Organizational Research Methods
2023, Vol. 26(4) 574–607
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10944281221115374
journals.sagepub.com/home/orm
Sage

Construct Development and Validation in Three Practical Steps: Recommendations for Reviewers, Editors, and Authors*

Lisa Schurer Lambert[1] **and Daniel A. Newman**[2]

Abstract

We review contemporary best practice for developing and validating measures of constructs in the organizational sciences. The three basic steps in scale development are: (a) construct definition, (b) choosing operationalizations that match the construct definition, and (c) obtaining empirical evidence to confirm construct validity. While summarizing this 3-step process [i.e., Define-Operationalize-Confirm], we address many issues in establishing construct validity and provide a checklist for journal reviewers and authors when evaluating the validity of measures used in organizational research. Among other points, we pay special attention to construct conceptualization, acknowledging existing constructs, improving existing measures, multidimensional constructs, macro-level constructs, and the need for independent samples to confirm construct validity and measurement equivalence across subpopulations.

Keywords

construct development, construct measurement, scale development, construct validity

The accuracy of tests of relationships between constructs rests on the foundation of sound construct development and measurement (Edwards, 2003; Nunnally & Bernstein, 1994; Schwab, 1980). Without evidence that measures represent their intended constructs, researchers run the risk that tests of theoretical relationships are biased, misleading, or simply wrong. We articulate contemporary

[1] Department of Management, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA
[2] School of Labor & Employment Relations, Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

*This manuscript was accepted by Editor-in-Chief Paul Bliese.

Corresponding Author:
Lisa Schurer Lambert, Department of Management, Spears School of Business, Oklahoma State University, Stillwater, Oklahoma, USA
Email: lisa.schurer.lambert@okstate.edu

best practices for selecting, developing, revising, and validating measures of constructs in the organizational sciences.

The activities involved in scale development can be organized into three general steps: (1) construct conceptualization, (2) operationalizing the construct, and (3) assessing evidence to confirm construct validity. These three steps capture the essence of approaches articulated in previous reviews that build upon the seminal work of Loevinger (Clark & Watson, 1995; 2019; DeVellis, 2003; Hinkin, 1998; Loevinger, 1957; MacKenzie, Podsakoff, & Podsakoff, 2011). These sets of authors, drawing from a common base of knowledge and practice, have described similar approaches to construct development. The three steps in construct development also apply (in an abbreviated way) when revising measures and when using existing measures. While summarizing this three-step process, we pay special attention to construct conceptualization and mapping operationalizations of constructs to their definitions. We also address issues of establishing construct validity for micro-level, macro-level, multilevel, and multidimensional constructs; measurement equivalence across time and subpopulations; and the need for independent samples to confirm construct validity.

The literature on construct development is large with many brilliant insights and considerable practical advice, and we direct readers to delve into the papers we cite instead of relying solely on this paper. We also acknowledge that researchers differ in their opinions and stances on a number of issues related to construct development. These divergences of opinion may be rooted in philosophical positions, or custom, but we urge researchers to develop their own principled stand as they specify their measurement models, collect and assemble empirical evidence to test their measurement theory, and present it to academic peers for inspection – just as we do for theories of relationships between constructs.

How researchers practice construct definition and operationalization, and then confirm validity, creates a foundation for accumulating knowledge in organizational science. Construct validity—i.e., the extent to which we are measuring what we believe we are measuring—is a *sine qua non* of organizational research, and knowledge of methods for establishing construct validity is therefore indispensable. Our recommendations are also presented in the form of a checklist that embodies the principles we discuss, which we hope succinctly captures the state of the art for reviewers, editors, and authors. Before we begin, it is helpful to direct the reader to the glossary of construct validity terms in the appendix.

Constructs are conceptual phenomena that facilitate our understanding of the world and how it operates. Thus, the nature of constructs varies substantially not only across disciplines (e.g., organizational behavior, strategic management) but from construct to construct. Constructs may differ considerably in the logical arguments that justify their conceptualization, in how they are manifested, and in the evidence that supports their plausibility. Despite the variety of constructs, there are common principles underlying the three steps in conceptualizing, operationalizing, and using evidence to confirm their validity.

Although the steps of conceptualization, operationalization, and confirmatory evidence are presented as a sequence, in practice these steps may not be followed in a strictly sequential fashion and are often iterative. For example, researchers may use available empirical evidence to clarify a definition and to revise items/indicators in the measurement model. Nor should the steps and procedures we describe be slavishly followed, because constructs may require deviations from our specific recommendations in order to adhere to the principles we espouse. Following a checklist is not prima facie evidence that a construct is well measured. We also remind readers that construct validity requires a unitary approach to establishing that the scores for a construct are valid for the intended purpose (e.g., theoretical research, decision making) rather than ticking off a list of types of validity. Whereas it may be useful to discuss evidence regarding content, construct, and criterion validity as these represent facets of validity, confirming construct validity requires a holistic assessment of the assembled evidence (American Education Research Association, 2014; Landy, 1986). The task is to construct a logical, theoretical argument for a construct, articulate the relationships between the construct and its measures, and obtain empirical evidence to support the plausibility of the hypothesized measurement model. The result should be a theoretical and empirical case for the measurement model that is convincing to a skeptical academic audience.

Step 1: Construct Conceptualization and Definition

A construct is an abstraction that helps us makes sense of our environment and is a useful aid to developing theories about relationships. Only by naming these abstractions as constructs (e.g., job satisfaction, organizational performance) can we theorize about relationships between them. We take the position that the phenomenon underlying the construct is real, even if our definition and understanding of it are flawed (Cronbach & Meehl, 1955; Messick, 1981). Construct definitions should correspond to the underlying phenomenon, and should distinguish not only what the construct is, but also what it isn't. Definitions should also clarify how a construct is different from, and similar to, other constructs. Constructs may be defined narrowly or broadly. For example, overall conscientiousness is a broad construct, whereas industriousness is a narrow construct or facet of conscientiousness. Likewise, corporate visibility is a slice of the wider construct of corporate reputation.

Review the Literature Thoroughly

The first step in conceptualizing a construct should be to review the relevant research literature to see if the construct is in use, perhaps under another name, or if multiple constructs with the same name have different definitions. Organizational science is vast and diverse with hundreds of existing constructs, and we recommend that researchers consider the extent to which their target construct is redundant with, or distinct from, the well-known and impactful construct domains in the field. When introducing a new construct, authors should take care to acknowledge whether the alleged new construct might be a relabeling or recombination of content sampled from other domains.

The enterprise of academic research presents powerful incentives for researchers to ignore existing constructs and to pretend that relabeled or reshuffled constructs are novel. Kelley's (1927) *jangle fallacy* occurs when two different construct names are used for the same phenomenon (or two labels for the same construct). As evidence of the jangle fallacy in strategy research, three measures (i.e., R&D intensity, patent counts, patent citations) have been used to assess a diverse array of constructs and construct labels (Ketchen, Ireland, & Baker, 2013), using multiple construct names for the same phenomenon (e.g., using patent counts to measure three distinct constructs: innovative productivity, knowledge stock, and technology expertise). Further, recent examples of highly-cited constructs and construct labels that have essentially ignored or downplayed their redundancy with pre-existing constructs can be found in the areas of grit [e.g., grit is correlated $r_{corrected} = .84$ with conscientiousness; (Credé, Tynan, & Harms, 2017) and work engagement (work engagement has high content overlap and is correlated $r_{corrected} = .77$ with a combination of job involvement, job satisfaction, and organizational commitment; Newman, Joseph, & Hulin, 2010)]. To address the jangle fallacy, Newman, Harrison, Carpenter, and Rariden (2016) surveyed the management literature and the editorial board of the *Academy of Management Journal* to enumerate seven cardinal construct domains in the field of OB/HR (i.e., general mental ability, core self-evaluations, overall job attitude, social exchange quality, behavioral work engagement, job complexity, and leader individualized consideration); noting these seven established construct domains have often been resampled and relabeled; and recommending future authors explicitly acknowledge these existing constructs rather than relabeling them.

Next, Kelley's (1927) *jingle fallacy* occurs when two different constructs are given the same name. For instance, the construct name 'strategic consensus' has been defined and measured in distinctly varied ways (Kellermanns, Walter, Lechner, & Floyd, 2005); and the construct name 'emotional intelligence' has also been used to refer to many different constructs (Mayer, Roberts, & Barsade, 2008). The problems of construct proliferation, empirical redundancy (Le, Schmidt, Harter, & Lauver, 2010; Schwab, 1980), relabeling of existing constructs (i.e., the jangle fallacy/ false differentiation; Kelley, 1927; Cardinal, Sitkin, & Long, 2010), and ignoring distinctions among constructs that are different (jingle fallacy) all contribute to confusion, lack of parsimony, and inefficiency in scientific research because of inadequate construct conceptualization.

Unfortunately, published research tends to minimize, or draw attention away from, measurement problems. Close inspection may reveal a variety of difficulties, including: inconsistencies in definitions, a lack of correspondence between definitions and operationalizations, weak discrimination from related constructs, troublesome items/indicators, or lack of substantial relationships when there is strong theory predicting such relationships. For example, in the area of leadership, items have been confounded with outcomes, supposedly distinct types of leadership have conceptual overlap, and similar items are used to measure different types of leadership (Shaffer, DeGeest, & Li, 2016; Van Knippenberg & Sitkin, 2013).

