




Commentaries on “Scale use and abuse: Toward best practices in the deployment of scales”

Constantine S. Katsikeas¹ | Shilpa Madan² | C. Miguel Brendl³  | Bobby J. Calder⁴  |
 Donald R. Lehmann⁵ | Hans Baumgartner⁶ | Bert Weijters⁷  | Mo Wang⁸ |
 Chengquan Huang⁸ | Joel Huber⁹

¹Marketing and International Management, Leeds University Business School, University of Leeds, Leeds, UK

²Pamplin College of Business, Virginia Tech, Blacksburg, Virginia, USA

³University of Basel, Basel, Switzerland

⁴Northwestern University, Evanston, Illinois, USA

⁵Columbia University, New York, USA

⁶The Pennsylvania State University, State College, Pennsylvania, USA

⁷Ghent University, Ghent, Belgium

⁸Warrington College of Business, University of Florida, Gainesville, Florida, USA

⁹The Fuqua School of Business, Duke University, Durham, North Carolina, USA

Correspondence

C. Miguel Brendl, University of Basel, Wirtschaftswissenschaftliche Fakultät, Universität Basel, Peter Merian-Weg 6, 4002 Basel, Switzerland.

Email: miguel.brendl@unibas.ch

Abstract

Five comments below provide strong and interesting perspectives on multi-item scale use. They define contexts and research areas where developed scales are valuable and where they are vulnerable. Katsikeas and Madan begin by taking a global perspective on scale use, demonstrating how the use and transferability of scales become even more problematic as researchers move across languages and cultures. They provide guidance for scale use that is particularly relevant to international marketing and marketing strategy research. Brendl and Calder acknowledge the use of well-formed scales as measured variables in psychological experiments, both as independent and dependent variables, but critique the use of multi-item scales to directly reveal latent unobservable constructs. As with any observed variable, scales should be used to test empirical predictions based on theoretical hypotheses about causal connections between theoretical constructs. Lehmann applauds the variability of multi-item scales and urges the exploration of the impact of various items within a scale. He advocates for flexibility and variation in multi-item scales related to psychological theories, simple three-item scales for manipulation checks, and one-item scales when measuring objective actions or beliefs. Baumgartner and Weijters focus on how to validate multi-item scales, particularly when used as mediators or moderators where a unique interpretation of the scale is so central. They recommend meta-analyses of scales that test relationships among measured scales. Like Lehmann, they worry about the impact of exhaustive scales on respondents and the impact of exhausted respondents on the scales themselves. In the final comment, Wang and Huang update our thinking on emerging ways to define and refine scales. They discuss ways to identify focal and orbital constructs and suggest item response theory as a way to adapt scales to subsets of items that best contribute to identifying individual differences between respondents. They support confirmatory factor analysis across different studies to assess scale equivalence across different contexts, cultures, and languages.

KEY WORDS

psychometrics, statistics

Accepted by Lauren Block, Editor; Associate Editor, Joel Huber

See relevant article: <https://doi.org/10.1002/jcpy.1320>, “Scale use and abuse: Towards best practices in the deployment of scales” by Kelly L. Haws, Kevin L. Sample and John Hulland.

BEST PRACTICES IN SCALE DEPLOYMENT: MAXIMIZING RELEVANCE TO CROSS-CULTURAL AND MARKETING STRATEGY RESEARCH

Constantine S. Katsikeas

Arnold Ziff Research Chair and Professor of
Marketing and International Management
Leeds University Business School
Maurice Keyworth Building
University of Leeds
Leeds LS2 9JT
UK
Email: csk@lubs.leeds.ac.uk

Shilpa Madan

Assistant Professor of Marketing
Pamplin College of Business
Virginia Tech
Blacksburg
VA 24061
USA
Email: shilpa.madan@vt.edu

An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer. But before an experiment can be performed, it must be planned—the question to Nature must be formulated before being posed.

Max Planck (Plank, 1949)

Haws et al. (2023) insightful paper on effectively using, adapting, and validating existing scales deserves the attention of social scientists in marketing, psychology, management, and beyond. It makes a compelling case for more rigorous measurement and transparent reporting of relevant scale-related decisions and provides a valuable toolkit of best practices for consumer behavior researchers to enhance the theoretical linkages of their findings with previous studies and to conduct and publish more systematic, comparable, and replicable research. While Haws et al. focus on scale deployment in experimental consumer behavior research, the conclusions and recommendations are applicable to broader marketing research and other business and management disciplines.

Compared with scale development, scale deployment has received less attention in the literature. Establishing the validity of existing measures or manipulations is not considered important unless it falls under the process of defining a new construct (Scopelliti et al., 2020). The tendency of editors, reviewers, and authors to casually accept previously used scales, whether originally validated or not, as valid may introduce confounds, lower statistical power, prevent the comparison of findings to prior research, and limit coherent cumulative knowledge

in the field. We seek to extend the relevance of Haws et al.'s recommendations to cross-cultural/international marketing and marketing strategy research by outlining unique scale usage and deployment challenges in these domains and providing recommendations that complement those of Haws et al.

Cross-cultural and international marketing research

Marketing research is increasingly expanding beyond Western, educated, industrialized, rich, and democratic populations. Because research is often conducted with different populations, with the ultimate goal of making informed comparisons across samples, cross-cultural and international consumer behavior researchers face unique measurement challenges with regard to scale usage and deployment. The importance of measurement in this area is underscored by a recent review that identified “psychometrically deficient measures” as the most pervasive methodological challenge in international business research (Aguinis et al., 2020).

When borrowing scales “As-Is” may not work

Given that most marketing scales were developed and validated only in the United States (de Jong et al., 2009), “as-is” use of validated scales can lead to measurement issues in international/cross-cultural marketing research for two main reasons. First, a scale validated in the United States may contain items that are not informative about the latent construct in other countries. Second, it may lack the relevant items to tap local cultural manifestations of the underlying construct. Thus, borrowing scales—even those that have been validated—may lead to invalid cross-national inferences.

For example, the consumer ethnocentrism scale (CETSCALE) was developed and validated with U.S. consumers (Shimp & Sharma, 1987). As evidence of nomological validity, consumers high in ethnocentrism were less positively disposed toward foreign products. However, although the scale was validated in a Western developed market, the correlation between ethnocentrism and consumers' attitudes toward foreign products was weaker in similar markets such as France and West Germany (Douglas & Nijssen, 2003). Furthermore, the relationship between ethnocentrism, assessed using CETSCALE, and negative attitudes toward foreign brands disappeared in a Polish sample (Supphellen & Rittenburg, 2001). While these different results may have several valid reasons (e.g., level of nationalism, availability/quality of foreign products in the country), it illustrates the core issue that the as-is use of validated scales may not yield theoretically expected results because validation occurred in a different context. Thus,

establishing construct equivalence is a crucial first step in rigorous cross-national and global marketing studies.

Measurement equivalence

Researchers have proposed different types of measurement equivalence at varying levels of abstraction. Consistent with Hui and Triandis' (1985) typology, conceptual equivalence requires latent constructs to have the same meaning across contexts. However, establishing conceptual equivalence may require modification of the scale for a given context. International marketing scholars argue that it may be necessary to include country-specific items in addition to, or in place of, cross-national or standardized items (Aaker et al., 2001) to ensure comprehensive coverage and specificity of the underlying construct across cultures.

While conceptual equivalence focuses on the construct, item and scalar equivalence refer to properties of the scale measuring the underlying construct. Item equivalence is the requirement that the response to a given scale item has the same meaning across contexts: "Each item should mean the same thing to subjects from Culture A as it does to those from Culture B" (Hui & Triandis, 1985, p. 134). Item equivalence may be negatively affected by culture-specific or ambiguously worded items. Through careful translation and pretesting, researchers must ensure that respondents across countries understand the scale items.

