desired values from substantive theory. This approach resonates with many areas of experimental research in which external manipulations are used to change the value of a latent variable.

5. Using calibration experiments to assess measurement validity in consumer research

Our proposed experimentalist complement to the MTMM approach exploits the idea that validity can be assessed by measurement of criteria that are believed to be closely related to the latent variable in question; for example, "tests measuring related […] constructs" or "criteria that the test is expected to predict" (AERA et al., 2014p. 16). Here, "constructs" refers to underlying latent traits. However, the calibration approach is broadly applicable to (multi-item) common factors, single observed variables, and other types of measurement (e.g., weighted composites) that researchers use to measure certain traits (e.g., Hair et al., 2024a, 2024b; Sarstedt et al., 2024). In addition, our proposal is based on the idea that observations can be taken in different temporal orders. Researchers may refer to "concurrent validity" when the two observations are taken at the same time, or "predictive validity" if a measurement of a latent variable is used to predict the criterion at a later time point (Cronbach & Meehl, 1955). In addition—crucially, for our argument—the criterion could also be measured at an earlier point in time; that is, could be an antecedent of the latent variable in question (Messick, 1987, p. 89). For latent variables that are subject to external manipulation, an effective experimental treatment would itself be an antecedent of the latent variable and could be used as a criterion to evaluate validity, as has long been suggested for educational measurement (Messick, 1987, p. 89): a valid test of mathematical skills should distinguish between students who did an intensive mathematical training, and those that did not. That is, the predicted effect of the experimental treatment (the independent variable) ought to correlate with a measurement (of the manipulated trait) that is taken at a later time point. Bach and Melinscak (2020) have referred to this correlation as "retrodictive validity," and this approach is essentially equivalent to experiment-based calibration in the physical sciences (Bach et al., 2020). The main difference is that psychological latent variables, at the present state of knowledge, have no natural scale. Hence, while calibration in the physical sciences would numerically compare individual measured values to individual values based on the accepted standard, calibration in the social sciences focuses on the correlation between "standard" values, which ought to be the outcomes of an experimental manipulation, and the actually measured values (Fig. 2).

To implement experiment-based calibration, researchers must seek an experimental procedure which impacts the measurand (the true value of the latent variable); for example, "confidence" in our running example (Fig. 2). At this stage, some scientific consensus is needed on how the treatment affects the latent variable. This could be an ordinal theory (e.g., a certain educational program improves literacy from "low" to "high"), or a numerical theory (e.g., doubling physical luminance of a certain colored object is expected to change the perceived lightness from

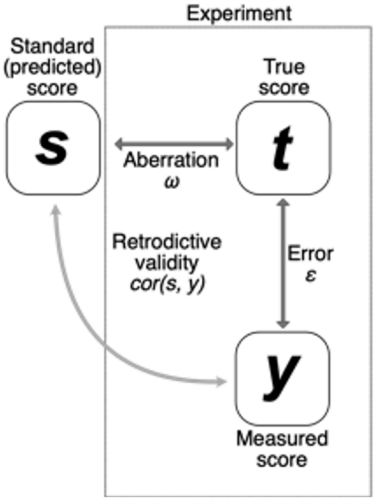


Fig. 2. Calibration experiment. Notes: The double-headed lines denote differences; the double-headed curve denotes a correlation. An experiment (grey box) is performed to impact the latent variable in question. The independent variable (not shown) is designed to achieve predicted standard values s, which are determined a priori. The difference between predicted and actually achieved (unknown) true scores is denoted experimental aberration. Several measurements are taken to quantify the value of the latent variable (only one shown in the figure). The correlation of this measured score with the predicted standard score is termed "retrodictive validity" and ranks different measurement methods by their accuracy.

value J_1 to value J_2). Based on this theory, researchers can specify predicted values for the change in the manipulated latent variable. These predicted or "standard" scores s are the values which the latent variable is expected to take on if the treatment performs as expected and if the actual value of the latent variable could be known. They are established with knowledge of the experimental design and the motivating theory, but before the experiment is conducted and thus without knowledge of the scores actually obtained, and so are exogenous to the outcome.

While prior knowledge and theory lead to predicted standard scores, the actual execution of the experimental procedure will generate unknowable true scores t of the latent variable, for example, true confidence scores (Fig. 2). Note that our definition of the true score is based on metrological convention and differs from classical true score theory. True score theory would define a "true score" in relation to a particular measurement method, as the expected value of measured scores across respondents or across hypothetical repeated samplings (Nunnally & Bernstein, 1994). In our usage, and consistent with metrology practice as well as other psychometric approaches, true scores are defined here as consequence of an experimental manipulation, independent of whether and how that consequence is measured (Borsboom, 2005; Haig & Evers, 2015). We define experimental aberration ω as the difference between standard scores, predicted on the grounds of prior knowledge and theory, and the unknown true value of the measurand. Like true scores themselves, experimental aberration is unobserved. Neither standard nor true scores depend on the measurement method, rendering experimental aberration independent of the measurement method and beyond the reach of any "method effects."

Measurements are then designed to capture the (unknown) true scores, but the results of measurements are measured scores y (Fig. 2). Measurements utilizing different methods will likely result in different measured scores. The measured scores that are obtained will be affected by the unknown true scores to some degree, but will deviate from those true scores due to systematic shortcomings of the specific measurement method as well as due to random variation. This difference is measurement error ε, which will likely vary across methods.