Traditional advice recommends using established measures whenever possible, yet problems with constructs and their operationalizations introduce bias into estimates of relationships between constructs. For example, the organizational commitment questionnaire (Mowday, Steers, & Porter, 1979) is a well-established measure, but has been criticized for being contaminated with items measuring turnover intentions, leading to upward bias in the correlation between organizational commitment and turnover (Bozeman & Perrewé, 2001). This raises the conceptual issue of whether turnover intentions should become part of the construct definition of organizational commitment (Klein, Molloy, & Brinsfield, 2012). When the literature review reveals inadequacies, ambiguities, or conflicts surrounding construct definitions, these issues must be resolved before proceeding. We argue that perpetuating poor measurement, in the face of documented weaknesses, does not advance science. Instead, we urge scholars to contribute toward building a solid foundation for generating knowledge by taking advantage of the opportunity to revise existing measures or develop new measures. In some instances, it may be necessary to define and develop a new construct. For example, the construct of firm risk taking has been conceived of as research and development (R&D) spending, and also as R&D intensity (spending to sales); but Bromiley, Rau and Zhang (2017) develop conceptual arguments for distinguishing between spending and intensity and offer supporting empirical evidence of the distinction. In another example, Brady, Brown and Liang (2017) expanded the concept of workplace gossip away from a narrow deviance perspective, to also include positive (and potentially prosocial) evaluative talk about another person who is not present.

Fortunate researchers will find that their focal construct is already adequately defined in past work. In many cases, prior definitions, accompanied by empirical validity evidence, may be sufficient to indicate that researchers can proceed with already-published scales or operationalizations. However, the mere existence of published definitions and scales is not sufficient to preclude the need for local construct validity evidence.

Conceptualizing a construct may lead to one of three decisions: adopting an existing construct and measurement, revising an existing construct and/or measurement, or developing a new construct and measurement. Our recommendations vary depending on the decision to adopt, revise, or newly develop a construct and measure. Regardless of this choice, the construct definition is arguably the central element of a construct conceptualization.

Formally Define the Construct: Characteristics of Good Construct Definitions

In a monumental work on developing construct definitions, Podsakoff, MacKenzie, and Podsakoff (2016) provide a list of issues reviewers and authors should consider when evaluating a construct definition. To our understanding, good definitions (a) clarify the type of property (e.g., feelings, perceptions, beliefs, behavior, or performance metrics) the construct represents, (b) clarify the entity/level of analysis (e.g., individual, group, organization, task, or event) to which the property applies, (c) note the construct's essential and unique attributes, or the attributes shared by cases of the concept, and (d) specify the dimensionality of the construct. In addition to the formal construct definition, construct conceptualization also involves detailing how the construct relates to existing constructs. It is nonetheless important to avoid circularity in the definition, as one should not embed antecedents or consequences in the definition (e.g., the construct work withdrawal should not be defined as a response to a dissatisfying job situation; rather the construct should be defined independently, and its antecedents should be empirically studied, rather than assumed and embedded in the construct definition itself). Constructs may be defined narrowly or broadly; narrow definitions are useful for fine-grained constructs (e.g., satisfaction with coworkers), whereas broad constructs (e.g., overall job satisfaction) may be more useful for theorizing at a more abstract or general level. The definition also should indicate whether the construct is stable or variable (e.g., over time, culture, organizational membership). The construct should be defined across its full expected range (Tay & Jebb, 2018). That is, is the low end of a construct defined as an absence of, or as a negative of the construct? (e.g., work engagement has sometimes been defined as the opposite of burnout, whereas positive affectivity and negative affectivity are two distinct constructs and not opposites of each other—as such, positive affectivity should be defined while specifying that its low end is not the same as negative affectivity).

Articulating the Nomological Net of the Target Construct

Constructs exist in a network of related constructs: antecedent causes, outcomes/criteria, and other variables that may be related to the target because they share a common cause. Theoretically predicted relationships between the target construct and other constructs clarify the nature of the target construct, and when tested serve to offer evidence of nomological validity. Literature review will indicate whether the nomological network was well specified and tested for existing constructs, but we recommend that researchers undertake this task anew for substantially revised and original construct measures. To increase the value of this step, we encourage researchers to be more precise in expressing both the direction (positive or negative) and the strength (weak, moderate, strong; or range) of relationships among constructs in the nomological network (Edwards & Berry, 2010). To elaborate, predicting that the relationship between a target construct and another construct will not only be positive and significant, but further that it will fall within a specified range, or that it will be of greater magnitude than the relationship with a different construct, will facilitate understanding of the target construct itself. For example, Shipp and colleagues specified both the direction and effect size magnitude for predicted relationships between their measure of temporal focus and other related constructs (Shipp, Edwards, & Lambert, 2009).

Documenting the Process and Evidence

The process for developing a revised or new definition must be adequately described. Researchers should identify the literature that was searched, include the steps used to revise and refine the attributes and definition of the construct (e.g., consulting dictionaries), and describe the role of subject matter experts, practitioner experts, or focus groups used to clarify matters central to the definition (Podsakoff et al., 2016). For example, O'Neill and Rothbard (2017) relied on extensive interviews with firefighters to develop the constructs of companionate love and joviality among coworkers.

When the choice is to adopt an existing construct, the definition must be reported with appropriate citations to critical past work on the construct. When researchers choose to revise a construct or to

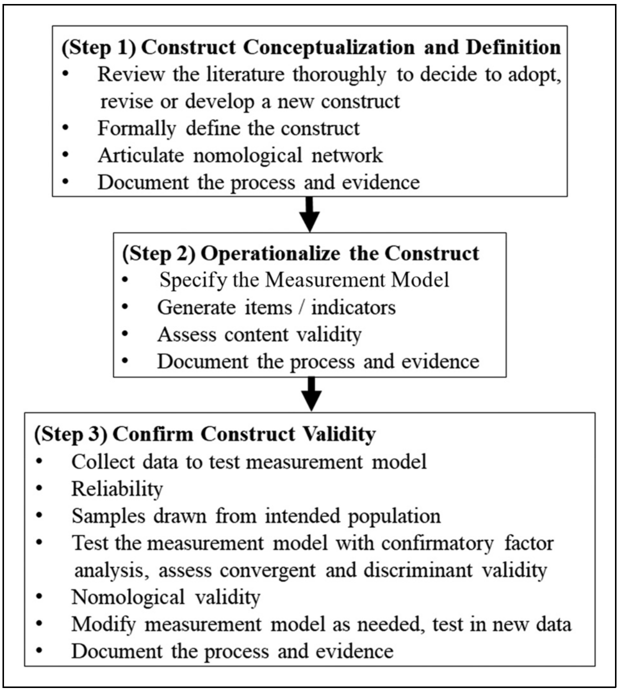
(Step 1) Construct Conceptualization and Definition
• Review the literature thoroughly to decide to adopt, revise or develop a new construct
• Formally define the construct
• Articulate nomological network
• Document the process and evidence
(Step 2) Operationalize the Construct
• Specify the Measurement Model
• Generate items / indicators
• Assess content validity
• Document the process and evidence
(Step 3) Confirm Construct Validity
• Collect data to test measurement model
• Reliability
• Samples drawn from intended population
• Test the measurement model with confirmatory factor analysis, assess convergent and discriminant validity
• Nomological validity
• Modify measurement model as needed, test in new data
• Document the process and evidence

Figure 1. The general steps of construct development.

develop a new construct, they must also fully describe and document the conceptualization. Keep in mind that the definition of the construct comes before, and may be independent of, decisions on how the construct will be measured. Figure 1 lists the steps associated with conceptualizing a construct. Table 1 is a checklist, in which we distinguish between information necessary when using an existing construct definition, vs. choosing to revise or develop a new construct.

Step 2: Operationalizing the Construct

Constructs are not directly observable. To infer the presence or degree/amount of a construct, we rely on signals of the construct as expressed in its items/indicators (e.g., items in a survey, responses to questions in an interview, accounting numbers associated with a firm's activity). The relationships between a construct definition and its' items/indicator(s) constitute a measurement theory, and it is up to researchers to articulate the theoretical logic linking constructs to the indicators of the construct. The operationalization of a construct must represent the definition of the construct (i.e., content validity). Also, the relationship between a construct and its items/indicator(s) must be theoretically described, because this specification will ultimately guide the development and selection of measures, as well as the process and standards for later confirming construct validity.

The source of the data does not define the relationship between the item/indicator and the construct. Indicators of a construct may come from self-reports, scores on word puzzles, others' reports of a target, or informed respondents (e.g., CFO or HR giving organization-level information). Counts and ratings of events, including behavioral observations in situ from videos, may be appropriate. Sources of archival information can include financial records, observable characteristics of groups or organizations, scores from content coding of emails or public speeches, web page material, and transcripts of presentations to analysts, among many others.

Table 1. Checklist for Construct Development and Validation.