Finally, scalar equivalence occurs "if a particular score on a scale represents the same degree, intensity, or magnitude of the construct across contexts regardless of the population of which the respondent is a member" (Hui & Triandis, 1985, p. 135). Given the popularity of Likert scales in consumer behavior research, this involves ensuring the appropriateness of scale anchors across contexts. For example, Chinese respondents may associate a different intensity with the word "moderately" compared with U.S. participants. To address this, adequate attention should be paid to the translation of scale anchors, which should (1) have the same rank order and (2) cover equidistant intervals on the scale across contexts (Szabo et al., 1997). Researchers can ascertain measurement equivalence using various analytical techniques (e.g., factor analysis, structural equation modeling, item response theory; see de Jong et al. (2009)).

Use of scales to measure Hofstede's cultural dimensions

Hofstede's (2011) six-dimensional typology of culture has been the dominant approach in cross-cultural research over the past four decades. Extant consumer behavior and marketing research has examined the

influence of individualism-collectivism, power distance, and, to a lesser extent, uncertainty avoidance, long-term orientation, and masculinity on various consumption-related outcomes. While Haws et al. highlight the issue with using national-level scores for individuals in a particular country, we consider the selection of appropriate scales for assessing Hofstede's cultural dimensions.

Given the variety of scales available to measure cultural dimensions at the individual-level, cross-cultural researchers must select the most appropriate measure for the study's purpose and context. For example, individualism-collectivism may be measured at the individual level using the 24-item independence/interdependence scale by Singelis (1994), the 13-item allocentrism-idiocentrism scale by Triandis et al. (1995), or the 6-item scale by Yoo et al. (2011), among others. While some are standalone single-dimension scales, others are part of larger multidimensional scale development. In line with Haws et al., besides prioritizing validated scales, research must emphasize (1) the contexts in which the measures were validated; (2) the measures' relevance to the research question, conceptual framework, and sample (e.g., not using scales/items phrased in a work context for student samples); and (3) the measures' parsimony. If the research question involves different cultural dimensions, researchers may opt for scales developed and validated together (e.g., Madan et al., 2022) rather than independently developed and validated scales for specific dimensions—which may undermine discriminant validity among the separate scales (Yoo et al., 2011).

Survey-based marketing strategy research

Haws et al.'s guidelines are also relevant for survey-based marketing strategy research. Because these studies often develop and test multiconstruct conceptual models, scale selection and length issues require focal attention. Because marketing strategy studies are often conducted with salespeople, frontline service personnel, managers, or other nonstudent samples, excessively long surveys are ill-suited and can lead to respondent attrition and, thus, incomplete data collection. In alignment with Haws et al.'s individual-level focus, we consider marketing strategy research focusing on attitudes, behaviors, assessments, and/or outcomes of individuals (e.g., salespeople) and not on research dealing with firm-level constructs.

Because marketing strategy research is often conducted in different industry contexts, it is easy to fall prey to both ends of the continuum—using less suitable existing measures that are not meaningful to the specific study context or overadapting existing measures to fit the study context—which may render results incomparable with those in previous literature and potentially impede

collective theory building. For example, sales(force) control, a construct extensively studied in marketing, is widely believed to drive salesperson performance. Thus, research has investigated various factors mediating the sales control–performance link, including salesperson learning (Katsikeas et al., 2018), customer orientation, and sales innovativeness (Evans et al., 2007), attributional dimensions (Fang et al., 2005), and job engagement and stress (Miao & Evans, 2013), as well as various sales outcomes, including salesperson job satisfaction and performance (e.g., Evans et al., 2007) across multiple industries. Despite the centrality of sales control in the literature, the construct has been operationalized in multiple ways with different underlying dimensions. Kohli et al. (1998) conceptualize control as a three-dimensional construct and employ 12 items. Fang et al. (2005) use 21 items and Katsikeas et al. (2018) employ eight items and one ratio indicator to tap these same dimensions, while Sarin et al. (2012) consider the outcome and process supervisory control actions, each comprising risk and reward dimensions measured by four three-item scales.

Such wide variation in operationalization makes the comparability of findings less meaningful and limits the potential for systematic replication studies that assess generalizability, lend credibility to extant knowledge, and advance theory. As Haws et al. argue, applying a less relevant scale can result in methodological confounds, loss of theoretical relevance, and lack of interpretation, thus undermining the synthesis of findings across studies for theory building.

While scholars have recommended using shorter multi-item scales to balance the need for high-quality responses and rigorous hypothesis testing, ad hoc reduction in the number of items can, as Haws et al. suggest, yield less-than-adequate coverage for theory testing and building and limit practical insights. Some techniques used to construct and validate scales may also be used to address these concerns. After identifying commonly used scales to measure a construct of interest and contexts in which these scales have been used, researchers may assess comparative applicability and relevance using an expert panel comprising academics familiar with research in the field and practitioners. Panelists should be informed about the research question, the focal construct, its dimensions, and the study context (Malhotra et al., 2012) to adequately assess which available measures are most suitable for the research.

This panel may assess the quality of individual items when researchers seek to modify items or remove/add items to an existing scale. Judges may be asked to evaluate items on their similarity, clarity, and representativeness to identify items that may be excluded without a significant loss of coverage of the construct's domain of content (Malhotra et al., 2012). Experts may also be used to evaluate the scale deployment of higher-order multidimensional constructs. Researchers could assess the level of agreement among experts on the extent to which items within each dimension

fall into the conceptual domain of a specific dimension and contribute to the focal construct. Subsequently, measures selected through expert panel screening should be submitted to rigorous validation procedures.

Conclusion

It may be argued that the key objectives of empirical research in marketing are to deepen our understanding of marketplace phenomena, advance and build theory, and generate insights that are valuable to various stakeholders/constituents. What is less debatable is that all these objectives depend on good measurement. Using the right measures is imperative to ensure that research conclusions, whether they contribute to theory or practice, are defensible. As Haws et al. and our commentary illustrate, marketing research can benefit from a critical eye on measurement, especially the use of scales, to ensure that our discipline can meaningfully extend old theories and build new ones and help stakeholders (e.g., managers, nongovernmental organizations, the public) understand and apply our research.

In their systematic review, Haws et al. make a compelling case for more rigorous measurement in consumer behavior research. Their broad-based pragmatic guidelines and their call for more transparent and detailed reporting of relevant scale-related decisions are essential for conducting research that inspires confidence and can be systematically and meaningfully extended and built on. Given the relevance of these guidelines for marketing researchers more broadly, in this commentary, we extend the relevance of Haws et al.'s recommendations to other subfields of marketing research, namely, cross-cultural/international marketing and marketing strategy research. We underscore the different challenges faced by researchers in these subfields regarding the use and deployment of existing scales and offer a set of recommendations, complementing those offered by Haws et al., to systematically address these concerns.

USING SCALES IN RESEARCH: MEASURED CONSTRUCTS OR MEASURED VARIABLES?

C. Miguel Brendl

University of Basel
Basel
Switzerland

Email: miguel.brendl@unibas.ch

Bobby J. Calder

Northwestern University
Evanston
Illinois
USA

Scales are regarded either as measured constructs or measured variables. Researchers should take these two interpretations into account. Haws et al. (2023) recommend the former and provide advice about deploying scales treated as measured constructs. There are, however, good reasons to treat scales as measured variables. We emphasize that scales are useful for theory testing, that is, for drawing inferences about construct-to-construct causal relations. The question is—how to use them for this purpose. Researchers face a choice between the two approaches to using scales.