Obviously, the validity question could be settled if the researcher could calculate the correlation between each set of measured scores and the true scores—but true scores are unavailable. However, the researcher *can* calculate the correlation between measured scores and predicted standard scores. The correlation between each set of measured scores and the corresponding set of predicted standard scores (i.e., retrodictive validity) depends on both experimental aberration and measurement error. Because experimental aberration (the difference between expected standard scores and true scores) is the *same* for all methods within a particular study, any difference in retrodictive validity between the measurement methods can *only* be due to differences in measurement error. Under the fairly general assumption that aberration and measurement error are not linearly related, it can be shown that the method with higher retrodictive validity—a higher correlation between standard score and measured score—minimizes overall measurement error variance, which is the sum of systematic and random components (Bach et al., 2020). Thus, the experimental manipulation and the exogenous standard scores provide a means to quantitatively determine the relative measurement error, and thus the relative validity, of the different measurement methods employed in a study. Retrodictive validity does not distinguish between random error variance and systematic error variance, but a variety of techniques are available for assessing the reliability of instruments, which reflects their random error variance.

The formal basis for experiment-based calibration evolved in the study of associative learning using psychophysiological indices. However, a number of calibration studies have been published in the wider field of experimental psychology (Table A1 in the Appendix), encompassing associative learning, emotion, and cognitive processes. In a typical calibration example, Xia et al. (2023) sought to compare validity of different measurement methods for a specific implementation of associative learning—that is, humans' learning to predict an aversive outcome (unconditioned stimulus, US) after a contingently preceding neutral cue (conditioned stimulus, CS), which plays a central role in explaining consumer behavior (e.g., Girard et al., 2019). Associative learning strength can be measured by the magnitude of a conditioned response (CR) when encountering the CS. Following (ordinal) learning theory, Xia et al. (2023) assumed that coupling a "CS+" with an aversive US would lead to a higher degree of associative memory than coupling another "CS-" with no US. Thus, they defined two standard values, "high" and "low," dummy-coded as 1 and 0. In an actual learning experiment, in which participants were exposed to sequences of CS/US couplings, they recorded five types of psychophysiological observables. For each of these, they formed an estimate of associative learning for the CS-/US relation, and for the CS+/US relation, which yielded two values per participant. These were correlated with the standard values (0 for CS-/US and 1 for CS+/US), yielding retrodictive validity values. For three of the five observables, retrodictive validity was not appreciably different from zero, and thus they did not appear to measure anything that is related to the standard values. For the remaining two, retrodictive validity, computed as $r_{SY} = \frac{d}{\sqrt{d^2+2}}$, where d is the average within-subject difference, divided by its standard deviation, were as follows:

Method 1, based on fear-potentiated startle: $r_{sy} = 0.46$
Method 2, based on pupil dilation: $r_{sy} = 0.60$

In summary, pupil dilation yielded the highest retrodictive validity, thus establishing that it was more closely related to true associative learning than any of the other measurement methods implemented.

Going beyond this formal approach of retrodictive validity, there is a plethora of published work using similar reasoning. For example, in applied educational measurement, there is a common argument going back to Messick (1987) that test scores should be higher after than before training, and that this establishes validity. Table A1 in the Appendix includes some example applications of this argument.

6. Simulated comparison of MTMM and calibration

In practice, because the MTMM approach and the calibration approach both constitute large-scale efforts, researchers will usually opt for one or the other, and not both. Furthermore, there are probably many latent variables that are not amenable to either calibration, because they defy external manipulation, or MTMM assessment, because there is no stable between-person variance (Bach, 2023b), or because of a lack of either convergent methods or unrelated latent variables. Recognizing this complication, here we illustrate the advantages and shortcomings of both methods with a simulated comparison, based on our initial example of confidence measurement (Fig. 1). Drawing on a large body of literature, we can assume that our example latent variable, "confidence," is amenable to external manipulation (i.e., it depends on the particular decision to be taken) and has relatively stable between-person variance when external factors are held constant. Thus, both validation approaches can in principle be used. Following Deer et al. (2025), we made our simulation code and results available on the Open Science Foundation platform at https://osf.io/2a6qp/.

First, we use the aforementioned paper-and-pencil procedure to establish four measurements: confidence with both the full-range procedure (Y_{11}) and the half-range procedure (Y_{12}), and political extremism with both the full-range and the half-range procedures (Y_{21}, Y_{22}). Because these variables have no natural scale, we assume all true scores to be mean-centered and to have unit variance. The simulation also added independently and identically distributed (iid) measurement noise and a method effect to the true scores, to simulate the observed values. Furthermore, we assume that the measurement of confidence relates to the true score non-linearly, and that this non-linearity is less pronounced for the full-range method than for the half-range method (i.e., the full-range method is more truthful than the half-range method). Table 1 shows the resulting MTMM matrix.