	Existing Construct	Revised Construct	New Construct	Key Methodological Sources
Step 1: Construct Conceptualization and Definition				
• Review Literature Thoroughly: Are key conceptual papers or related constructs ignored?	x	x	x	Clark & Watson, 1995; 2019; Cronbach & Meehl, 1995; Edwards 2003; Loevinger, 1957; Newman et al., 2016; Pedhazur & Schmelkin, 1991; Podsakoff, MacKenzie & Podsakoff 2016; Schwab, 2005
• Formally Define the Construct: Characteristics of Good Construct Definitions	x	x	x	
Is the definition present, cited, and explained? a) clarify the type of property (e.g., attitude, intention, performance), b) clarify the entity/level of analysis (e.g., individual, team, firm, event), c) note the construct's essential and unique attributes, d) specify the dimensionality of the construct, (e) definition not circular (e.g., presumed antecedents and consequences not embedded in the definition), f) specify whether construct is stable or varying (e.g., trait vs. state), h) defined across the full range of the construct.	x	x	x	
• Articulate the Nomological Network of the Target Construct: Were relationships with several external constructs predicted in terms of strength and direction? (e.g., "strongly positively associated with", "weakly negatively associated with")		x	x	
• Document the Process and Evidence: Is the process for revising or developing the definition adequately documented? (e.g., citations, consulting with experts, focus groups/interviews)		x	x	
Step 2: Operationalizing the Construct				
• Specify Measurement Models: Was the measurement model clearly articulated? a) unidimensional vs. multidimensional vs. hierarchical?, b) if multidimensional, are factors correlated, and how?, c) which indicators correspond to which factors?, d) uniquenesses uncorrelated?, e) Are the levels of analysis of the construct and levels of measurement clearly identified (e.g., group-level leadership or climate, measured via individual perceptions)?	x	x	x	Anderson & Gerbing, 1991; Bradburn, Sudman, & Wansink, 2004; Clark & Watson, 1995; 2019; Colquitt et al., 2019; DeVellis, 2017; Edwards & Bagozzi, 2000; Heggestad et al., 2019; Hinkin & Tracey, 1999; Krosnick & Presser, 2010; Nunnally, 1970; Tourangeau, Rips, & Rasinski, 2000; Willis, 2005

(*continued*)

Table 1. (continued)

	Existing Construct	Revised Construct	New Construct	Key Methodological Sources
• Generate Items or Indicators: a) simple language (avoid jargon, slang, difficult vocabulary, ambiguous terms), b) not double-barreled, c) specify whether items generated inductively vs. deductively, d) minimum of 3 items per construct, e) appropriate response scale (e.g., frequency vs. amount vs. agreement scales)		x	x	
• Assess Content Validity: Which systematic approaches were used to demonstrate content validity (e.g., expert or naïve ratings, or cognitive interviewing, to assess correspondence between items/indicators and the intended construct definitions, distinctiveness from unintended construct definitions, items/indicators representative of entire construct domain)?		x	x	
• Documenting the Process and Evidence: Were the following reported: indicator(s), instructions to participants, procedures for administration, scoring guidelines, response scale, relevant citations? Was the rationale for revising a measure reported?	x	x	x	
Step 3: Evidence to Confirm Construct Validity				
• Collect Data to Test Measurement Model: Were samples and procedures described?	x	x	x	Anderson & Gerbing, 1988; Brown, 2015; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Hinkin, 1998; Jackson, Gillapsy & Purc-Stephenson, 2009; Lance & Vandenberg, 2002; Pedhazur & Schmelkin, 1991
• Reliability: Was the appropriate reliability (α or ω) information reported?	x	x	x	
• Test the Measurement Model with Confirmatory Factor Analysis: Were the desirable features present? (a) convergent validity, (b) discriminant validity, (c) simple structure, (d) no correlated uniquenesses (residuals)	x x	x x	x x	
• Was nomological validity tested?		x	x	
• Modifying the Measurement Model:				

(continued)

Table 1. (continued)

	Existing Construct	Revised Construct	New Construct	Key Methodological Sources
If any data-driven modifications, were they tested on new data?	x	x	x	
Was parceling avoided when testing the measurement model?	x	x	x	
Avoid Exploratory Factor Analysis (EFA), unless either: a CFA has already been attempted and failed, or a confession is made that the author does not know what they are trying to measure. When using EFA, do not use Principal Components Analysis, and avoid orthogonal rotations.	x	x	x	
• Documenting the Process and Evidence:				
Were reporting standards for CFA followed? [e.g., software version and estimation, how missing data were handled, table of means and correlations, fit indices, standardized factor loadings and factor intercorrelations (or at least the average and range of these), diagnostics of model misfit for future confirmation]?	x	x	x	
Construct validity without CFA? (e.g., necessary for single-indicator measurement, text analysis, etc.): Was construct validity evidence presented (e.g., convergent and discriminant validity, simple structure, nomological validity)?		x	x	
Other Considerations				
• Measurement Equivalence: If comparing groups or comparing across time, was measurement equivalence shown?	x	x	x	Vandenberg & Lance, 2000
• Reflective vs. Formative Indicators: Most measurement models should use reflective indicators, not formative indicators. (Is it reasonable to assume that there is no measurement error in the indicators? Was a strong theoretical rationale provided for selecting the reflective indicators and constructs that were used for identifying the formative constructs [as tested with a MIMIC model or identified via reflective constructs, respectively]?)	x	x	x	Edwards, 2001; Mackenzie, Podsakoff & Jarvis, 2005

(continued)

Table 1. (continued)

	Existing Construct	Revised Construct	New Construct	Key Methodological Sources
• Forced Choice/Ranked Measures: Unless using IRT, this approach should be avoided.	x	x	x	Meade, 2004; Brown & Maydeu-Olivares, 2013
• Single Item/Indicator Measures: Is there strong theoretical logic linking the construct and the indicator? Were content validity and nomological validity assessed?				Shadish, Cook & Campbell, 2002
• Levels of Analysis: Is the level of analysis of each construct consistent with the measurement? Were appropriate conceptual and statistical procedures followed (e.g., multilevel CFA, within group reliability)?	x	x	x	Muthen, 1994; Tay, Woo, & Vermunt, 2014; Bliese, 2000; Chan, 1998b
• Method Variance: If the same construct can be measured via multiple methods or multiple sources, were trait variance and method variance appropriately modeled?	x	x	x	Podsakoff et al., 2003; Williams & McGonagle, 2016
• Algebraically Combined Measures: Difference scores, ratios, and other indices that conflate the relationships of distinct constructs with external variables should be avoided when possible.	x	x	x	Certo et al., 2020; Edwards, 2002

When a review of relevant research (Step 1: Define) reveals that a construct has existing measure(s) that have been theoretically justified, explicitly articulated in a measurement model, and the measurement model has been empirically confirmed; then it makes sense to adopt an existing measure. Using an existing measure facilitates comparing results across studies and thus enables the creation of new knowledge that can be integrated with past research. Revising an existing operationalization may be necessary when evidence indicates prior problems or inconsistencies (e.g., item content does not match definition, construct deficiency/measures do not capture all aspects of the construct, construct contamination/measures capture surplus content that is not part of the construct definition). Operationalizing the construct entails specifying the measurement model.

Specify Measurement Models

We believe that construct definitions (Step 1) and operationalizations (Step 2) can be more easily confirmed via construct validity methods (i.e., see Step 3 below) if the researcher begins with the outcome in mind, and then thinks backward from that outcome. To this end, we note there are several types of relationships between indicators and constructs, which can be specified in various types of measurement models (for examples see Figure 2).

Figure 2a shows a measurement model that is unidimensional, with one construct and multiple measures/items/operationalizations/indicators. For example, the construct might be job satisfaction, and the indicators might be survey items from the Brayfield and Rothe (1951) overall job satisfaction scale. Job satisfaction would be considered a latent construct (not directly observed, but rather inferred from each person's scores on the measures/observable items/manifest indicators). The latent job satisfaction construct causes or gives rise to the observed scores on the items/indicators, which is why the measurement model is drawn with arrows pointing from the construct to its

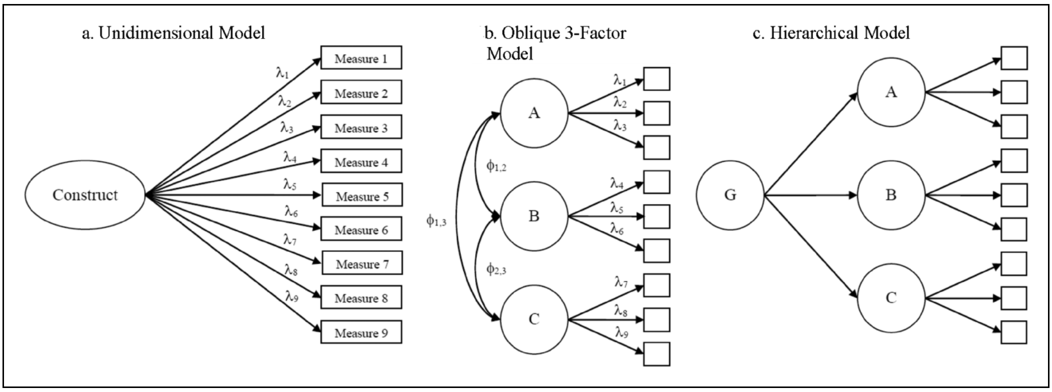

Figure 2. Examples of Measurement Models. **2a.** Unidimensional Model. **2b.** Oblique 3-Factor Model. **2c.** Hierarchical Model. The ϕ (phi) parameters are the *factor correlations* (correlations between constructs), the λ (lambda) parameters are called *factor loadings* (relations between a measure/item/indicator and a construct/common factor), and the ϵ (epsilon) parameters are called *item uniquenesses* (variance in an item/indicator that is unique to that item and not shared with the latent construct). Each measurement model thus asserts that variance in a measure can be decomposed into: (a) a portion of variance accounted for by a common factor/latent construct, and (b) a portion of variance unique to the measure. If two measures have shared variance in a measurement model, it should be because they measure the same construct (e.g., Figure 2a), or because they measure related constructs (e.g., Figure 2b). Specifying the measurement model involves specifying which items/indicators measure which constructs (i.e., the pattern of factor loadings, λ), as well as specifying whether constructs are allowed to correlate with each other (e.g., the pattern of factor correlations, ϕ, from the oblique factor model [Figure 2b], or the existence of a higher-order construct [Figure 2c]).

items/indicators. In this construct validity paradigm, the survey items are written with the goal of quantifying each individual's standing on the latent concept of job satisfaction. Another example is the unidimensional measurement model for the construct of board of directors control (Boyd, 1994; Boyd, Gove, & Hitt, 2005), which is measured by five indicators: % of stock owned by board, number of directors representing ownership groups, proportion of insiders on the board, director pay and CEO duality (negative indicators).