Haws et al. “...assert that scales can and should be effectively utilized to measure key constructs.” (p. 227) They point out that their assertion is opposed to the position articulated in Calder et al. (2021). Below we discuss where exactly we see disagreements between these two positions. Haws et al. focus on the deployment of scales that were previously developed via construct validation, a process that aims to establish a measurement relationship between scales and constructs. We will question that the current practice of construct validation during scale development results in a straightforward interpretation of this relationship.

What are measured constructs and what is construct validation?

We assume researchers would agree with Calder et al. that observable variables are things that can be measured or manipulated. A response to an item on a scale is an observable measured variable, therefore so is a set of average responses to multiple items that form a scale. This is the case irrespective of why several items are combined into a scale. We also think it is uncontroversial that the reason for using the latent unobserved construct distinction, as in Haws et al., is to identify something that is different from an observed variable—something that is by definition not observable. Latent refers to something that is hidden or concealed. This implies that something unobservable is revealed, by its measurement. Hence the term measured construct. We note, as discussed below, that Calder et al.’s view of constructs differs from this use. This difference is not about terminology. If a word other than construct is used, the same issue arises.

The key question is, what is the relationship between observable variables and constructs? We do not repeat the discussion in Calder et al. here. Their key conclusion is that, in theory-testing, cause-effect relationships among constructs serve to explain associations among observable variables. We emphasize that constructs are part of explanations. They are useful in a theory to the degree they have explanatory value and to the degree they can explain observations. In order to have explanatory value, they need to be more abstract than observations (more universal according to Popper, 1935/2004).

However, many researchers (e. g., Flake & Fried, 2020; Vazire et al., 2022), advocate the use of scales as measured

constructs based on construct validity, but what is the underlying logic of this? Haws et al. describe the typical procedure that arrives at assigning construct validity to a scale. (1) Scale development begins with articulating an unobservable theoretical construct. Researchers generate an initial pool of questionnaire items, often based on face validity, that is, on expert intuition. (2) Then, the researchers select a subset of items (observed variables) based on criteria of construct validation. These criteria are a very high correlation of the items with each other (internal reliability), a moderately high correlation with other items meant to measure something similar (convergent validity), and a low correlation with items that are thought to measure something different (discriminant validity). (3) Finally, Haws et al. refer in passing to nomological validity as articulating the relation of the construct to other constructs. In sum, it is common practice to treat questionnaire items that have “survived” this procedure as a *measure* of a construct. We remind here that our discussion is exclusively about constructs as unobservable, latent concepts. According to Haws et al., the scale can then be deployed in theory-testing research if the scale items are not changed too much. They offer extensive recommendations on how much is too much. We agree with their important point that modifying scale items risks changing their statistical properties.

What is it that scale values measure?

Step 2 of the above procedure aggregates multiple observations from one respondent into a single scale such that each respondent is assigned a scale value. The scale value still is an observed variable. But how do scale values become measured constructs? Or put differently, what do scale values measure if they are not the same as measured variables? We present two alternative views about this and evaluate each.

Scales as measured constructs

In this view, scale values make visible unobservable constructs. Scales allow inferring constructs independently of what role the construct plays as an explanation in any particular study. Sometimes researchers use the term construct merely as a semantic label that describes observations (scale values), where essentially the scale and construct are the same (cf. Calder et al.). Sometimes they additionally devote more or less effort to “defining” the construct.

Scales as measured variables

According to a second view, scale values are measured variables. Scales are used to determine empirical

effects by relating them to other observable variables. These empirical effects in turn can test theoretical explanations that involve constructs, where constructs are abstract and unobservable. In this view, scales and constructs are different. One cannot reveal constructs through scales, but can nevertheless employ them as measured variables in testing theory. Specifically, one can predict how scale values (i.e., observed variables) should co-vary with other observed variables based on theory. According to this view, scales are measured variables and researchers should use them in the same way as manipulated variables (though due to the lack of random assignment, scales are subject to selection and history). Scales can serve as both independent or dependent variables in psychological experiments. Again, in either case, the role of scale values lies in testing a prediction a theory makes for observable variables—thus, observable scale values should be used to test unobservable theoretical explanations.

Next, we evaluate each of the two views as to what it is that a scale value measures. We begin with the view that scale values make visible something invisible. This idea involves a challenge. On the one hand, a set of answers to a scale is an observable variable. On the other hand, these answers capture a construct that is not observable. The scale developer assumes that the observations—after a process of aggregation—somehow reveal the construct, i. e., make it accessible to an observer. But on the face of it, we have a contradiction—something that is unobservable but latently observable. If one asserts that constructs can be revealed by observing answers to items, one cannot also assume that constructs are fully unobservable. This may underlie the notion of latent unobservability. However, we find this notion mysterious because it does not spell out how one moves from the observable to the unobservable.

Possibly, proponents of this perspective attribute the emergence of the unobservable to statistical aggregation techniques. Step 2 of construction validation involves multivariate statistical techniques like factor analysis. In statistical parlance, factors explain co-variance among their underlying variables. This meaning of explanation in statistics is very different, however, from its meaning in a theory of cause and effect (Alexandrova & Haybron, 2016). Statistical aggregation processes during construct validation do not reveal constructs that serve as causes or effects in theoretical explanations. Consider an example that there exist particular symptoms that regularly co-occur in patients. The symptoms load high on a factor, with their variation being statistically explained by the factor. However, whereas this knowledge suggests that the symptoms may have a common cause, it alone does not identify the cause, that is, does not explain why the symptoms occur, what the underlying mechanism is. The term “explanation” as used in statistics and in theory could thus be a source of misunderstanding.

More likely, proponents of this perspective have faith in their ability (Step 1) to define the “construct” and its “domain,” as endorsed by Haws et al. In practice, the criteria for such construct definitions and domains, however, are extremely vague; often definitions seem little more than intuitions or lay psychology terms. As with the oft-used Positive and Negative Affect Scale (PANAS), which is based on everyday English mood terms, construct definitions are often little more than the name of the scale. In these cases, the construct is “defined” as whatever the scale items are assumed to measure. This way of defining constructs is reflected in Haws et al.’s contention that some constructs, purchase intentions, or general attitudes, for instance, are so straightforward that scales may not be necessary because their definitions are obvious. These measured constructs essentially define themselves. They do so, however, in terms of everyday meaning.

At the extreme, the exercise of defining constructs implicitly allows for a view where anything can be a measured construct. For example, consider a set of items related to how one feels about cleaning the kitchen. It is likely that one can find a set of items (such as a set of items related to taking out the garbage) that is moderately correlated with this, and a set of items (such as items related to brushing one’s teeth) that is uncorrelated with it. If one then offers a plausible interpretation of the kitchen items, one meets the criterion of face validity, and the scale must be regarded as a valid measure of a construct, maybe the construct of Kitchen Cleaning Affect. We do not think that is an absurd exaggeration. The number of possible constructs is infinite as long as the scale construction requires only a vague definition.

Alexandrova and Haybron (2016) point to yet another challenge of the construct validation process, one concerning convergent and discriminant validity. Consider two measures of life satisfaction. A cognitive measure correlates more highly with material life outcomes (e.g., good governance correlates with more life satisfaction), and an affective measure correlates with relationship outcomes (e.g., better relationships correlate with more life satisfaction). By merely defining life satisfaction in a plausible manner, we could turn either measure into the better one, and then, each correlation could serve either as evidence for convergent validity or discriminant validity. Neither statistics nor definitions based on plausibility can solve this problem. In contrast, a measure can take on meaning in the context of a theory that spells out cause-effect relationships. This is in line with the second view as to what scales values measure.