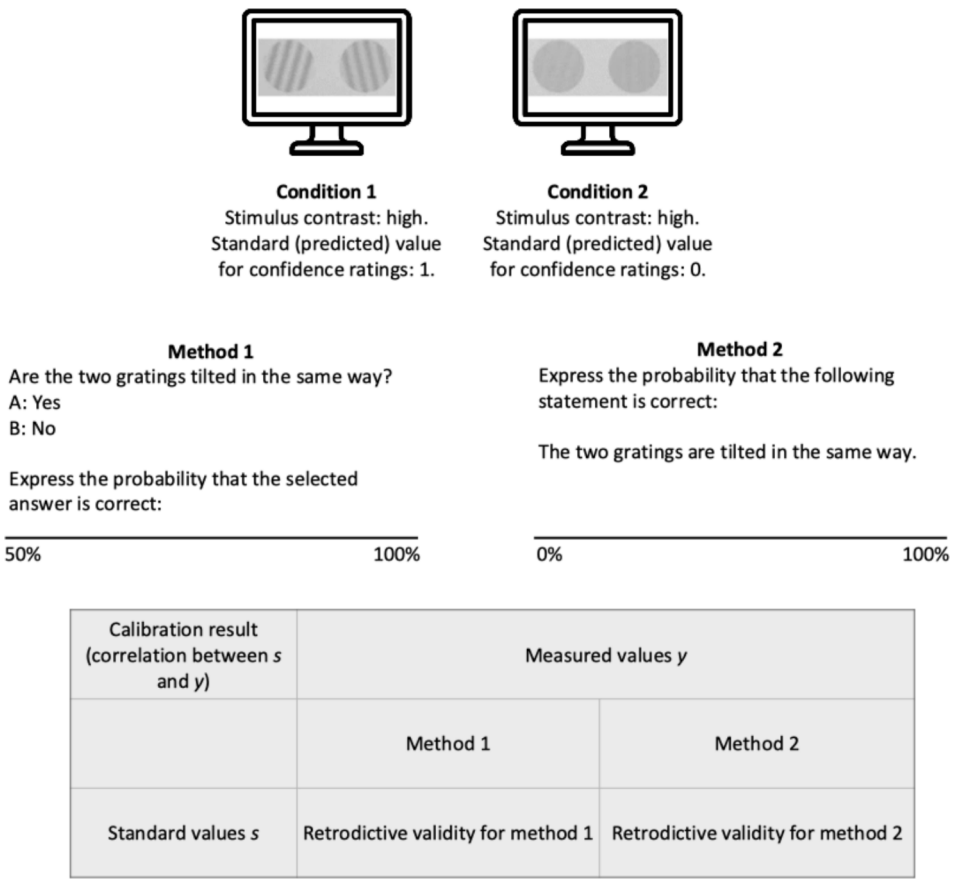
The simulation produces a pattern of correlations among the observed variables which would be consistent with a positive evaluation of measurement validity. The simulated results indicate that, as methods for measuring confidence, the two methods have a substantial degree of both convergent validity and discriminant validity. Notably, reliability assessment (diagonal entries in Table 1) yields similar results for both methods, because the observation noise is identical in our example, and the higher simulated inaccuracy for the half-range methods stems from a non-linearity.

Next, we simulate a calibration assessment (Fig. 3). Doing so requires an experimental procedure that uncontroversially impacts the latent variable in question. In our confidence example, we know from much previous work that stimulus properties in perceptual decisions impact confidence (Vickers, 1979), and a variety of methods for manipulating

Table 1
Numerical example for the proposed approach.

MTMM assessment		Confidence		Political extremism	
		Full-range	Half-range	Full-range	Half-range
Confidence	Full-range	0.958	0.877	0.064	0.014
	Half-range		0.999	−0.040	−0.004
Political extremism	Full-range			0.653	0.647
	Half-range				0.634
Calibration assessment	**Retrodictive validity**	**0.473**	**0.270**		

Notes: Upper part: classical MTMM assessment shows high convergent validity between the methods, high discriminant validity, and a small degree of methods variance. Diagonal entries show test–retest reliability. Lower part: retrodictive validity (correlation with experimental criterion) is appreciably higher for full-range than half-range method. In our simulations, this is due to a non-linearity in the measurement which is more pronounced for the half-range method. Simulation code available at https://osf.io/2a6qp/.



Calibration result (correlation between *s* and *y*)	Measured values *y*	
	Method 1	Method 2
Standard values *s*	Retrodictive validity for method 1	Retrodictive validity for method 2

Fig. 3. Example of a calibration experiment for the assessment of choice confidence.

stimulus properties are available in the psychophysics literature. Manipulating stimulus properties, then, constitutes an experimental procedure that impacts trial-by-trial decision confidence. A common perceptual decision-making task is to ask participants whether two sets of parallel lines ("visual gratings"), presented next to each other on a screen for a short interval, have the same or different orientation. For relatively similar orientations, this task is easier when the brightness contrast between lines and background is greater (e.g., black lines on a white background renders the task easier than dark grey lines on a light grey background). Whenever the task is easier, participants are likely to express more confidence in their decision (Bang et al., 2019). Thus, we can use brightness contrast to manipulate confidence. The experimental procedure defines the predicted standard values for the measurand (which might just be on two levels: "high confidence" and "low confidence," dummy-coded as 0 and 1). These standard values serve as an exogenous validity criterion.

Now the experiment is performed with many trials per participant in four conditions. On each trial, the participant is shown two high or low contrast gratings. They are asked whether each pair of gratings have the same orientation and how confident they are about this, either on the half-range scale or on the full-range scale. Confidence values from each condition and participant are then averaged. For each of the two measurement methods, high- and low-contrast confidence values from all participants are then correlated with the predicted standard scores.

The resulting retrodictive validity values are shown in the bottom row of Table 1. Both methods have substantial correlations with the standard scores, indicating some degree of validity. This resonates with the results of the MTMM matrix assessment. However, retrodictive validity substantially differs between the methods, and is appreciably higher for the (theoretically more accurate) full-range method. Notably, the numerical values of the retrodictive validity coefficient have no direct interpretation in terms of measurement error alone because they depend on both measurement error and experimental aberration, and the latter is unrelated to measurement method. The values of the retrodictive validity coefficient are meaningful only in a categorical sense (i.e. that they differ from zero establishes validity for both methods), and importantly, in relation to each other. This latter type of ranking was not possible in the MTMM approach. Crucially, it also did not require the assessment of political extremism as a control variable.

7. Advantages of the calibration experiment approach

It is easy to see partial parallels between experiment-based calibration and the MTMM matrix approach, but the calibration approach solves several of the difficulties associated with the classical technique, while opening the way to address additional goals.

7.1. Removing common methods bias

First and foremost, the calibration approach better manages the potential confound of common method bias. In calibration, the experimental treatment plays the role of one of the methods. Common method bias across the experimental treatment and the instruments being tested seems less likely than it does across different instruments being administered to respondents, often simultaneously. In the calibration approach, common methods variance across competing measurement *instruments* may blur distinctions between the alternative instruments, but it cannot make the weaker measure seem like the stronger one, as can happen when convergent validity is assessed in the conventional MTMM procedure.