Next, Figure 2b shows a measurement model with more than one construct (i.e., oblique 3-factor model). The three constructs (also called *factors* in factor analysis) are labeled A, B, and C; and each of these constructs/factors is reflected with its own unique set of measures. The model is called *oblique*, which means the constructs (or factors) are correlated with each other (there is no theoretical constraint requiring the constructs to be uncorrelated, or *orthogonal*). Note the correlations between constructs (ϕ, called *factor correlations*), and the relationships between each indicator and its corresponding construct/factor (λ, called *factor loadings*). Examples of Figure 2b might include the measurement model for job satisfaction, organizational commitment, and job involvement (which are 3 correlated constructs, but are measured via different items/indicators; Mathieu & Farr, 1991).

Another measurement model is shown in Figure 2c (hierarchical model): (a) there are 3 constructs/factors (A, B, and C), (b) each construct/factor has its own unique items/indicators, but (c) the pattern of relationships among the 3 constructs/factors is modeled as a more general or abstract higher-order factor. One example of a higher-order construct/factor is general mental ability (e.g., Spearman's *g*), which is reflected by lower-order factors such as numerical ability, verbal ability, and spatial ability (Ackerman, Beier, & Boyle, 2005), each of which has its own indicators/items/operations.

Generate Items or Indicators

After the construct is defined and its conceptual properties and distinctiveness from related constructs is clarified, it is time to begin choosing items/indicators. One principle of measurement models is that the content of the items/indicators should correspond to the content of the construct. The process of selecting or creating particular items or measures from a universe of possible indicators, in order to represent a particular hypothetical construct domain, is called *domain sampling* (Nunnally, 1970).

Global versus facet approaches to domain sampling. In order to assess most constructs, one can simply sample indicators (e.g., write survey items or choose archival indicators) from the given construct domain. But in order to assess broad constructs, there are two available strategies for domain sampling: (a) global domain sampling - directly sampling the broad construct domain, or (b) facet domain sampling - sampling the lower-order specific construct domains and then combining responses across narrow facet domains to assess the broader construct. If seeking to measure conscientiousness, a global item would be, “I am conscientious” or “I am careful”. A facet composite approach to domain sampling would be to use such items as, “I am detail-oriented”, “I am industrious”, and “I am responsible,” and then mathematically combine these items (e.g., by averaging). *Global domain sampling* asks the *respondent* to average across narrow domains prior to answering the item, whereas *facet domain sampling* asks the *researcher* to average across narrow domains after the item responses are collected.

The global approach requires that each of the indicators, or items, fully reflects the content of the construct at the level of abstraction that is used to define the construct. Continuing the job satisfaction example, global survey items would reference the idea of overall satisfaction with the job in its entirety (Ironson, Smith, Brannick, Gibson, & Paul, 1989). In contrast, the facet-composite approach uses facet domain sampling by identifying important facets of a broad, higher-order construct. Items/indicators are chosen to reflect each facet-level construct, theoretically reasoning that the facet constructs themselves are specific reflections of the higher-order construct. Regardless whether global or facet domain sampling is used, Clark and Watson (1995; 2019) recommend oversampling the content domain to include both items/indicators directly assessing the target construct and items/indicators tangentially related to the target construct, to enable distinctions to be drawn in later analyses.

When choosing a certain number of items/indicators to measure a construct, researchers can face a bandwidth-fidelity dilemma (Cronbach, 1960; Ones, Viswesvaran, & Reiss, 1996). For example, if we are constrained to use only 3 items on a survey measure, then we must make the choice between *measuring a narrow facet construct (e.g., industriousness) reliably* (by writing 3 very-similar items), versus *measuring a broad construct (e.g., conscientiousness) unreliably* (by using one item to measure each facet; industriousness, attention to detail, responsibility). If a researcher claims to be measuring a broad Big Five personality trait with only 2 or 3 items, then we know that researcher has chosen to either: (a) only measure one facet of the trait, reliably, or (b) measure the broad trait, unreliably. In order to measure a broad trait reliably, one could simply use more than 2 or 3 items.

Adherents of the global domain sampling approach argue that the hierarchical facet domain sampling approach is problematic. First, if you measure a broad construct (job satisfaction) by oversampling one part of the content domain (satisfaction with working conditions) and under-sampling another part of the content domain (satisfaction with compensation), then your unbalanced domain sampling can produce over- or underrepresentation of particular lower-order constructs in the resulting measurement instrument. Second, it is difficult to theoretically or empirically determine that the facet sampling approach has captured the "right" facets (Scarpello & Campbell, 1983). In contrast, those favoring facet domain sampling argue that the global approach is not applicable for some constructs, because it is difficult to sample some constructs at the proper level of abstraction (e.g., effective items/indicators exist for the facet constructs of verbal intelligence and spatial intelligence, but it is difficult to find good indicators for the higher order construct of general intelligence).

We, the authors, differ in our views on the utility and appropriateness of the global and facet approaches—specifically, Lambert believes that domain sampling should occur at the same level of abstraction or construct breadth where the analysis occurs, whereas Newman believes that facet domain sampling can support analyses at both narrow and broad levels of abstraction. We recognize the value in both perspectives, particularly as regards the untested assumption that broad constructs can be assessed by combining items/indicators designed to measure narrower constructs (is the whole meaningfully different from the sum of its parts?; cf. Chang, Ferris, Johnson, Rosen, & Tan, 2012; Scarpello & Campbell, 1983). We urge you to understand the distinctions between these different approaches and to choose based on your own reason and logic.

Item/indicator generation. Advice for generating effective survey items is familiar to many but bears repetition (Bradburn, Sudman, & Wansink, 2004; Krosnick & Presser, 2010; Tourangeau, Rips, & Rasinski, 2000), because it is still frequently ignored. Items/indicators should be clearly written with unambiguous interpretations. For example, the question "How often did you seek medical care last year?" is ambiguous because medical care is not defined (e.g., are dentists and urgent care clinics included?), and last year may mean the prior calendar year or the prior twelve months. Items/indicators should be unidimensional referring to one idea only. "Do you like cake and ice cream?" is problematic because a respondent may like one and not the other. Items/indicators should correspond to the definition and should not be contaminated with content from related constructs (e.g., "How satisfied are you with your pay and your supervisor" confounds pay satisfaction with supervisor satisfaction). Simple language is always preferred, even when items/indicators will be administered to highly-educated respondents. Jargon and slang should be avoided because some respondents may not understand the terms, and the language may become dated.

Items/indicators may be sampled from the content domain both *deductively* (items/indicators drawn directly from the construct definition, prior theory, and/or prior measurement; (e.g., Colquitt, 2001), or *inductively* (perhaps by asking members of the target population or experts to generate examples of items/indicators; e.g., Bennett & Robinson, 2000). The deductive approach rests on theory and acknowledges the expertise of past scholars, and the inductive approach may incorporate the perspectives of knowledgeable stakeholders and members of the population being studied, facilitating the development of new theory (Hinkin, 1995). Both the deductive and inductive approaches to indicator generation are useful, and may be employed singly or in combination.

When developing items/indicators to measure a construct, how many are desirable? According to Hinkin's (1998) rule-of-thumb, the initial data collection (prior to validation) should contain approximately 8 to 12 items/indicators per construct, and the final data collection should contain 4 to 6 items/indicators per construct. We suggest the absolute minimum number of items/indicators is three, because that number is required to algebraically identify a measurement model when testing construct validity via structural equation modeling (SEM). In order to ensure adequate internal consistency reliability (e.g., Cronbach's $\alpha > .7$), the number of items required would depend upon average inter-item correlations (i.e., $\alpha = \frac{j\bar{r}_{item,item}}{1+(j-1)\bar{r}_{item,item}}$). As interitem correlations decrease (due to item measurement error, broader constructs being measured, or dichotomous scoring), more items/indicators are needed. For instance, measures with interitem correlations $\bar{r}_{item,item} = .3\text{-}.4$ (e.g., typical measures of broad job attitudes and work behavior) often need 4 to 7 items/indicators per construct, measures with $\bar{r}_{item,item} = .2\text{-}.3$ (e.g., typical measures of broad Big Five personality traits) often need about 6 to 12 items/indicators per trait, and measures with $\bar{r}_{item,item} = .1\text{-}.2$ (e.g., intelligence constructs) often need 10 to 27 items/indicators per construct, to maintain adequate internal consistency reliability and construct coverage. Beyond ensuring adequate Cronbach's α, considerations of scale length and scale shortening practices should preserve scale unidimensionality (McDonald's ω), representativeness of the construct by the set of items/indicators (Messick, 1995), and adequate part-whole correlations for the shortened scale (see Cortina et al., 2020).

Response scales (e.g., with response options referring to the extent of agreement, to frequency, or amount) should be carefully selected to accurately map onto respondents' ability to discriminate between response options (e.g., neither too few nor too many points in the scale), and in terms they understand (e.g., Americans generally are more familiar with Fahrenheit temperature than with Celsius, and in the range of 0°F to 100°F rather than to 300°F; Krosnick & Presser, 2010). Moreover, the meaning of the verbal anchors (i.e., response format) for response scales should align with the wording of the items/indicators, and should capture the full range of respondents' intended answers (Tourangeau, Conrad, & Couper, 2013, p. 78). Results are somewhat mixed and depend on the question and the sample, but 5-point, 7-point, and 9-point (odd-numbered) scales are common and may be preferable; and reliability can be higher when the points are accompanied by verbal anchors rather than just labeling the endpoints (Alwin & Krosnick, 1991).