Calder et al. contend that this second view, scales as measured variables testing theoretical explanations, is more defensible. More specifically, scale values are measured variables that should be deployed to determine empirical effects and test theoretical explanations

using constructs that are abstract and unobservable. This brings us to the points Calder et al. make, concerning the question of how to decide on the merits of a theory. They argued current experimental research practice calls for adhering to a list of experimental procedures, which they called the Verification Approach. These procedures favor constructs that have a close fit with observations at the expense of the explanatory power of the theory, at times even to a degree that the constructs are mere descriptions of the observations. The construct validation approach in scale development, if implemented by means of a focus on statistical procedures and a disregard for cause-effect relationships at the construct level, leads to a similar outcome. The constructs that scales measure turn into mere re-descriptions of observed variables.

Is the Verification Approach on stronger ground when scales are used as moderators? Our view is that scales still function as measured variables. Consider two communications, A and B. A study finds that A is more persuasive when PANAS-scores are low, whereas B is more persuasive when PANAS-scores are high. PANAS moderates the effect of communication type. A conceptualization that is merely a summary label of the PANAS items cannot explain this, but a hypothesis about construct-to-construct relations could. Treated as an observed variable, the scale would be conducive to testing this hypothesis. As discussed by Calder et al. (p. 195), the same logic holds for mediator variables.

A scale that has undergone “construct” validation could have been developed from either of the two perspectives outlined above. The scale user has to deal with this ambiguity rather than merely accepting that such scales are measured constructs.

Recommendation for scale deployment

We do agree with Haws et al. that in testing theory it can be very fruitful to use scales. However, we suggest treating these scales as measured variables, not as measured constructs. The theory to be tested, then, needs to do the work of predicting how scale values will change, for instance, in conjunction with experimental manipulations or as a response to experimental manipulations. This falls under nomological validity, which then, contrary to Haws et al.’s position, would be central to scale deployment. In theory testing, for every use of the scale, a theory needs to be able to make predictions for the scale values, it needs to explain why particular values are expected. [Corrections made on 15 November 2022, after first online publication: This commentary has been updated with minor changes throughout in order to rectify a mistake made during the production process. The updates do not change the meaning of the work.]

BENEFITS AND LIMITATIONS OF MULTI-ITEM SCALES

Donald R. Lehmann
Columbia University
New York
New York
USA

In their paper, Haws et al. (2023) have done a wonderful job documenting and categorizing how scales have been employed in marketing and detailing a procedure for considering whether and how to use them. What follows, then, is not a critique but rather some additional thoughts on scale development and use.

Improving interval validity

The basic rationale for including multiple measures of a construct is to reduce the variance of its measure. This leads to an interesting dilemma. If the items are not sufficiently related to each other due to errors in responses, assessing multiple subaspects of a construct, or respondent fatigue or lack of interest, it will take a larger number of items to capture it (thus increasing fatigue and the impact of response style).

Consider two items and their overlap. If the two items are uncorrelated, then the variance of the average follows the “standard variance of a single item divided by n ” rule, i.e., it reduces variance. Unfortunately, if they are uncorrelated this suggests they are neither formative nor reflexive indicators of the same construct. On the contrary, if they are perfectly correlated, the variance of the average of the items equals that of a single item, meaning adding items does not improve precision. This suggests there is a “just right” level of correlation among items. However, the tendency is to focus on maximizing internal consistency or selecting items that are maximally correlated with each other to increase coefficient alpha. Unfortunately, this will produce scales with semantically similar items and is not consistent with the notion of multi-methods. This is especially an issue in developing original scales where dropping less-correlated items can noticeably overstate/inflate coefficient alpha (Kopalle & Lehmann, 1997), as will simply adding more items.

Also, I have observed that after a few (say, seven) items, respondents tend to look for a way to finish a survey as much as to answer a question, so response style and carryover effects (deJong et al., 2012) increase, giving one an inflated sense of construct coherence. Putting correlated items next to each other, while easing the respondent burden, further increases “artificial” correlations between items.

Why should the original scale have special status?

Many authors do try to use something close to the “original.” This gives a “first mover” advantage to the initial scale. However, the first is not necessarily the best. In a sense, this is related to the tendency in replication work to treat the first results as the “gold standard” and deviations from it in terms of lack of significance as evidence of failure to replicate. (N.B. Failure to replicate should be based on the significance, statistical and in terms of the size of the effect, of the difference between the original and subsequent studies, not on the significance of the subsequent study.) Put differently, some imperfectly-replicated scales may be upgraded.

The value of imperfect scale use

Exact scale replication use confounds measures and constructs. If the exact same items are always used to measure a construct, any effects found can be attributed to either (a) the construct or (b) the specific items being used. To develop general knowledge, some variation in all aspects, including measures, is desirable. In other words, the logic of “conceptual replication” (Lynch et al., 2015) extends to scales as well. Also, as Haws et al. (2023) point out, the meaning of constructs changes over time and varies across cultures. Consider the constructs “cool” and authentic. What made someone cool in the 1950s or 1970s (think disco and bell-bottom pants) is definitely not cool presently. Re. authentic, this has emerged as a key construct in the U.S. only recently.

In summary, just as imperfect replication benefits knowledge development in general (Lynch et al., 2015), so does (minor) variation in scale use increase (if the results are similar) or decrease (if the results differ) confidence in estimated construct effects. Varying measures and accounting for this variation in a meta-analysis is the best way to produce generalizations.

The impact when measuring multiple constructs

Similar to but even more important than the impact of using a large number of items to measure a single construct, using multiple constructs with multiple items has multiple unintended consequences. These include fatigue and irritation (and increased error/carelessness in each question's response), reliance on response style to provide responses, e.g., yea-saying, carryover effects (DeJong et al., 2012), and premature terminations. Measuring ten constructs with ten items each means participants have to answer one hundred questions.

This means it makes sense to make initial scales contain 3–5 items; their correlations with bigger ones are likely to be 0.90 or above and one often has to “stretch” to come up with additional items, many of which

are minor tweaks on the initial ones (Böckenholt & Lehmann, 2015). This is especially true when the goal is to assess average values (Hulbert and Lehmann, 1975).

Construct objectivity

Some constructs are fairly vague and not consciously available, while others are quite specific. Both independent variables such as income (which of course respondents may be reluctant to provide or provide accurately) and dependent ones (dollars donated to a particular cause) can typically be assessed with a single (hopefully top-of-mind) question. This manifested itself in a meta-analysis of Fishbein Model results (Farley et al., 1981), which found that in marketing single-item attitude dependent measures worked better than multiple-item ones, whereas in social psychology multiple-item measures worked better. After the fact, it became clear that the marketing attitude DVs were simpler, while the social psychology papers had studied more nuanced constructs.

Institutional constraints

Some research platforms such as Suzy have significant restrictions on the number of questions that can be asked. Similarly, a firm may be willing to allow a few questions to be added to one of their studies. For these, the use of existing scales is not possible, and severely shortened (or even one item) scales may be the only option.

Limitations on who can be studied

While long surveys may be tolerated by students and MTurks, they are not likely to be so for executives, people who are busy or in a hurry, etc. This leads to an important “selection”/sample frame issue in interpreting results and perhaps more important discourages research on “important” people and messy topics.

Valuable data may lack strong psychometric properties

Related to the previous point, considerable useful data exist for exploring consumer behavior issues (e.g., The American Consumer Satisfaction Index, Y&R's BAV scales, YouGov). Almost none of them have ideal psychometric properties, but they do contain valuable information. Similarly, text mining can capture useful information but not in the form of six-point scales. The fact that they do not allow for the use of an existing scale—or the opportunity to explore their psychometric properties—does not necessarily make them un-useful. Here the criterion is whether they are “close enough” to

the construct in question to provide (imperfect) information about it and its relation(s) to other constructs.

“Scales” often are collections of scales

Many scales have subscales that are logically distinct but related to each other. An important decision is whether they are logically distinct and hence should be treated as different (and potentially causally related constructs) or are in fact measures of the same construct. In any case, including multiple measures of each subscale, even if they are conceptually and/or statistically different (load on separate factors/are discriminately distinct) clearly increases the respondent burden.