7.2. Dispensing with unrelated latent variables

Second, and as a consequence, the assessment of discriminant validity becomes obsolete. Assessing discriminant validity is primarily driven by concern about common method covariance. Calibration is not subject to this limitation. While method variance can appear to increase convergent validity metrics, it will reduce retrodictive validity (covariance with the experimental treatment). This disadvantages any measurement instrument affected by construct-irrelevant variance in a direct comparison of several methods. Hence, an explicit assessment of method variance is not necessary for the interpretation of retrodictive validity metrics.

The problem of common method variance in the MTMM matrix approach corresponds to the problem of experimental confounds in experiment-based calibration. If an experimental manipulation is not entirely selective, then one might erroneously end up with a measurement method that measures the confound, rather than measuring the latent variable in question (Bach, 2023a). This, however, is a classic problem in experimental design (Lipsitch et al., 2010). Assessing and controlling for confounds is relevant for any substantive experimental research, whereas finding unrelated latent variables and discriminant methods might arguably be more relevant for psychometric assessment only. We would therefore venture to suggest that in experimental research, solutions to control confounds are more developed than solutions to the unrelated latent variable / unrelated method problem.

In cases where confounds are not sufficiently controlled within one experimental manipulation, a common solution is the use of negative controls: experimental manipulations that are supposed to *not* have the desired effect. Experimental research is rife with examples of this sort (often compiled into series of experiments, some of which are supposed to show an effect and others the absence of an effect). In the framework of experiment-based calibration, one could then perform negative-control calibration and favor methods that show high retrodictive validity in the main calibration experiment and low retrodictive validity (i.e., little covariation with the confound) in the negative-control experiment. In our example, one might speculate that stimulus contrast—or, generally, decision difficulty—impacts not only confidence but also perceived social expectations as a confound (Nascimento & Loureiro, 2024): respondents might feel that they are expected to demonstrate lower confidence for more difficult decisions, even if their actual level of confidence is unchanged by the experimental treatment. One might then execute a negative control experiment in which social expectations are manipulated in a different way; for example, expressing confidence ratings alone in a room with assurance that they are only going to be seen by the computer in one treatment, versus expressing confidence ratings with several researchers watching, in the other. If a measurement instrument measures only decision confidence and not social expectations, then it would show low retrodictive validity for the social desirability latent variable in this experiment.

7.3. Direct comparison of convergent methods

Third, calibration provides a quantitative metric, retrodictive validity, to compare even closely related measurement methods. This is a crucial advantage over the classical MTMM matrix approach. In our example, we wanted to know whether the half-range method or the full-range method for measuring choice confidence is more appropriate, because they support diverging theories. The MTMM matrix approach finds no answer to this question, but the calibration approach does: if measured scores from one method have a closer relation to the standard values, then retrodictive validity will be higher. This closer relation might stem from more precise measurement (i.e., less random variation across repeated measurement of the same true score) and/or from more truthful measurement (i.e., closer association of averaged measurements with the true scores). To distinguish between these two scenarios, one might complement experiment-based calibration with reliability assessment, which addresses measurement precision on its own.

Because the calibration approach ranks methods according to their retrodictive validity, it affords the possibility of incremental and iterative improvement in measurement method. Such incremental development has so far been restricted to improvements in reliability, where a clear ranking of methods is possible. However, improved reliability might come at the cost of lower validity. The calibration approach opens a means to make this crucial distinction and to enable long-term strategies to improve measurement practice. In this vein, experiment-based calibration can also be used for combining several related measurement methods (e.g., different questionnaire items) into a weighted score in order to derive an even more accurate estimate of the value of a trait. In this strategy, the optimal weights can be normatively derived by the criterion of maximizing retrodictive validity (Mancinelli et al., 2024).

8. Discussion, future directions, and conclusion

The calibration approach offers several potential advantages over the MTMM matrix approach. It dispenses with the need for unrelated latent variables and enables optimization of the psychometric procedures themselves. Calibration can thus complement classical psychometrics for those latent variables that are amenable to external manipulation. The calibration approach could become part of an integrative validity argument in consumer research.

While calibration appears conceptually plausible for many latent variables that can be externally manipulated, the prevalence of studies assessing such latent variables differs across the various branches of consumer and marketing research and, more broadly, in psychology—see Table 2 for an overview. We suggest that experiment-based calibration is applicable widely in the areas of experimental and intervention research across these fields. To facilitate the choice between MTMM and calibration assessment, we summarize our main arguments in a decision tree shown in Fig. 4.

Despite its usefulness, calibration is not free from limitations. Most notably, the calibration approach requires that the latent variable is amenable to experimental manipulation, thus excluding stable enduring traits. Secondly, there ought to be some form of agreement in the field about a suitable experimental procedure (for an example, see Bach et al., 2023). This limits the approach to a subset of all latent variables (see Table 2)—but for those it provides a significant advantage.

A second limitation is that in the current form, retrodictive validity does not distinguish systematic measurement error (which relates to a common conceptualization of validity) from random error (which relates to reliability). Thus, we see the calibration approach as a complement to reliability assessments.

Finally, calibration assumes that experimental aberration and error are uncorrelated. This is often plausible because the experimental treatment and the measurement method will use conceptually different and independent procedures. However, there might be cases where the systematic components of each are negatively correlated. In this case, a method's higher retrodictive validity can be due to a higher negative correlation of error and aberration (Bach et al., 2020; Bach, 2023a). This could happen if predicted scores of the experiment are misspecified, and the measurement instrument is also misspecified in an inverse manner. This serves to remind us that the process of improving theory and measurement is broadly iterative, not only in the calibration approach (Cote & Buckley, 1988), and it resonates with recent proposals that any measurement should be grounded in substantive theory, rather than relying entirely upon a particular measurement instrument (Borgstede & Eggert, 2023).