Assess Content Validity

We address content validity in this step, rather than step 3, because of its central role in choosing effective operationalizations of constructs. All measures, regardless of their type, should exhibit content validity (Aguinis & Vandenberg, 2014; Anderson & Gerbing, 1991; Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993). Remembering that the relationships between items/indicators and constructs represent a measurement theory necessitates presenting a theoretical rationale to justify the operationalization of the construct. The point of content validity analysis is to demonstrate that the indicators correspond to the construct definition.

Multiple approaches can be used to bolster the case for content validity of items/indicators used to measure a construct (see J. C. Anderson & Gerbing, 1991; Colquitt, Sabey, Rodell, & Hill, 2019; Hinkin & Tracey, 1999; Schriesheim et al., 1993). Briefly, the approaches described by these authors involve asking a sample of judges to either (a) classify each item/indicator as matching one construct definition more than it matches other construct definitions (Anderson & Gerbing, 1991; Colquitt et al. label this *definitional distinctiveness*), and/or (b) rate the degree of correspondence between each item and a set of various construct definitions (Hinkin & Tracey, 1999; Colquitt et al. label this *definitional correspondence*). Ideally, each item/indicator can be correctly classified as belonging to its intended construct definition, and/or can be rated to have a high degree of correspondence with its intended construct definition. For example, subject matter experts (e.g., faculty, doctoral students, advanced undergraduates, or members of the population being studied) can rate the extent to which draft items/indicators are consistent with definitions of a new construct (Wolfson, Tannenbaum, Mathieu, & Maynard, 2018). Colquitt et al. (2019) systematically tested and developed norming standards for evaluating the probability of correct item categorization, and for evaluating the magnitude of item-definition correspondence (norming standards for both of these depend upon correlations between the focal construct and related constructs). Another content validity approach, sometimes called cognitive interviewing or a "think aloud" technique, involves prompting respondents to report every thought that occurs to them as they respond to survey questions or other measures (Willis, 2005). For example, Grégoire et al. (2010) asked experienced entrepreneurs to report their thoughts as they completed an opportunity recognition exercise, supporting the logical argument for content validity. As a result of content validity analysis, items/indicators that do not correspond to their intended constructs may be deleted, prior to collecting data for confirmatory factor analysis.

It is difficult to overstate the importance of content validity assessment. Miller et al. found that 66% of a sample of published papers lacked correspondence between the construct definition and the operationalization of organizational performance (e.g., organizational performance is defined as a broad latent construct, but often operationalized as a single dimension of performance; Miller, Washburn, & Glick, 2013). Likewise, individual employees' job performance has also been operationalized in ways that diverge from construct conceptualizations (J. P. Campbell, Gasser, & Oswald, 1996). As both sets of authors point out, the lack of correspondence between measures and constructs renders the interpretation of results from a single study meaningless, and obstructs the cumulation of results across studies.

In our experience, measures that routinely suffer from weak confirmatory factor analysis (CFA) results have likely not been subjected to content validity assessment (either by the Anderson-Gerbing 1991 approach, or the Hinkin-Tracey 1999 approach); and such assessment greatly improves the chances that one's measurement model will exhibit good fit after data are collected. Further, content validity assessment can be used for scale revision with existing scales, to improve CFA results (Carpenter, Son, Harris, Alexander, & Horner, 2016).

Documenting the Process and Evidence

When selecting operationalizations to match a construct definition, the choices must be documented. The requirements for documenting an *existing measure* are relatively straightforward. Researchers should report: (a) representative example items/indicators (or the full scale if the scale is new), (b) notes about instructions to participants, (c) scoring guidelines (including standardization decisions), (d) the response scale [e.g., (1-not at all) to (5-a great deal)], (e) the measurement model to be tested (e.g., which items/indicators load onto which constructs, and which constructs are specified to be correlated; see Figure 2), and (f) relevant citations for the measure and any existing theory that supports the chosen operationalization.

The instructions to respondents, response scale, instructions for scoring and coding, description of the training for behavioral raters, required procedures (e.g., for web-scraping), and other materials integral to interpreting data accurately are all part of the measure, and must be reported (DeVellis, 2003; Tourangeau et al., 2013). Instructions, to both respondents and to researchers, can be theoretically important by creating context. For example, the same behavioral items/indicators may be used to measure followers' perceptions of their leaders' behavior today (Tepper et al., 2018), or on average (Judge & Piccolo, 2004); and the instructions are necessary to clarify which information was requested.

The theoretical rationale supporting *changes to an existing measure* should also be described (Heggestad et al., 2019). For example, when items/indicators originally referring to supervisors as the target are changed to refer to the organization as the target, this should be mentioned. However, when troublesome items/indicators are revised (e.g., wording changes), or a lengthy scale is trimmed, the content validity process should be repeated with the modified scale (including empirical evidence from an independent sample for why an item/indicator was dropped, the resulting part-whole correlation for the shortened scale, and new validity evidence to support any changes to item wording; Heggestad et al., 2019). This supports the correspondence between the construct definition and its modified operationalization/items/indicators.

Documenting the process for proposing a *new measure* requires a lengthier and more detailed description. This includes the theoretical rationale for the items/indicators, specifying the measurement model, and steps taken to document content validity.

Step 3: Evidence to Confirm Construct Validity

Once a construct has been defined (Step 1: Define) (see Pedhazur & Schmelkin, 1991; Podsakoff et al., 2016), and after measures/items/indicators have been selected and screened for their subjective content-based connection to the construct definition (Step 2: Operationalize) (Anderson & Gerbing, 1991; Colquitt et al., 2019; Hinkin & Tracey, 1999), the final step (Step 3: Confirm) is to begin the iterative process of confirming and refining the measurement model in a series of independent samples. Construct validity is not determined by pointing to a specific statistic, but is a plausible conclusion that is based on an array of evidence consistent with the proposed theoretical measurement model (Jackson, Gillaspy, & Purc-Stephenson, 2009; McDonald & Ho, 2002). Because construct validity is not a property of a scale, but rather a property of the specific application of the scale in a particular sample, evidence for construct validity must be examined anew for each sample (Messick, 1995; Nunnally & Bernstein, 1994). The specific type of evidence that may be persuasive varies depending on the definition of the construct, but often includes assessing the reliability of the measures, testing the measurement model with CFA, and assessing the nomological validity of the focal construct(s).

In our discussion of construct validity, we report common rules of thumb to orient readers to desirable standards; but rules of thumb are only coarse approximations of truth, can be easily misapplied, and might be useless in specific circumstances (Lance & Vandenberg, 2009). Rules of thumb should not be applied thoughtlessly, and are not hard and fast. As with any rule of thumb, it is the researcher's underlying logic and strength of argument that should be paramount, not the numerical rule per se.

Collect Data to Test Measurement Model

When collecting data to evaluate one's measurement model, the researcher should sample data from the population of interest (e.g., working adults, top management team members, customer service representatives, firms in a dynamic environment). Convenience samples (e.g., MBA students, MTurk workers/online panels/crowdsourced data, undergraduate students) might well represent one's population of interest. However, we advise researchers to use a diverse selection of convenience samples—that is, to avoid using three MTurk samples or three student samples only—and to attempt to validate constructs across different samples from the population.

Planning an adequate sample size is a complicated issue in CFA (Gerbing & Anderson, 1985; Jackson, 2003; MacCallum, Browne, & Sugawara, 1996; MacCallum, Lee, & Browne, 2010; Muthén & Muthén, 2002). In summary, it is important to secure an adequate sample size in order to maintain adequate statistical power for CFA hypothesis tests, as well as to limit convergence failures and error in parameter estimates (factor loadings, factor intercorrelations) and model fit indices (χ^2, CFI, TLI, RMSEA). Sample size is not the only important factor, as the quality of CFA outcomes is also enhanced by larger magnitudes of factor loadings, having multiple indicators per variable, and avoiding model misspecification. A general rule of thumb for CFA sample size might not make sense, but without one researchers may push the boundaries by using tiny samples. The closest we can find to an empirically-grounded rule of thumb is Jackson's (2001) result, showing sample sizes of $N =$ 200-400 produced better CFA results than samples of $N = 100$, but that there are diminishing returns after $N = 400$ (as summarized by Jackson, 2007). Sample sizes smaller than 200 might be adequate when factor loadings are large (e.g., greater than .7) or when there are many items/indicators per factor. The rule of thumb we advocate ($N \approx 200$ or greater) is the same as Hoelter's (1983) tentative suggestion that sample size should exceed $N = 200$ per group, to "indicate that a particular model adequately reproduces an observed covariance structure" (p. 331).

Further, the population from which one samples might influence item/indicator variance, item/indicator means/base rates, normality, or whether the item/indicator makes sense. Prior to conducting CFA, it is important to inspect and report item/indicator distributions to ensure the items/indicators exhibit adequate item variance and are not extremely skewed [e.g., Bennett & Robinson, 2000, removed items with standard deviations < 1.2 (on a 1-to-7 scale) from their workplace deviance measure]. In the long run, the measurement model can ultimately be tested across different populations using item-level meta-analysis (Carpenter et al., 2016).

Test the Measurement Model with Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is an elegant methodology for confirming the construct validity of one's measures. Many sources describe how to conduct CFA (e.g., Brown, 2015; Lance & Vandenberg, 2002; Pedhazur & Schmelkin, 1991). CFA is the first step in the two-step approach to structural equation modeling advocated by Anderson and Gerbing (1988), where the first step is to test the measurement model and the second step is to test the substantive structural model. That is, the *a priori* measurement model is a hypothesis about the relationships between indicators and constructs, and should be tested before proceeding to subsequent tests of substantive hypotheses as specified in the theoretical model of interest. Because the goodness-of-fit of a substantive structural equation model is often driven by the fit of its measurement model component (O'Boyle & Williams, 2011), it is essential to assess the measurement model independent of the overall theoretical model. CFA is not limited to survey data alone, but can be used for any multiple-indicator micro or macro construct, regardless of the source of the data (archival, self-report, etc.).