Exploratory research

In order to test a given theory, one wants to make sure the measures match the constructs. This is enhanced by using existing scales or carefully adapted ones following the suggestions of Haws et al. (2023). On the contrary, in exploratory work, it is desirable to include a number of variables to test post hoc whether they “matter” as alternative explanations, mediators, or moderators, etc. In these cases, a single item (or at most three items) is probably adequate as long as the need for follow-up work is acknowledged.

Summary

Over 40 years ago, Churchill (1979) produced a seminal contribution to scale development and use. Since then, the field has refined it, developed compendiums of existing scales (e.g., Beardon et al., 2010), and paid more attention to it. The paper by Haws et al. (2023) follows in and expands on this tradition. It is commendable for its thoughtfulness, logic, and suggestions. It is a definite “must read” for PhD methods courses. However, it would be unfortunate if readers consider it a bible/formula for getting papers published. Rather they should treat it as an important input into their research decisions. Most important, full and accurate reporting of what was done allows subsequent researchers to better interpret results and, in time, include them in meta-analyses.

EXTERNAL VALIDITY IS MORE IMPORTANT THAN INTERNAL CONSISTENCY

Hans Baumgartner
The Pennsylvania State University
University Park
Pennsylvania
USA

Bert Weijters
Ghent University
Ghent, Belgium

Based on a review of measurement scale usage in four recent issues of the *Journal of Consumer Psychology*, Haws et al. (2023, this issue) offer a set of practical guidelines on how to deploy existing scales for measuring constructs of interest, with the goal of improving current scale use practices. In doing so, they address a blind spot in the scaling literature, as most attention so far has focused on scale construction and validation, rather than actual scale usage. The purpose of this commentary is twofold. First, we offer a critical reflection on some aspects of the approach proposed by Haws et al. and call for more research devoted to meta-analytic evaluations of scales (including applications of scales following their initial development). Second, we try to broaden the discussion of scale deployment, which is currently mainly focused on measurement instruments and tends to ignore the respondents who are asked to complete these instruments.

Haws et al. review 66 scale deployments and classify them along two dimensions: whether researchers used a validated or improvised scale, and whether the scale was used as is or modified. Surprisingly, and maybe disconcertingly, only 16 scale deployments (or 24%) used a validated scale as is. A total of 33 scale deployments (or 50%) involved an improvised scale, where an ad hoc scale was used as is or in modified form (in either case the usage is ad hoc), and in 17 cases a validated scale was modified. Based on this evidence, the authors propose best practice guidelines for identifying, and assessing the fit of, a scale for the purpose at hand; modifying scales and validating modified scales; and reporting evidence about the reliability and validity of chosen scales.

Little empirical evidence about Haws et al.’s proposed first step (scale selection and fit assessment) is available, but we imagine that many researchers already do what Haws et al. suggest (esp. since researchers themselves are presumably the best experts to evaluate the appropriateness of a scale for the purpose at hand). If no validated existing scale can be found, modifications and subsequent validation studies are necessary, according to Haws et al. This puts an additional onus on researchers, many of whom already feel overburdened by the ever-increasing demands that have to be met in terms of data sharing, replicability, care for privacy, ethical research practice regulations, preregistration, transparency, etc. It is therefore necessary to be pragmatic about when a scale can be judged to be validated, when scale modifications may or may not be problematic, and what type of additional validation is necessary.

Supposedly thorough scale development and validation efforts usually contain numerous studies devoted entirely to the assessment of the internal psychometric properties of an instrument (dimensionality, reliability and convergent validity, discriminant validity), using

mechanistic and ritualistic procedures outlined in various methodological articles. If the external validity (e.g., nomological validity) of a scale is investigated at all, which is frequently not the case in any meaningful sense of the word (e.g., is it surprising that brand love is correlated with repeat purchase intentions or intentions to recommend the brand to others?), an unweighted average of the items in the instrument is usually computed, which makes elaborate investigations of the multidimensional structure of a scale and the discriminant validity of the dimensions essentially superfluous. An instrument such as the need for cognition scale (even in its abbreviated version; see Cacioppo et al., 1984) performs terribly in an examination of the internal psychometric properties of the scale, and the initial scale development and validation reported by the authors would nowadays be unlikely to be acceptable in any reputable marketing journal, but the scale has proven valuable in numerous applications and we would undoubtedly prefer that scale to one in which the items in the scale are minor rewordings of the same idea, even if the latter scale has undergone a "rigorous" scale development and validation process. In other words, the meaning of rigorous validation is often not very clear. In light of all this, the validation procedures proposed by Haws et al. are somewhat narrow, in that they are predominantly focused on the internal psychometric properties of measures (see Clifton, 2020, for a similar point).

The required elaborateness of validation will also depend on the purpose that the scale serves. For example, a manipulation check that is closely aligned with the manipulation itself probably does not require rigorous validation. Many response scales used as dependent variables in experiments (e.g., ad or brand evaluations) are also unlikely to require extensive validation. The situation is different in the case of mediators and moderators, for which reliable and valid measurement is a *sine qua non*. However, reliability and convergent validity of the items is not the most important issue in this case. Much more important is the discriminant validity of the measure from the antecedents and consequents (not only in a statistical sense, which is easily obtained, but also in a conceptual sense).

With regard to modifications of existing scales, not all modifications are created equal. Measuring product involvement does not require the full 20-item synonym scale originally proposed by Zaichkowsky, or even the shorter 10-item version (Zaichkowsky, 1994); three or four of the better synonyms will do just fine. Adapting a scale to suit a particular product category or changing the response scale from a 5-point to a 7-point format, or even changing the response category labels slightly, will probably not materially influence the reliability or validity of the scale. In contrast, using one of the dimensions of the scale as a measure of the overall construct (e.g., the happiness dimension of the Material Values Scale to measure materialism overall), retaining only a

subset of the most similar items in a scale (and eliminating all reversed items), or using only items that increase the likelihood that a predicted relationship with another construct will be obtained are modifications that should not be tolerated.

Since it is unrealistic to expect that the measure validation procedures that would be optimal in theory can be implemented in a single paper, we call for more dedicated research that reviews and validates existing measurement instruments within specific domains, with the aim of providing recommendations on which scales to use to measure specific constructs. Meta-studies of this kind could start with an inventory of existing scales for a given construct, summarize the extant evidence on the reliability and validity, and evaluate the track record of existing scales in predicting relevant outcomes. Although various scale handbooks are available that list scales suitable for measuring different constructs, they are usually focused on describing the initial evidence on the reliability and validity presented by the developers of the scale. The fact that many scales do not perform well in subsequent evaluations implies that the initial evidence is not always trustworthy. In particular, evidence about the external validity of different scales to measure a certain construct is a particularly relevant piece of information for potential users of a scale.

A good example of the proposed approach is the recent paper by Lange and Dewitte (2019), in which they surveyed different tools for measuring pro-environmental behavior (PEB), including self-report measures, field observations, and laboratory observations. In particular, Lange and Dewitte (2019) reviewed 20 general and 13 domain-specific measures of PEB and report information about the number of items, the dimensionality of the scale, coefficient alpha, example items, and various correlates. The authors also provide recommendations about which measurement tools to use depending on the purpose of the research (e.g., when a researcher is interested in the personality correlates of PEB, Lange and Dewitte recommend a particular self-report measure of PEB or multiple methods, including informant reports, to avoid common method bias).