To summarize, the standard approach for evaluating the quality of measurements in the physical sciences involves calibration. For a long time, this has not received particular attention in the social sciences, possibly because psychometrics is largely developed in the context of stable traits, which are not amenable to short-term experimental manipulation. In many disciplines of behavioral science, however, it is

Table 2
Potential areas of application in consumer research and psychology.

Field	Example latent latent variables	MTMM possible	Calibration possible	Comparison
Experimental consumer research	Attention, emotion, learning, memory, motivation, stimulus percepts	Rarely performed in experimental fields	Published examples exist in psychology	Calibration preferable due to frequent lack of stable between-participant variance (Bach, 2023b)
Individual differences & personality	Attitudes, intelligence, personality	Frequently performed	Only for latent variables that are malleable by external manipulation	MTMM often preferable
Neuromarketing	Arousal, stress response	For stable latent variables (e.g., effects of brain lesions)	For latent variables that are defined as responses to external events (e.g., arousal and stress responses)	Depending on stability of latent variable
Social marketing	Social influence, social percept	For stable latent variables	For latent variables that are malleable by external manipulation	Calibration preferable
Educational psychology	Performance, skills	For stable latent variables, frequently performed	For most performance latent variables, published examples	Calibration complementary, potentially preferable

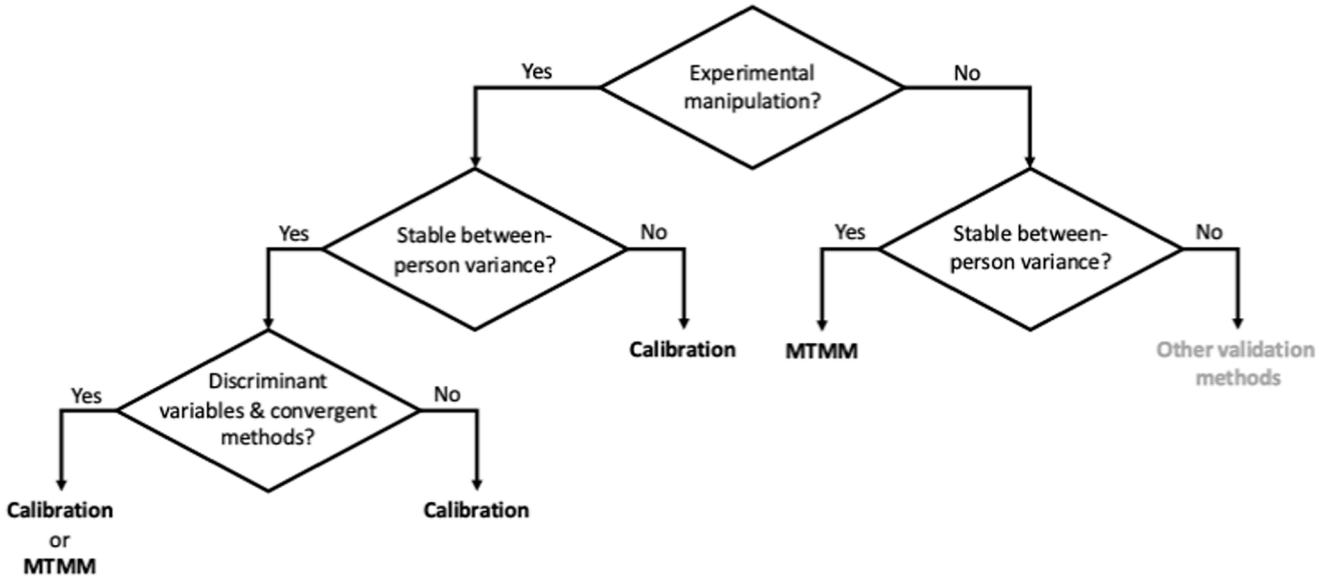


Fig. 4. Decision tree for arbitrating between MTMM and experiment-based calibration. Notes: The option "Other validation methods" denotes cases in which neither MTMM nor experiment-based calibration are applicable. In this case, researchers may have to rely on other validity arguments, such as face validity.

quite possible to manipulate the value of latent variables on time-scales that can be realized in a matter of minutes or hours. This makes calibration practical and applicable, and it solves several of the problems of the classical MTMM matrix approach in the domain of experimentally manipulable latent variables.

Declaration of generative AI use in scientific writing

The authors declare that they did not use any generative AI tools in the writing process of this article.

CRediT authorship contribution statement

Dominik R. Bach: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Edward E. Rigdon:** Writing – review & editing, Writing – original draft, Validation, Conceptualization. **Marko Sarstedt:** Writing – review & editing, Writing – original draft, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Dominik R. Bach receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2018 CoG-816564 ActionContraThreat). The Hertz Chair for Artificial Intelligence and Neuroscience in the Transdisciplinary Research Area Life and Health, University of Bonn, is funded as part of the Excellence Strategy of the German federal and state governments. The authors thank Lukas Kornemann, Josie Linnell, and Olivier de Vries, for commenting on a manuscript draft.

Appendix

Table A1
Examples of published calibration studies from different fields of experimental psychology and educational research.

Latent latent variable	Experimental manipulation	Standard values	Observable(s)	Examples
Associative fear learning & memory	Pavlovian conditioning	high, low	Psychophysiological recordings, self-reports	Greaves et al. (2024); Khemka et al. (2017); Kuhn et al. (2022); Privratsky et al. (2020); Staib et al. (2015); Wehrli et al. (2022); Xia et al. (2023)
Stimulus-evoked arousal	Pictures, sounds, pain, target identification task	high, low	Psychophysiological recordings	Bach (2014); Bach et al. (2013)
Anxiety	Public speaking	high, low	Psychophysiological recordings	Bach et al. (2010a, 2010b); Bach & Staib (2015)
Mental effort	Mental arithmetic	high, low	Psychophysiological recordings	Bach & Staib (2015)
Interprofessional collaborative competency	Training	High, low	Self-report questionnaire	Lunde et al. (2021)
Team skills	Training	High, low	Examiner rating scale	Wright et al. (2013)
Knowledge and attitude towards a teaching program	Exposure to the teaching program	High, low	Self-report questionnaire and knowledge test	Taylor et al. (2001)

Data availability

The simulation code and results are available on the Open Science Foundation platform: https://osf.io/2a6qp/

References

AERA, APA, & NCME. (2014). *The standards for educational and psychological testing.* American Educational Research Association.