It is worth mentioning at this point that *exploratory* factor analysis (EFA) is unnecessary and rarely appropriate, until after a CFA has already been attempted and failed. When the researcher assigns items/indicators to a construct, this constitutes a hypothesized relationship that should be tested. Thus, CFA should be used first. The attempt to use an EFA first would be tantamount to a confession that the researcher did not know what s/he was trying to measure when the data were collected (had no hypothesized measurement model). To restate, EFA only makes sense after CFA has failed, or if data being used were collected without any construct definitions guiding the selection of items/indicators (e.g., archival data). In such instances, we still recommend undertaking Step 2 (Operationalize the Construct, content validity analysis) as described above in order to specify one's measurement model, prior to implementing any factor analysis.

Is EFA ever appropriate? Yes. Exploratory, inductive approaches can complement deductive approaches (Aguinis & Vandenberg, 2014). Thus, EFA may be useful, under two circumstances: (a) when the researcher legitimately does not know what they are trying to measure (e.g., when the Big Five personality traits were originally derived; see Cattell, 1947; Fiske, 1949; Norman, 1963), and/or (b) after CFA has failed (i.e., EFA, including EFA with parallel analysis to detect inadvertent multidimensionality, or exploratory uses of CFA with post hoc model modifications, can be appropriate, but only if used for the purpose of specifying a model that is then immediately tested using CFA in an independent dataset). When assessing dimensionality with limited information about what may be a complex measurement structure, both of the above conditions might be met. Whether using traditional EFA or more recent exploratory procedures in SEM (Asparouhov & Muthén, 2009; Brown, 2015; Conway & Huffcutt, 2003; Fabrigar et al., 1999; Morin, Arens & Marsh, 2016; Zickar, 2020), we reiterate the requirement to test the obtained measurement model with CFA using an independent sample. Further, we emphasize that principal components analysis (PCA) is not factor analysis, and should be avoided whenever the goal is to measure latent constructs (see discussion by Conway & Huffcutt, 2003; Ford et al., 1986). Also, when using EFA, orthogonal rotations should be avoided because most constructs are theoretically correlated, and if constructs are indeed uncorrelated the oblique techniques will still reveal that.

Reliability

A common reliability index is Cronbach's coefficient α, which assesses internal consistency across items/indicators (Cortina, 1993; cf. Cho & Kim, 2015). Cronbach's α is weak evidence of construct validity because it: (a) is strongly influenced by the number of items/indicators in the measure, (b) is a lower bound estimate of reliability, (c) does not address convergent or discriminant validity between constructs, and (d) assumes that the combined items/indicators are unidimensional, tau-equivalent (i.e., have equal factor loadings), and have uncorrelated item errors (Cho & Kim, 2015; Cortina, 1993; McNeish, 2018). Alternative internal consistency reliability indices include coefficient omega (ω_h, which relaxes the assumption of tau equivalence, and also indexes unidimensionality) and composite reliabilities (which are appropriate for a construct of multiple related dimensions for either tau equivalent or congeneric measurement models; Cho, 2016). Cronbach's α assesses reliability across items/indicators on a measure, but reliability can also be estimated across raters (LeBreton & Senter, 2008), across occasions (Schmidt, Le, & Ilies, 2003), or across items/indicators, raters, and occasions simultaneously (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; DeShon, 2002; Woehr, Putka, & Bowler, 2012); generalizability theory). For example, an approach for using trained raters to code CEO narcissism on the basis of video clips was also evaluated by comparing ratings of video clips by doctoral students to both self-reports and to others' reports of narcissism (Petrenko, Aime, Ridge, & Hill, 2016). Content analysis of text can also be assessed for reliability. For instance, multiple kinds of measurement error for constructs developed from computer-aided text analyses can be estimated (McKenny, Aguinis, Short, & Anglin, 2018).

Six desirable features when confirming a measurement model. When confirming a measurement model, one should pursue the following desirable features: (a) convergent validity, (b) discriminant validity, (c) simple structure, (d) no correlated uniquenesses (residuals), (e) any measurement model modifications tested on new data, and (f) nomological validity.

Convergent validity. Convergent validity is supported when two items/indicators of the same construct are related. To elaborate, Campbell and Fiske (1959) describe convergent validity using the correlation between two maximally dissimilar methods used to measure the same construct, whereas Bagozzi (1981) posits that Campbell and Fiske's convergent validity is a special case of the more general notion of *convergence in measurement*—which applies to convergence between any two indicators of the same construct (e.g., different items on the same survey), regardless whether they are taken from dissimilar methods. This broader form of convergent validity can be exhibited by large factor loadings (λ; see Figure 2), suggesting that multiple items/indicators are largely measuring the same thing. Indeed, Widaman (1985, p. 9) pointed out that factor loadings can be a mathematical transformation of Campbell and Fiske's (1959) validity diagonal itself, and therefore "loadings in Λ_T are structural modeling analogues of, and are measures of, convergent validation of measures." We note a common rule of thumb that standardized factor loadings should be $\lambda \geq .4$, consistent with a similar heuristic in the context of exploratory factor analysis (Ford, MacCallum, & Tait, 1986). This cutoff is arbitrary, but is also a fairly low standard. Factor loadings $\lambda \geq .4$ mean that the latent factor accounts for at least $\lambda^2 \geq .4^2 = 16\%$ of variance in the measure (it also implies that interitem correlations are at least $r = .16$). All else equal, larger loadings are generally better (e.g., Fornell & Larcker, 1981, propose a rule of thumb similar to standardized $\lambda \geq .7$), however if factor loadings are too high (standardized $\lambda > .9$), it means items/indicators are empirically redundant and may not be providing enough unique information to justify the additional length of the survey instrument.

Discriminant validity. Discriminant validity is supported when the items/indicators of two different constructs are not too strongly correlated (Campbell & Fiske, 1959). Discriminant validity should be assessed for both items/indicators and for intact measures (sets of indicators). A coarse rule of thumb we endorse is that factor intercorrelations among theoretically distinct constructs should typically be $\phi < .7$ (see Figure 2b). Rönkkö and Cho (2020) avoid dichotomous judgments and offer a graduated approach to evaluating the extent of discriminant validity. We point out that testing a measurement model in a piecemeal fashion (e.g., one construct at a time) offers no evidence for discriminant validity and should thus be avoided.

Additionally, researchers often seek to support discriminant validity inferences by comparing their hypothesized measurement model against theoretically plausible alternative models (e.g., comparing an oblique factor model with two constructs against a unidimensional model in which the two constructs are constrained to be perfectly correlated: $\phi = 1.0$, or a single-factor model). If the model-data fit for the unidimensional model is worse than model-data fit for the oblique multifactor model (e.g., if $\Delta \text{CFI} > .01$), this is treated as initial evidence for discriminant validity. However, this sort of model comparison evidence can be quite weak, because with large sample sizes even a factor correlation of $\phi = .8$ or $\phi = .9$ can be empirically distinguished from $\phi = 1.0$. There are circumstances where two constructs may be very highly correlated: such as a very strong causal (nearly deterministic) effect, the existence of a higher-order construct, or a slightly different form of the same construct (same content but different targets; e.g., perceived organizational support and perceived supervisor support).

Simple structure. When specifying and testing a measurement model, typically each item/indicator should load onto only one construct/factor (each item/indicator should be assigned to directly assess a particular construct). Cross-loading or double-loading items/indicators are guilty of "obfuscating the meaning of the estimated underlying constructs" (p. 417, Anderson & Gerbing, 1988). Nonetheless, one possible use of double-loading items/indicators would be in multitrait-multimethod (MTMM) CFA models (Widaman, 1985), in which a single item/indicator loads onto both a trait factor and a method factor. In general, items/indicators that assess multiple constructs should be removed or replaced with items/indicators intended to cleanly assess one construct.

No correlated uniquenesses/residuals. An assumption of CFA in general is that the indicator/item residuals (item error/uniqueness/variance not shared in common with the latent factor) are uncorrelated with the residuals of other indicators. This assumption is stringent, and might be a reason that many measurement models are rejected. The general rule is that indicator/item residuals should not be allowed to correlate and that doing so is cheating (Cole, Ciesla, & Steiger, 2007). As such, one motive for specifying correlated uniquenesses in a *post hoc* fashion is to improve model fit.

However, this practice leads to inaccurate reporting of fit indices (Cortina, Green, Keeler, & Vandenberg, 2016), and is a form of capitalizing on chance that renders fit indices meaningless.

Exceptions could be made when there is a strong *a priori* theoretical reason for expecting correlated residuals, for example longitudinal designs where an item/indicator is repeated and its uniqueness correlates with itself over time, or perhaps because the content of the items/indicators is influenced by a construct other than the one in question (Cole et al., 2007). Much of the time when researchers want to allow correlated uniquenesses among items/indicators, it is because they believe there exists a lower-order or specific factor that two or more items/indicators have in common (based on similar item wording, etc.). In such cases, the researcher should specify the lower-order factor and confirm its existence in a new dataset.