Our second major point in this commentary is that the perspective on measurement should be broadened to encompass the respondent, in addition to the measurement instrument. Hawes et al. focus solely on the measurement instrument as a source of unreliability and invalidity. This is of course only part of the story because of respondents' goals (accuracy vs. self-presentation), respondents' ability and willingness to answer questions accurately (optimizing vs. satisficing), and threats to the integrity of surveys that can arise during the various stages of the survey process (comprehension, judgment, and response) are also important determinants of the reliability and validity of data collected using measurement instruments (see Baumgartner & Weijters, 2019, for details on these points). In Baumgartner and Weijters (2019), we

distinguished three related but distinct senses in which one can think about measurement: (a) conceptualizing theoretical variables of interest and choosing appropriate observable indicators of the intended construct; (b) collecting the data necessary for an empirical examination of the theoretical issues under study; and (c) constructing a model that relates the data collected in the second step to the latent factors representing the concepts the researcher is interested in (as specified in the first step). The methodological literature tends to emphasize the first sense (measurement as indicator specification and selection) and the third sense (measurement modeling). However, there are good reasons to pay more attention to the second sense of measurement (collecting data from respondents).

Details about important issues related to the second meaning of measurement are provided in Baumgartner and Weijters (2019, chapter 3). Here we would like to draw attention to one of the more pressing concerns related to respondents in this digital age of data collection. On the one hand, researchers tend to formulate and test increasingly complex (and supposedly causal) models in which multiple constructs, either sequentially or in parallel, mediate the effects of a treatment on an outcome, and yet other constructs moderate some of these mediated effects. This necessitates the measurement of numerous, often abstract constructs that must be assessed with multiple items each. But simultaneously, the volume of data needed, the desire to obtain data quickly, resource constraints, and other factors drive researchers to collect data under suboptimal conditions. In particular, markets of sorts have emerged in which respondents “sell” their time to researchers for minimal pay (mediated by platforms such as Amazon Mechanical Turk or Prolific Academic). If a researcher pays an MTurk worker a dollar or two (if even that), constraints are bound to be placed on which scales can be deployed, and the pressure to modify scales is likely high. Professional survey takers who have to complete numerous questionnaires per day to earn even a minimum wage are likely satisficing rather than optimizing, multi-tasking is apparently common, and response quality will be low.

To improve data quality, many papers now contain a section on attention checks. In a recent paper (Baumgartner & Weijters, *in press*), we classified the methods available to identify careless responders along two dimensions (whether dedicated measures have to be included in the questionnaire to assess careless responding or whether careless responding is inferred from the measures of the substantive constructs; whether respondents' tendency to minimize the time and effort spent on the survey is measured relatively directly or inferred from its presumed consequences on data quality) and, based on an empirical study involving 880 students, recommended three methods (a yes/no question asking respondents to state honestly whether they thought their data should be included in the study; instructed

response items asking respondents to select a particular response category on the rating scale or leave the answer blank; and a page time measure to be used for identifying speeders). However, it should be acknowledged that even these recommended measures have weaknesses because some rely on respondents' honesty to admit that they just wasted the researcher's time; professional survey takers know how to deal with instructed response items; and even sophisticated page time measures as proposed in Baumgartner and Weijters (*in press*) may only flag the most flagrant speeders. Instead of eliminating respondents after the fact based on imperfect careless responding checks, would it not make more sense for researchers to select better respondents to begin with? Among other things, this would involve choosing interested respondents completing tasks that do not involve hypothetical scenarios; incentivizing participants properly; avoiding (as much as possible) unproctored studies administered online with no control over the survey setting; and not letting convenient and low-cost access to survey participants dictate the use of improvised and modified measures.

In conclusion, Haws et al. have provided a valuable service to the research community by shining a light on the neglected issue of scale deployment and by reminding researchers to pay greater attention to the selection of validated scales and to validate ad hoc scales or scales that have been modified. However, it is also necessary to be pragmatic and to be cognizant of the fact that researchers will not be able to satisfy the stringent requirements proposed by Haws et al. in a single paper. As a solution to this conundrum, we recommend that methodologically oriented researchers devote more time and effort to analyzing previous applications of scales (with an emphasis on tracking scale deployment in different contexts, instead of simply describing the reliability and validity evidence reported in the initial scale development paper). In addition, we want to remind researchers that reliable and valid measurement is not only a function of the instrument used to measure a construct of interest. The respondents to whom the scale is administered are a very important contributor to the integrity of measurement, and researchers should be concerned not only with the selection of appropriate scales but also with the selection of appropriate respondents.

USING STATE-OF-THE-ART PSYCHOMETRICS TO SUPPORT SCALE DEPLOYMENT

Mo Wang and Chengquan Huang
Warrington College of Business
University of Florida
Gainesville
Florida
USA

Haws, Sample, and Hulland (this issue) provide an integrated review of recently published research in the *Journal of Consumer Psychology*. They categorize four types of scale deployment (i.e., "as-is, validated," "as-is, improvised," "modified, validated," and "modified, improvised" usage) and discuss the possible problems of scale validity in each situation. They also categorize four types of scale modification (i.e., wording modification, length modification, dimension modification, and multiple modifications) and provide a decision tree to guide better scale deployment. We applaud the guidance offered by Haws et al. but are concerned that the article's recommended psychometrics for scale deployment do not reflect state-of-the-art practices. Hence, in this commentary, we provide an update on psychometrical approaches for scale modification and deployment, thereby enriching the readers' methodological toolbox.

The organization of this commentary is as follows. We first discuss the methodologies for content validation, which can be used for supporting all types of scale modification. Next, we discuss the sample size requirement for confirmatory factor analysis (CFA), which is important to consider given that CFA is a useful tool for scale validation. Further, we introduce item response theory (IRT) as a useful psychometric technique for scale length modification. Finally, we discuss measurement equivalence for scale deployment in different cultures and contexts. It should be noted that though we focus on scale modification as Haws et al. (2023), many of these tools can be used for scale validation in the measurement development process.

State-of-the-art content validation procedures

Although all types of validity (e.g., convergent validity, discriminant validity, nomological validity) are important to establish for scale deployment, here we would like to focus on the procedure for establishing content validity. This is because the approaches Haws et al. (2023) suggested for examining "face validity" do not seem quite rigorous for this purpose. While Haws et al. acknowledge that face validity, as "the mere appearance that a measure has validity," is only "one aspect of content validity" (see table 1 in Haws et al.), their recommended procedures leave content validity undiscussed. Indeed, face validity is about whether the items "look like" the construct of interest, whereas content validity is the extent to which the items adequately sample the content universe of a construct. A recent review suggests that content validity is an important but less examined type of validity in applied behavioral and psychological research (Colquitt et al., 2019). Such a problem points to a lack of *content validation* in our research practice.

To improve in this regard, two methods are recommended (Colquitt et al., 2019): the Anderson and Gerbing (1991) approach and the Hinkin and

Tracey (1999) approach. One commonality of these two approaches is that they both ask a group of judges to independently read the definitions of multiple constructs (including the focal construct of interest and some orbiting constructs) and all items measuring these constructs. The choice of orbiting constructs is important for the rigorous examination of item distinctiveness, as they should be conceptually relevant to the focal construct (e.g., if the negative mood is the focal construct, then positive mood can be used as an orbiting construct) but cannot be in a "part-whole" relationship with the focal construct (e.g., anger and negative mood). Another commonality of the two approaches is that they both prefer naïve judges, rather than experts as recommended by Haws et al. (2023). This is because experts may use their expertise to facilitate content judgment, failing to represent the typical responses from a layperson who has no such professional knowledge.