Bach, D. R. (2014). A head-to-head comparison of SCRalyze and Ledalab, two model-based methods for skin conductance analysis. *Biological Psychology, 103*, 63–68. https://doi.org/10.1016/j.biopsycho.2014.08.006

Bach, D. R. (2023a). *Experiment-based calibration in psychology: Foundational and data-generating model* [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/4t32a

Bach, D. R. (2023b). Psychometrics in experimental psychology: A case for calibration. *Psychonomic Bulletin & Review, 31*, 1451–1470. https://doi.org/10.3758/s13423-023-02421-z

Bach, D. R., Daunizeau, J., Friston, K. J., & Dolan, R. J. (2010a). Dynamic causal modelling of anticipatory skin conductance responses. *Biological Psychology, 85*(1), 163–170. https://doi.org/10.1016/j.biopsycho.2010.06.007

Bach, D. R., Friston, K. J., & Dolan, R. J. (2010b). Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations. *International Journal of Psychophysiology, 76*(1), 52–55. https://doi.org/10.1016/j.ijpsycho.2010.01.011

Bach, D. R., Friston, K. J., & Dolan, R. J. (2013). An improved algorithm for model-based analysis of evoked skin conductance responses. *Biological Psychology, 94*(3), 490–497. https://doi.org/10.1016/j.biopsycho.2013.09.010

Bach, D. R., & Melinscak, F. (2020). Psychophysiological modelling and the measurement of fear conditioning. *Behavior Research and Therapy, 127*, Article 103576. https://doi.org/10.1016/j.brat.2020.103576

Bach, D. R., Melinscak, F., Fleming, S. M., & Voelkle, M. C. (2020). Calibrating the experimental measurement of psychological attributes. *Nature Human Behaviour, 4*, 1229–1235. https://doi.org/10.1038/s41562-020-00976-8

Bach, D. R., Sporrer, J., Abend, R., Beckers, T., Dunsmoor, J. E., Fullana, M. A., Gamer, M., Gee, D. G., Hamm, A., Hartley, C. A., Herringa, R. J., Jovanovic, T., Kalisch, R., Knight, D. C., Lissek, S., Lonsdorf, T. B., Merz, C. J., Milad, M., Morriss, J., & Schiller, D. (2023). Consensus design of a calibration experiment for human fear conditioning. *Neuroscience & Biobehavioral Reviews, 148*, Article 105146. https://doi.org/10.1016/j.neubiorev.2023.105146

Bach, D. R., & Staib, M. (2015). A matching pursuit algorithm for inferring tonic sympathetic arousal from spontaneous skin conductance fluctuations. *Psychophysiology, 52*(8), 1106–1112. https://doi.org/10.1111/psyp.12434

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General, 148*(3), 437–452. https://doi.org/10.1037/xge0000511

Borgstede, M., & Eggert, F. (2023). Squaring the circle: From latent variables to theory-based measurement. *Theory & Psychology, 33*, 118–137. https://doi.org/10.1177/09593543221127985

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Coote, L. V. (2011). Measurement properties of ranking and ratings. *Journal of Business Research, 64*(12), 1296–1302. https://doi.org/10.1016/j.jbusres.2010.12.006

Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research, 14*(4), 579–582. https://doi.org/10.1086/209137

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

Czakon, W., Klimas, P., Kawa, A., & Kraus, S. (2023). How myopic are managers? Development and validation of a multidimensional strategic myopia scale. *Journal of Business Research, 157*, Article 113573. https://doi.org/10.1016/j.jbusres.2022.113573

Deer, L., Adler, S., Datta, H., Mizik, N., & Sarstedt, M. (2025). Toward open science in marketing research. *International Journal of Research in Marketing, 42*(1), 212–233. https://doi.org/10.1016/j.ijresmar.2024.12.005

Deaves, R., Lüders, E., & Luo, G. Y. (2009). An experimental test of the fmpact of overconfidence and gender on trading activity. *Review of Finance, 13*(3), 555–575. https://doi.org/10.1093/rof/rfn023

Eid, M., & Nussbeck, F. W. (2009). The multitrait-multimethod matrix at 50! *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*(3), 71. https://doi.org/10.1027/1614-2241.5.3.71

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*(3), 230–253. https://doi.org/10.1037/a0013219

Estler, W. T. (1999). Measurement as inference: Fundamental ideas. *In CIRP Annals, 48*(2), 611–632. https://doi.org/10.1016/S0007-8506(07)63238-7

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324–327. https://doi.org/10.1016/j.jbusres.2009.05.003

Fiske, D. W. (1982). Convergent-discriminant validation in measurements and research strategies. *New Directions for Methodology of Social & Behavioral Science, 12*, 77–92.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Franke, G. R., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: A comparison of four procedures. *Internet Research, 29*(3), 430–447. https://doi.org/10.1108/IntR-12-2017-0515

Franke, G. R., Sarstedt, M., & Danks, N. P. (2021). Assessing measure congruence in nomological networks. *Journal of Business Research, 130*, 318–334. https://doi.org/10.1016/j.jbusres.2021.03.003

Girard, A., Lichters, M., Sarstedt, M., & Biswas, D. (2019). Short- and long-term effects of nonconsciously processed ambient scents in a servicescape: Findings from two field experiments. *Journal of Service Research, 22*(4), 440–455. https://doi.org/10.1177/1094670519842333

Greaves, M. D., Felmingham, K. L., Ney, L. J., Nicholson, E., Li, S., Vervliet, B., Harrison, B. J., Graham, B. M., & Steward, T. (2024). Using electrodermal activity to estimate fear learning differences in anxiety: A multiverse analysis. *Behaviour Research and Therapy, 104598*. https://doi.org/10.1016/j.brat.2024.104598

Haig, B. D., & Evers, C. W. (2015). *Realist inquiry in social science.* Sage.