Nomological Validity

The understanding of a construct is facilitated by knowing its relationships with other constructs. Testing the nomological network (as predicted in Step 2 above) entails showing that a construct is related to other variables as expected (Cronbach & Meehl, 1955; Schwab, 2005). As an example, when Klein, Cooper, Molloy and Swanson (2014) proposed their new measure of organizational commitment, they predicted and found relationships between their new measure and theoretically-related constructs (i.e., positive relationships with job satisfaction, organizational identification, extra-role behavior, and in-role effort; and a negative relationship with turnover intentions). Likewise, Danneels (2016), as a part of developing new measures for dynamic capabilities, tested the nomological net showing that R&D competency predicted concurrent and subsequent accumulation of technological resources. Further, in the rare circumstance that it might be necessary to use a single-item/indicator measure of a construct, assessing the nomological network can bolster the empirical case for its construct validity.

Modifying the Measurement Model: Data-Driven Modifications Require Collecting New Data

When the results of the CFA suggest that the measurement model fits the data and there is evidence of convergent validity, discriminant validity, and nomological validity; then it is reasonable to proceed to hypothesis testing. However, when the measurement model exhibits poor fit to the data, then steps should be taken to identify the problem(s) and to modify the model. Such steps typically include deleting an item/indicator, or specifying an item/indicator to load onto a different factor than was originally hypothesized. However, we strongly caution that, *when estimating measurement model fit, one should not use modification indices or other CFA information to change the model in any way after looking at the data*. Changing the model then reporting the modified fit on the same data renders the model fit indices meaningless in the current dataset. Only after CFA fails (e.g., poor model fit, standardized factor loadings < .4, standardized factor correlations > .7), then CFA or EFA may be used in an exploratory fashion (i.e., by inspecting *both* the model modification indices/standardized residuals *and* the results of a content validity assessment—see Step 2 above) to identify the source of misfit. We emphasize that the modified measurement model must be tested on a new, independent sample. *Post hoc* model modifications to improve model fit often fail to replicate in future samples, because they capitalize on chance characteristics of the dataset at hand (MacCallum, Roznowski, & Necowitz, 1992). A new dataset must be collected for each revised measurement model.

When CFA fails, researchers might be tempted to improve model fit by using item *parceling*. An item parcel is a subset of items/indicators aggregated (usually averaged together) to form an indicator (e.g., instead of using 15 items as indicators, one might use 5 parcels of 3 indicators each). The advantages of parceling compared to using single items/indicators (Bandalos, 2002; Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013; Williams, Vandenberg, & Edwards, 2009) include: (a) parcels are more reliable (smaller uniquenesses), more normally distributed, have smaller correlations between residuals, and have more intervals between scale points, (b) parceling greatly reduces the number of parameters being estimated, enabling CFA to converge with much smaller sample sizes, and (c) parceling generates better model goodness-of-fit. Regardless of the parceling strategy employed (e.g., based on item factor loadings, random assignment of items/indicators to parcels, or *a priori* theoretical facets), the major disadvantage of parceling is that a CFA with parceled indicators hides the very information that is necessary to evaluate the relationships between a construct and its item-level indicators (Little, Rhemtulla, Gibson, & Schoemann, 2013; Marsh et al., 2013; Meade & Kroustalis, 2006). Parcels may or may not reflect the construct, but they are not diagnostic of the construct-item relationship for each item/indicator. As such, parceling may be appropriate when testing structural models, but should be avoided when testing measurement models. Parceling strategies, when used to assess measurement models, signal that more item/indicator-level construct validation work is needed.

Documenting the Process and Evidence

Reporting standards for CFA. Researchers should report the model(s) tested and analyses performed with enough detail to satisfy the questions and concerns of a skeptical academic audience. We encourage transparency regarding the results and any flaws discovered in the evidence, recognizing that less-than-perfect validity evidence need not doom a study from making a useful contribution. When CFA is conducted on a dataset to test one's hypothesized measurement model, the results should include (Jackson et al., 2009): (a) the software version and estimation routine (typically maximum likelihood), (b) the missing data treatment (usually FIML, which currently is the default approach in lavaan, LISREL, and Mplus), (c) sample description, (d) sample size, (e) list of items/indicators and response formats, and (f) whether items/indicators were screened for normality, low variance or high skewness. The results should also include a table containing means, standard deviations, N's, Cronbach's α's, and correlations among all measures. There should be a description of model(s) estimated, which specifies: factor loadings, factor correlations, and no correlated uniquenesses. The model fit indices should be reported (i.e., χ^2, *df*, RMSEA, CFI, TLI, and SRMR), how the latent variables were scaled (by setting either a loading or a factor variance to 1), as well as standardized factor loadings for each indicator (or a mean and range of factor loadings for each factor) and the standardized latent factor correlations. Model fit indices are often interpreted according to rules of thumb (good fit = RMSEA < .06; CFI & TLI > .95; SRMR < .08; Hu & Bentler, 1999; cf. Fan & Sivo, 2007; Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004); although rules of thumb are often inappropriately generalized beyond the models that generated these guidelines. Model fit indices by themselves are not evidence of construct validity; model fit, parameter estimates, and nomological validity (in the form of correlations between latent variables) should be considered together.

Construct validity without CFA? For instances in which CFA is not possible [e.g., with necessarily small samples (where the level of analysis is countries/nations, or Fortune 100 companies), single-item archival measures (often used in strategic management research), or natural language processing (Pandey & Pandey, 2019)], it still remains important to attempt to establish construct validity. In such instances, the researcher should still present evidence for the measures used, in terms of the desirable features described earlier: (a) convergent validity – positive correlation with other measures of similar constructs, (b) discriminant validity – correlation less than .7 with other constructs, (c) simple structure – each measure is intended to assess a single construct, (d) any measurement model modifications tested on new data, and (f) nomological validity – correlations in expected directions with external variables.

Other Considerations: Method Variance, Measurement Equivalence, Formative Constructs, Single Item Measures, Forced Choice Measures, Multilevel Constructs, and Algebraically Combined Measures

In addition to the basic methods of construct validation reviewed in the preceding sections, there are a few other considerations to keep in mind when addressing construct validity.

Method variance and multitrait-multimethod analyses (MTMM). Adapted from Campbell and Fiske's (1959) classic paper on construct validation via convergent and discriminant validity, the premise of multitrait-multimethod analyses is that the same construct can be measured via different methods [e.g., different raters (self-rating vs. supervisor-rating vs. coworker rating of employee job performance), or different instruments (MOAQ vs. JIG as different measures of job satisfaction)]. Analyses of a multitrait-multimethod matrix provide an elegant way to assess how much of the variance in a measure is due to trait variance versus method variance, as well as to estimate the extent to which observed covariance between measures is due to true trait covariance versus common method variance. Nonetheless, MTMM CFA analyses sometimes suffer mathematical problems that often prevent the model from being estimated (Brannick & Spector, 1990); although promising alternative estimation methods have been proposed (Helm, Castro-Schilo, & Oravecz, 2017). Other approaches to assess common method variance are reviewed by Podsakoff et al. (2003); Richardson, Simmering, and Sturman (2009); and Williams and McGonagle (2016). This is still an active area of research, with major implications for determining both the substantive trait constructs (e.g., personality, attitudes) and the method constructs (e.g., self-rating bias) that underlie our measures.

Measurement equivalence. When constructs are measured in multiple groups, or in multiple instances across time, it is essential to establish that the meaning of the construct is invariant over time or equivalent in each of the groups studied (Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Vandenberg & Lance, 2000). Measurement equivalence/invariance analyses involve establishing that the factor loadings, item intercepts, and item uniquenesses are essentially the same across groups, or across time, to support the claim that the same construct is being measured across different groups or different occasions. Measurement equivalence analyses are a prerequisite for mean comparisons (e.g., comparing average vocational interests between women and men, cross-cultural comparisons using different language translations of a survey, test bias), and are needed for understanding whether groups differ on the actual underlying construct of interest. Measurement equivalence analyses are also a prerequisite to longitudinal analysis, to ensure that observed changes in measures are due to actual changes in the underlying construct rather than simply due to changes in instrumentation or to shifts in how the scales or raters are calibrated over time (Chan, 1998a; Schmitt, 1982).

Reflective versus formative indicators. The operationalization of a construct, as expressed in a measurement model, represents a theory because it states the causal relationship between the construct and its indicators. One critical choice is between a construct that causes the scores on the measures (i.e., reflective indicators) or vice versa, meaning that the scores on the measures combine to cause the construct (i.e., formative indicators). There is a recurring, and sometimes raging, debate among methodologists regarding formative variables (Bollen & Diamantopoulos, 2017; Edwards, 2001, 2011; Howell, Breivik, & Wilcox, 2007; MacKenzie et al., 2005). We give more attention to reflective variables in the current paper, not only because they are in greater use in the organizational sciences, but because we find the methodological arguments in their favor to be superior (Edwards, 2011). Information on formative measurement is presented primarily to educate readers on problems with formative item/indicator models, and to equip them to assess the challenges posed by formative measures.

In a formative measure, the items/indicators may be uncorrelated with each other, and may correspond to unrelated facets of the construct (meaning that the items/indicators need not correspond to the entire construct definition). If researchers choose to develop a formative measure, it is not sufficient to generate a set of items/indicators where each measure refers to different content, and then to simply declare the measurement model as formative. Instead, it is essential to justify the theoretical logic for why and how items/indicators cause the latent construct, why it is unnecessary to account for measurement error in the item/indicator(s), to specify the measurement model, to demonstrate statistical evidence required for validity, and to specify how the model will be identified for estimation in SEM. To identify a formative model, it is necessary to include at least two reflective measures, or two endogenous outcome constructs (MacKenzie et al., 2005); but the loadings on the formative indicators will vary depending on which reflective constructs or measures are chosen, changing the meaning of the formative construct itself (Edwards, 2011). Accordingly, researchers who choose formative measurement should develop theoretically sound qualifying criteria for specifying which reflective variables or indicators, and not others, are appropriate for identifying the model when assessing construct validity (Bollen & Diamantopoulos, 2017; Edwards, 2001, 2011; Howell et al., 2007; MacKenzie et al., 2005). One of the better examples of validating a formative measurement approach is the research on entrepreneurial orientation by Anderson et al. (2015).