The difference between the two approaches is that Anderson and Gerbing (1991) ask the judges to sort the items into the appropriate construct definition, whereas Hinkin and Tracey (1999) ask the judges to rate how well each scale item corresponds to the construct definition using a Likert-style rating process (e.g., from 1 = not at all to 5 = completely). An advantage of these approaches over Haws et al. (2023) is that they offer psychometric indices to evaluate content validity. With the Anderson and Gerbing (1991) approach, two indices can be calculated: the proportion of substantive agreement (p_{sa}) and the substantive validity coefficient (c_{sv}). With the Hinkin and Tracey (Hinkin & Tracey, 1999) approach, ANOVA can be performed to examine whether items reflect the definition of an intended construct better than other orbiting construct items. Researchers may also calculate the Hinkin Tracey correspondence (htc) and the Hinkin Tracey distinctiveness (htd) coefficients, as suggested by Colquitt et al. (2019).

Sample size consideration for confirmatory factor analyses

We agree with Haws et al. (2023) that confirmatory factor analysis (CFA) is a useful tool for scale validation. However, we would like to point out that their recommended sample size of 50 is unlikely to be appropriate for CFA. Indeed, there is a long history of research on the appropriate sample size for CFA (MacCallum et al., 1999). Some early research tried to offer an absolute minimal sample size (N) and proposed numbers ranging from 100, 200, to 500. However, this approach was fiercely criticized because it ignored that the factor structure and model complexity differ across studies/scales. Another line of research tried to propose a minimum $N:p$ ratio (i.e., the sample size [N] to the number of items [p] ratio), such as 5:1 and 10:1 (MacCallum et al., 1999). A third line of research, focusing on estimation accuracy and efficiency,

tried to offer a minimum $N:q$ ratio (i.e., the ratio of the sample size $[N]$ to the number of parameters that require statistical estimates $[q]$). Kline (2015) suggested that an ideal $N:q$ ratio would be 20:1, and an acceptable $N:q$ ratio would be 10:1.

Compiling these research findings, we offer the following recommendations for the sample size requirement of CFA. First, a large sample size should always be preferred when possible. Second, rather than use an absolute rule of thumb for minimal sample size (e.g., 50 or 100), researchers should use the $N:q$ or $N:p$ rule to facilitate their decision about minimum sample size, considering the complexity of their measurement model. Third, researchers should also consider whether the desired sample size for CFA would satisfy the statistical power requirement for their main analyses. If their main analyses require a larger sample size to meet the statistical power requirement, then this larger sample size should be adopted instead.

Item response theory for supporting length modification

As Haws et al. (2023) point out, researchers may also reduce the length of a measurement scale. They recommended using the confirmatory factor analysis (CFA) and examining the factor loadings to facilitate the decisions of which items to retain or remove. Although their recommended CFA and the traditional correlational approach (i.e., examining the correlation between the original scale and the shortened scale) are useful methods for validating the shortened scale, the use of Item Response Theory (IRT) may offer a more systematic approach for how to shorten an existing scale.

Item response theory is a probabilistic nonlinear modeling technique that can identify each scale item's unique contribution to the measurement of a latent construct. The IRT calculates the respondents' probability of selecting particular response options for each item and estimates each item's ability to differentiate the respondents. Therefore, the IRT can be used to select the most informative items from a long scale to form a shorter version. For example, Huang et al. (2017) reported a study using IRT to shorten a long safety climate scale. Specifically, using IRT, they calculated the percentage of the total information that each item contributed to capturing the underlying measurement construct. They found that for the original 16-item organizational safety climate scale, the most informative eight items retained 56.94% of the total test information, and the four most informative items retained 30.29% of the total test information. These results can thus help the researchers to decide which items to retain and how many items to retain when shortening the length of a scale. Interested readers can refer to Lang and Tay (2021) for more details about IRT.

Measurement equivalence consideration for “as-is, validated” usage

So far, our discussion of psychometric approaches has focused on supporting scale modification. However, previously validated scales (i.e., the “as-is, validated” usage) may also need further support for deployment in different contexts, especially in different cultures or demographic groups. As Haws et al. (2023) point out, researchers may deploy a measurement scale in a different culture where the norms and meanings of items may shift. To ensure that an instrument captures the same theoretical construct of interest in a different culture, researchers may examine the *measurement equivalence* of their scale in different cultures. Here we offer a reader-friendly overview of what measurement equivalence is and how researchers can examine it.

Measurement equivalence examines whether the scale measures the same construct in the same way across different groups of participants. As Wang and Russell (2005, p. 710) mentioned, “an instrument yields cross-cultural measurement equivalence if individuals across different cultures with identical latent construct scores also have the same expected raw scores at the item level, the total score level, or both.” It should be noted that measurement equivalence does not mean there are no cross-cultural differences at the population level. Instead, it means that respondents from different cultural groups who have the same score for the latent construct should have similar responses to the scale. Measurement equivalence is worth researchers' attention even if the scale has been carefully translated (e.g., using the translation and back-translation procedure of Brislin, 1980). This is because culture may shift people's understanding of the items and thus lead to measurement biases (Davidov et al., 2014). For example, social desirability biases have been found to be more prevalent in collectivist countries; familiarity with stimuli across cultures may also impact the respondents' ratings (Davidov et al., 2014).

Measurement equivalence researchers have proposed to test measurement equivalence in a hierarchical manner via CFA. The first level of equivalence—*configural equivalence*—requires the factor structure to be the same across compared groups (e.g., different cultures). For example, if a measure is unidimensional in its original culture, its factor structure should also be unidimensional when deployed in another culture. Configural equivalence is often the first step in the examination of measurement equivalence. The second level of equivalence is *metric equivalence*. Metric equivalence requires the factor loadings of the items to be equal for the compared groups (e.g., the factor loading of item #1 in culture A should be the same as that in culture B). If metric equivalence is satisfied, it indicates that one unit increase of the scale in culture A has the same meaning as one unit increase of the scale in culture B. The third level of

equivalence is *scalar equivalence*, which requires not only the factor loadings but also the intercepts of the items to be equal for compared groups. This warrants that any observed mean differences in the scale scores between the compared groups reflect “apple-to-apple” comparisons between the two groups.

A useful tool to examine measurement equivalence is multi-group confirmatory factor analysis (MGCFA). The MGCFA framework is useful to test all of the three types of measurement equivalence because researchers can set parameter constraints to the model (e.g., fixing the factor loadings of two groups to be equal to assess metric equivalence), examine the model fit indices, and compare the model fit indices to determine whether measurement equivalence is achieved. Interested readers can refer to Davidov et al. (2014) and Vandenberg and Lance (2000) for detailed tutorials, and Wang and Russell (2005) for empirical reference.

The MGCFA mentioned above is mainly for the case when researchers can collect data from two contexts to compare. However, researchers can also examine measurement equivalence even if they only have data for their context of scale deployment. To examine configural equivalence, researchers can use CFA to evaluate the intended factor structure and examine whether the model fit is acceptable; to examine metric equivalence and scalar equivalence, researchers can compare item factor loadings and intercepts with those reported in published articles (e.g., the original scale development studies) and assess similarities. Taken together, testing measurement equivalence with CFA is a useful approach for supporting the “as-is” usage of a validated scale in a different context.

Conclusion

Haws et al. (2023) represent a valuable attempt to tackle the measurement validity problems of scale deployment. To supplement their proposed procedures of validating scales in such circumstances, we review state-of-the-art psychometric approaches that can help researchers establish construct validity when they modify scales in research or use scales in a different culture. It is our hope that future researchers can utilize these psychometric tools to rigorously support their scientific inquiry.