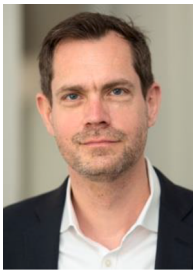
Hair, J. F., Sarstedt, M., Ringle, C. M., Sharma, P. N., & Liengaard, B. D. (2024a). Going beyond the untold facts in PLS-SEM and moving forward. *European Journal of Marketing, 58*(13), 81–106. https://doi.org/10.1108/EJM-08-2023-0645

Hair, J. F., Sharma, P. N., Sarstedt, M., Ringle, C. M., & Liengaard, B. D. (2024b). The shortcomings of equal weights estimation and the composite equivalence index in PLS-SEM. *European Journal of Marketing, 58*(13), 30–55. https://doi.org/10.1108/EJM-04-2023-0307

Helm, J. L. (Ed.). (2022). *Advanced multitrait-multimethod analyses for the behavioral and social sciences.* Routledge, Taylor & Franics Group.

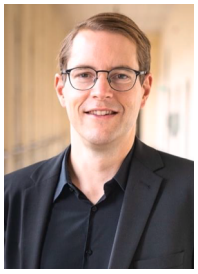
Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the*

Academy of Marketing Science, 43, 115–135. https://doi.org/10.1007/s11747-014-0403-8
Höfling, V., Schermelleh-Engel, K., & Moosbrugger, H. (2009). Analyzing multitrait-multimethod data: A comparison of three approaches. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*(3), 99–111. https://doi.org/10.1027/1614-2241.5.3.99
Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. https://doi.org/10.1007/BF02291366
Kane, M. T. (2016). Explicating validity. *Assessment in Education-Principles Policy & Practice, 23*(2), 198–211. https://doi.org/10.1080/0969594x.2015.1060192
Kenny, D. A. (2021). Multitrait-multimethod matrix: Method in the madness. In J. L. Helm (Ed.), *Advanced multitrait-multimethod analyses for the behavioral and social sciences* (pp. 16–27). Routledge.
Khemka, S., Tzovara, A., Gerster, S., Quednow, B. B., & Bach, D. R. (2017). Modeling startle eyeblink electromyogram to assess fear learning. *Psychophysiology, 54*(2), 204–214. https://doi.org/10.1111/psyp.12775
Kline, R. B. (2011). Convergence of structural equation modeling and mulitlevel modeling. In M. Williams, & W. P. Vogt (Eds.), *Handbook of methodological innovation* (pp. 562–589). Sage.
Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. Academic Press.
Kuhn, M., Gerlicher, A. M. V., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM. *Psychophysiology, 59*(9). https://doi.org/10.1111/psyp.14058. Article e14058.
Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology, 21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb
Lunde, L., Bærheim, A., Johannessen, A., Aase, I., Almendingen, K., Andersen, I. A., Bengtsson, R., Brenna, S. J., Hauksdottir, N., Steinsbekk, A., & Rosvold, E. O. (2021). Evidence of validity for the Norwegian version of the interprofessional collaborative competency attainment survey (ICCAS). *Journal of Interprofessional Care, 35*(4), 604–611. https://doi.org/10.1080/13561820.2020.1791806
Maas, C. J. M., Lensvelt-Mulders, G. J. L. M., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*, 72–77. https://doi.org/10.1027/1614-2241.5.3.72
Mancinelli, F., Sporrer, J. K., Myrov, V., Melinscak, F., Zimmermann, J., Liu, H., & Bach, D. R. (2024). Dimensionality and optimal combination of autonomic fear-conditioning measures in humans. *Behavior Research Methods, 56*(6), 6119–6129. https://doi.org/10.3758/s13428-024-02341-3
Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2), i–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x
Mishra, D. P. (2000). An empirical assessment of measurement error in health-care survey research. *Journal of Business Research, 48*(3), 193–205. https://doi.org/10.1016/S0148-2963(98)00088-5
Moore, D. A. (2022). Overprecision is a property of thinking systems. *Psychological Review, 130*(5), 1339–1350. https://doi.org/10.1037/rev0000370
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502
Nascimento, J., & Loureiro, S. M. C. (2024). Understanding the desire for green consumption: norms, emotions, and attitudes. *Journal of Business Research, 178*, Article 114675. https://doi.org/10.1016/j.jbusres.2024.114675
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* ((3rd ed.).). McGraw-Hill.
Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research, 67*(8), 1766–1770. https://doi.org/10.1016/j.jbusres.2014.03.002
Ong, C.-S., Chang, S.-C., & Lee, S.-M. (2015). Development of WebHapp: Factors predicting user perception of website-related happiness. *Journal of Business Research, 68*(3), 591–598. https://doi.org/10.1016/j.jbusres.2014.09.002
Oort, F. J. (2009). Three-mode models for multitrait-multimethod data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 5*(3), 78–87. https://doi.org/10.1027/1614-2241.5.3.78
Pecot, F., Vasilopoulou, S., & Cavallaro, M. (2021). How political ideology drives anti-consumption manifestations. *Journal of Business Research, 128*, 61–69. https://doi.org/10.1016/j.jbusres.2021.01.062
Phillips, S. D., Estler, W. T., Doiron, T., Eberhardt, K. R., & Levenson, M. S. (2001). A careful consideration of the calibration concept. *Journal of Research of the National Institute for Standards in Technology, 106*(2), 371–379. https://doi.org/10.6028/jres.106.014
Pillai, K. G. (2014). Range of confidence scale and consumer knowledge calibration. *Psychological Reports, 114*(1), 149–155. https://doi.org/10.2466/03.01.PR0.114k15w5
Privratsky, A. A., Bush, K. A., Bach, D. R., Hahn, E. M., & Cisler, J. M. (2020). Filtering and model-based analysis independently improve skin-conductance response measures in the fMRI environment: Validation in a sample of women with PTSD. *Innternational Journal of Psychophysiology, 158*(December), 86–95. https://doi.org/10.1016/j.ijpsycho.2020.09.015