ORCID

C. Miguel Brendl  <https://orcid.org/0000-0002-9111-2491>
Bobby J. Calder  <https://orcid.org/0000-0002-0120-4641>
Bert Weijters  <https://orcid.org/0000-0002-8590-0088>

REFERENCES

- Aaker, J. L., Benet-Martinez, V., & Garolera, J. (2001). Consumption symbols as carriers of culture: A study of Japanese and Spanish brand personality constructs. *Journal of Personality and Social Psychology, 81*(3), 492–508.
- Aguinis, H., Ramani, R. S., & Cascio, W. F. (2020). Methodological practices in international business research: An after-action review of challenges and solutions. *Journal of International Business Studies, 51*(9), 1593–1608.
- Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science, 83*(5), 1098–1109.
- Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*(5), 732–740. <https://doi.org/10.1037/0021-9010.76.5.732>
- Baumgartner, H., & Weijters, B. (2019). Measurement in marketing. *Foundations and Trends in Marketing, 12*(4), 278–400. <https://doi.org/10.1561/17000000058>
- Baumgartner, H., & Weijters, B. (in press). How to identify careless responders in surveys. In H. Baumgartner & B. Weijters (Eds.), *Measurement in marketing, vol. 19 of review of marketing research*. Emerald Publishing.
- Böckenholt, U., & Lehmann, D. R. (2015). On the limits of research rigidity: The number of items in a scale. *Marketing Letters, 26*(3), 257–260.
- Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (pp. 349–444). Allyn & Bacon.
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307.
- Calder, B. J., Brendl, C. M., Tybout, A. M., & Sternthal, B. (2021). Distinguishing constructs from variables in designing research. *Journal of Consumer Research, 3*(1), 188–208.
- Clifton, J. D. (2020). Managing validity versus reliability trade-offs in scale-building decisions. *Psychological Methods, 25*(3), 259–270.
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(10), 1243–1265. <https://doi.org/10.1037/apl0000406>
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75. <https://doi.org/10.1146/annurev-soc-071913-043137>
- de Jong, M. G., Steenkamp, J.-B. E. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-term marketing scales. *Marketing Science, 28*(4), 674–689. <https://doi.org/10.1287/mksc.1080.0439>
- DeJong, M. G., Lehmann, D. R., & Netzer, O. (2012). State-dependence effects in surveys. *Marketing Science, 31*(5), 838–854.
- Douglas, S. P., & Nijssen, E. J. (2003). On the use of “borrowed” scales in cross-national research: A cautionary note. *International Marketing Review, 20*(6), 621–642.
- Evans, K. R., Landry, T. D., Li, P.-C., & Zou, S. (2007). How sales controls affect job-related outcomes: The role of organizational sales-related psychological climate perceptions. *Journal of the Academy of Marketing Science, 35*(3), 445–459.
- Fang, E., Evans, K. R., & Landry, T. D. (2005). Control systems' effect on attributional processes and sales outcomes: A cybernetic information processing perspective. *Journal of the Academy of Marketing Science, 33*(4), 553–574.
- Farley, J. U., Lehmann, D. R., & Ryan, M. J. (1981). Generalizing from ‘imperfect’ replication. *Journal of Business, 54*, 597–610.
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465.
- Haws, K., Hulland, J., & Sample, K. (2023). Scale use and abuse: Towards best practices in the deployment of scales. *Journal of Consumer Psychology, 33*(1), 226–243.

- Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods*, 2(2), 175–186. <https://doi.org/10.1177/109442819922004>
- Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1014>
- Huang, Y. H., Lee, J., Chen, Z., Perry, M., Cheung, J. H., & Wang, M. (2017). An item-response theory approach to safety climate measurement: The liberty mutual safety climate short scales. *Accident Analysis & Prevention*, 103, 96–104. <https://doi.org/10.1016/j.aap.2017.03.015>
- Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, 16(2), 131–152.
- Katsikeas, C. S., Auh, S., Spyropoulou, S., & Menguc, B. (2018). Unpacking the relationship between sales control and salesperson performance: A regulatory fit perspective. *Journal of Marketing*, 82(3), 45–69.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
- Kohli, A. K., Shervani, T. A., & Challagalla, G. N. (1998). Learning and performance orientation of salespeople: The role of supervisors. *Journal of Marketing Research*, 35, 263–274.
- Kopalle, P. K., & Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes*, 70, 189–197.
- Lang, J. W., & Tay, L. (2021). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, 8, 311–338. <https://doi.org/10.1146/annurev-orgpsych-012420-061705>
- Lange, F., & Dewitte, S. (2019). Measuring pro-environmental behavior: Review and recommendations. *Journal of Environmental Psychology*, 63(6), 92–100.
- Lynch, J. G., Jr., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing*, 32(4), 333–342.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84–99. <https://doi.org/10.1037/1082-989X.4.1.84>
- Madan, S., Savani, K., & Katsikeas, C. S. (2022). Privacy please: Power distance and people's responses to data breaches across countries. *Journal of International Business Studies*, 1–24. <https://doi.org/10.1057/s41267-022-00519-5>
- Malhotra, N. K., Mukhopadhyay, S., Liu, X., & Dash, S. (2012). One, few or many? An integrated framework for identifying the items in measurement scales. *International Journal of Market Research*, 54(6), 835–862.
- Miao, C. F., & Evans, K. R. (2013). The interactive effects of sales control systems on salesperson performance: A job demands–resources perspective. *Journal of the Academy of Marketing Science*, 41(1), 73–90.
- Plank, M. (1949, September 30). The meaning and limits of exact science. *Science*, 110(2857), 325.
- Popper, K. (2004). *The logic of scientific discovery*. Routledge. (Original work published in Gein 1935)
- Sarin, S., Challagalla, G., & Kohli, A. K. (2012). Implementing changes in marketing strategy: The role of perceived outcome- and process-oriented supervisory actions. *Journal of Marketing Research*, 49, 564–580.
- Scopelliti, I., Vosgerau, J., & Huh, Y. E. (2020). Response to commentaries on the exerting self-control ≠ sacrificing pleasure research dialogue. *Journal of Consumer Psychology*, 30(1), 215–216. <https://doi.org/10.1002/jcpy.1141>
- Shimp, T. A., & Sharma, S. (1987). Consumer ethnocentrism: Construction and validation of the CETSCALE. *Journal of Marketing Research*, 24(3), 280–289.
- Singelis, T. M. (1994). The measurement of independent and interdependent self-construal. *Personality and Social Psychology Bulletin*, 20(5), 580–591.
- Supphellen, M., & Rittenburg, T. L. (2001). Consumer ethnocentrism when foreign products are better. *Psychology & Marketing*, 18(9), 907–927.
- Szabo, S., Orley, J., & Saxena, S. (1997). An approach to response scale development for cross-cultural questionnaires. *European Psychologist*, 2(3), 270–276.
- Triandis, H. C., Chan, D. K. S., Bhawuk, D. P., Iwao, S., & Sinha, J. B. (1995). Multimethod probes of allocentrism and idiocentrism. *International Journal of Psychology*, 30(4), 461–480.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70. <https://doi.org/10.1177/109442810031002>
- Vazire, S., Schiavonne, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162–168.
- Wang, M., & Russell, S. S. (2005). Measurement equivalence of the job descriptive index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educational and Psychological Measurement*, 65(4), 709–732. <https://doi.org/10.1177/0013164404272494>
- Yoo, B., Donthu, N., & Lenartowicz, T. (2011). Measuring Hofstede's five dimensions of cultural values at the individual level: Development and validation of CVSCALE. *Journal of International Consumer Marketing*, 23(3–4), 193–210.
- Zaichkowsky, J. L. (1994). The personal involvement inventory: Reduction, revision, and application to advertising. *Journal of Advertising*, 23(4), 59–70.

How to cite this article: Katsikeas, C. S., Madan, S., Brendl, C. M., Calder, B. J., Lehmann, D. R., Baumgartner, H., Weijters, B., Wang, M., Huang, C., & Huber, J. (2023). Commentaries on “Scale use and abuse: Toward best practices in the deployment of scales”. *Journal of Consumer Psychology*, 33, 244–258. <https://doi.org/10.1002/jcpy.1319>