Radomir, L., & Moisescu, O. I. (2019). Discriminant validity of the customer-based corporate reputation scale: Some causes for concern. *Journal of Product & Brand Management, 29*(4), 457–469. https://doi.org/10.1108/JPBM-11-2018-2115
Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research, 54*(3), 429–443. https://doi.org/10.1080/00273171.2018.1535420
Rigdon, E. E., Sarstedt, M., & Becker, J. M. (2020). Quantify uncertainty in behavioral research. *Nature Human Behaviour, 4*, 329–331. https://doi.org/10.1038/s41562-019-0806-0
Rigdon, E. E., Sarstedt, M., & Moisescu, O. (2023). Quantifying model selection uncertainty via bootstrapping and Akaike weights. *International Journal of Consumer Studies, 47*(4), 1596–1608. https://doi.org/10.1111/ijcs.12906
Ringle, C. M., Sarstedt, M., Sinkovics, N., & Sinkovics, R. R. (2023). A perspective on using partial least squares structural equation modelling in data articles. *Data in Brief, 48*, Article 109074. https://doi.org/10.1016/j.dib.2023.109074
Rönkkö, M., & Cho, E. (2022). An update guideline for assessing discriminant validity. *Organizational Research Methods, 25*(1), 6–14. https://doi.org/10.1177/109442812096861
Sarstedt, M., Adler, S. J., Ringle, C. M., Cho, G., Diamantopoulos, A., Hwang, H., & Liengaard, B. D. (2024). Same model, same data, but different outcomes: Evaluating the impact of method choice in structural equation modeling. *Journal of Product Innovation Management, 41*(6), 1100–1117. https://doi.org/10.1111/jpim.12738
Suoniemi, S., Terho, H., Zablah, A.l., Olkkonen, R., & Straub, D. W. (2021). The impact of firm-level and project-level IT capabilities on CRM system quality and organizational productivity. *Journal of Business Research, 127*, 108–122. https://doi.org/10.1016/j.jbusres.2021.01.007
Spearman, C. (1904a). 'General intelligence,' objectively determined and measured. *American Journal of Psychology, 15*(2), 201–292. https://doi.org/10.1037/11491-006
Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.1037/11491-005
Staib, M., Castegnetti, G., & Bach, D. R. (2015). Optimising a model-based approach to inferring fear learning from skin conductance responses. *Journal of Neuroscience Methods, 255*, 131–138. https://doi.org/10.1016/j.jneumeth.2015.08.009
Taylor, R., Reeves, B., Mears, R., Keast, J., Binns, S., Ewings, P., & Khan, K. (2001). Development and validation of a questionnaire to evaluate the effectiveness of evidence-based practice teaching. *Medical Education, 35*(6), 544–547. https://doi.org/10.1046/j.1365-2923.2001.00916.x
Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.
Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.
Wehrli, J. M., Xia, Y., Gerster, S., & Bach, D. R. (2022). Measuring human trace fear conditioning. *Psychophysiology, 59*(12). https://doi.org/10.1111/psyp.14119. e14119.
Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1–26. https://doi.org/10.1177/014662168500900101
Wright, M. C., Segall, N., Hobbs, G., Phillips-Bute, B., Maynard, L., & Taekman, J. M. (2013). Standardized assessment for evaluation of team skills: Validity and feasibility. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 8*(5), 292–303. https://doi.org/10.1097/SIH.0b013e318290a022
Xia, Y., Wehrli, J., Gerster, S., Kroes, M., Houtekamer, M., & Bach, D. R. (2023). Measuring human context fear conditioning and retention after consolidation. *Learning & Memory, 30*(7), 139–150. https://doi.org/10.1101/lm.053781.123
Zumbo, B. D., & Chan, E. K. (2014). Reflections on validation practices in the social, behavioral, and health sciences. In B. D. Zumbo, & S. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 321–327). Springer.



Dominik Bach is a professor of artificial intelligence and neuroscience at University of Bonn (Germany) and an honorary professor at University College London (UK). His main research interest is in understanding the computing architecture of the mind and brain in critical situations, to which end he is developing experimental and data analysis methods. His work has been published in journals such as *Nature Human Behaviour*, *Nature Communications*, *Nature Reviews Neuroscience*, *Behavior Research Methods*, and others.

Edward E.Rigdon is Marketing RoundTable Professor in the Robinson College of Business at Georgia State University in Atlanta, Georgia, USA. His research on structural equation modeling and related topics in applied statistics has been published in premier journals in marketing, in information systems, and in quantitative psychology. Dr. Rigdon has been offering his seminar on structural equation modeling since 1990.



MarkoSarstedt is a Chaired Professor of Marketing at the Ludwig-Maximilians-University Munich (Germany) and an Adjunct Research Professor at Babeș-Bolyai-University Cluj-Napoca (Romania). His main research interest is the advancement of research methods to further the understanding of consumer behavior. His research has been published in *Nature Human Behavior*, *Journal of Marketing Research*, *Journal of the Academy of Marketing Science*, *Multivariate Behavioral Research*, *Organizational Research Methods*, *MIS Quarterly*, and *Psychometrika*, among others. Marko has been repeatedly named member of Clarivate Analytics' Highly Cited Researchers List, which includes the "world's most impactful scientific researchers."