Single item/indicator measures. Irrespective of the content of a construct, it is important for each construct to have several measures associated with it, to avoid construct contamination and construct deficiency that can plague single-indicator construct measurement (i.e., mono-operation bias; Shadish, Cook, & Campbell, 2002). For example, if the construct 'job performance' is measured with a single operationalization (e.g., quarterly sales in $), then this measure suffers from construct contamination/measures irrelevant constructs beyond individual job performance (e.g., the wealth of the neighborhood where sales are made and the national economy are both reflected in dollar sales, but are not part of the job performance construct); and also suffers construct deficiency/underrepresents the construct of interest by not measuring specific attributes that are central to the definition of the construct (e.g., not measuring specific behaviors that constitute job performance; J. P. Campbell et al., 1996). In short, construct contamination and deficiency can be better controlled when there is more than one operationalization (e.g., more than one item or one indicator) per construct.

Despite the clear advantages of using multiple indicators of a construct, single item/indicator measures are sometimes unavoidable, for instance when relying on archival measures. Using single item/indicator measures confers a special obligation to clearly articulate the theoretical rationale linking item/indicator and construct, and to assess content and nomological validity. Using single item/indicator measures is better than no measures at all, but not as desirable as using multiple items/indicators to indicate a construct.

Forced-choice measures. The procedures we describe in this paper are not appropriate for measures that instruct respondents to choose one of two or more options, or to rank a set of items/indicators. These kinds of forced-choice measures are designed to reduce biases (e.g. social desirability, halo effects, faking) and to uncover respondents' true preferences. Forced-choice measures are ipsative, in that the scores (e.g. 0 for response A, 1 for response B) for each respondent: (a) sum to the same value, creating dependence in the data, (b) disregard the absolute scores (considering only relative scores), and (c) violate assumptions of analyses based on correlations/covariances (Cornwell & Dunlap, 1994; Hicks, 1970; Meade, 2004; Schriesheim, Hinkin, & Podsakoff, 1991). More recent research has avoided the problems of ipsativity in data from forced-choice measures by applying item response theory (IRT) to the scoring (Brown & Maydeu-Olivares, 2013; Stark, Chernyshenko, & Drasgow, 2005; Zhang et al., 2020).

Level of analysis in constructs. Constructs can reside at multiple levels of analysis (e.g., industries, firms, departments, teams, individual employees, or days nested within individuals). For example, individual job performance may be nested within team performance, which is nested within organizational or firm performance. Performance may be averaged within industries, or assessed daily across workdays/occasions. Such constructs are useful for theorizing multilevel phenomena (e.g., cross-level moderation effects where a group-level construct moderates the relationship between two individual-level constructs; Jex & Bliese, 1999). Validating constructs at different levels of analysis requires additional attention. Kozlowski and Klein have done brilliant work showing how researchers must specify the *what* (phenomenon of interest), *how* (top down vs. emergence processes), *where* (levels or units involved), *when* (role of time), and *why* (causal reasoning) of multilevel constructs (Klein & Kozlowski, 2000a). Just as findings from one level of analysis cannot be assumed to generalize to another level (Ostroff, 1993), a theoretical construct conceptualized and measured at one level of analysis may become a different construct at a lower or higher level of analysis (Chan, 1998b; Klein & Kozlowski, 2000b).

Establishing construct validity for aggregate or group-level constructs requires matching the nature of the construct to the appropriate type of empirical evidence. Many widely-studied organizational phenomena inherently require multilevel measurement and analysis, because the phenomena are often measured via individual-level perceptions but the constructs are conceptualized to reside at the group-level of analysis (e.g., organizational climate, leadership; (James & Jones, 1974; Klein & Kozlowski, 2000b; Rousseau, 1985). These constructs require considering within-group agreement and reliability (Bliese, 2000; LeBreton & Senter, 2008; Newman & Sin, 2020); consideration of the item referent (e.g., "I am satisfied" vs. "My team is satisfied"; Chan, 1998b; Klein & Kozlowski, 2000b), as well as potential considerations of measurement equivalence across levels of analysis (psychometric isomorphism; Tay, Woo, & Vermunt, 2014), and nomological network homology across levels of analysis (Chen, Bliese, & Mathieu, 2005). In many cases (e.g., when measuring team-level constructs or organization-level constructs that are assessed via individual-level perceptions or surveys), multilevel CFA is often appropriate (Muthen, 1994)—which estimates two sets of factor loadings and factor correlations (at both the within-group and between-group levels of analysis) simultaneously.

Algebraically combined measures. Finally, there is a class of formative measures in widespread use that are constructed by combining one or more sources of data using mathematical operations. Difference scores and ratios are two common examples of algebraically combined measures; the problems associated with both have been well documented and alternative estimation strategies proffered (Bergh & Fairbank, 2002; Certo, Busenbark, Kalm, & LePine, 2020; Edwards, 2002; Finkel, 1995; Kronmal, 1993; Wiseman, 2009). As another example, social network measures may combine ratings from team members to capture centralization or density of ties within a group. These algebraically combined scores ignore the separate effects of the individual pieces of information that are combined into a single score intended to stand in for the meaning of the construct. The difficulty is that algebraically combined scores embody assumptions that are rarely described or tested. For example, using ratios implies that only relative scores rather than absolute scores of the components matter. Algebraically combined measures pose interpretational difficulties, and the embodied assumptions should be exposed and evaluated for their plausibility.

Summary and Conclusion

It is important to remember—whether using established, revised, or newly developed measures—that the relationships between items/indicators and the constructs they are intended to represent is a measurement theory that must be tested. We have endeavored to provide practical guidance to reviewers, editors, and authors in the form of a checklist with supporting explanations. Our advice is no guarantee that the measure of a construct is valid but should be viewed as a guide to improving measurement practice. The reader should keep in mind that our paper is simply an overview of current recommendations and is subject to future revision. Moreover, it was necessary to give short shrift to many construct measurement and validity topics, and our review of recommended practices cannot address all contingencies that will arise in research practice. Instead, our overarching recommendation is to develop a sound, theoretically-derived approach to measuring constructs and to provide substantial and persuasive evidence that the measurement model should not be rejected.

We remind readers that confirming construct validity does not certify that a measure is validated for all time and for all purposes. The extent to which tests require local validity analyses varies. Tests that have been developed for specific and well defined purposes, and have been extensively validated (e.g., well-known tests for educational, vocational or clinical purposes) may not need validity assessment for a specific application. Yet, many, if not most, tests used for research purposes and academic publications lack recommended and extensive validity assessment (American Educational Research Association, 2014). Moreover, the purpose of the research is related to the required precision of the instruments; scales used for making decisions about peoples' lives (e.g., hiring, admissions) often require more and different validity evidence than empirical contributions to theoretical work. Construct validity is an ongoing process (American Educational Research Association, 2014). Even if using measures that exhibited adequate construct validity in prior studies, local construct validity evidence may be critical. Measures for a construct never reach standards of validity such that further testing is unnecessary - construct validity must be revisited each time the construct is used.

Appendix A: Glossary of Construct Validity Terms

Construct (latent construct, concept, factor) – an attribute, process or disposition of people, groups, or firms (Cronbach & Meehl, 1955, p. 283; Messick, 1981, p577).

Measure (operationalization, item, indicator) – "an observed score gathered through self-report, interview, observation, or some other means (DeVellis, 2003; Edwards, 2003, p. 329; Edwards & Bagozzi, 2000; Lord & Novick, 1968; Messick, 1995)."

Construct validity – "the correspondence between a construct and a measure" as evaluated by cumulative evidence (Cronbach & Meehl, 1955; Edwards, 2003, p. 329; Nunnally, 1978; Schwab, 1980).

Content validity – "the degree to which a measure represents a particular domain of content" (Anderson & Gerbing, 1991; Edwards, 2003, p. 330).

Construct domain – theoretical definition of the content area of a particular construct (Hinkin, 1995, p. 969; Nunnally, 1970; Podsakoff, MacKenzie & Podsakoff, 2016; Schwab, 1980; Schriesheim et al., 1999). The notion of a construct domain is useful for understanding the practice of domain sampling.

Domain sampling – choosing particular items or measures from a universe of possible items, in order to represent a particular hypothetical construct domain (Nunnally, 1970, p. 546).

Measurement model – specifies the relationships of indicators/items to their assigned constructs, typically with freely correlated constructs (Anderson & Gerbing, 1988).

Note. Adapted from Newman, Harrison, Carpenter, & Rariden (2016).

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

Lisa Schurer Lambert https://orcid.org/0000-0001-5167-8442
Daniel A. Newman https://orcid.org/0000-0002-5876-4498

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*(1), 30-60. Retrieved from https://doi.org/10.1037/0033-2909.131.1.30

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*, 569-595.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attitudes. *Sociological Methods and Research*, *20*(1), 139-181. https://doi.org10.1177_0049124191020001005

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (2014). *Standards for educational and psychological testing*. AERA.

Anderson, B. S., Kreiser, P. M., Kuratko, D. F., Hornsby, J. S., & Eshima, Y. (2015). Reconceptualizing entrepreneurial orientation. *Strategic Management Journal*, *36*(10), 1579-1596. Retrieved from https://doi.org/10.1002/smj.2298

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411-423. https://doi.org/10.1037/0033-2909.103.3.411

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. Retrieved from https://doi.org/10.1037/0021-9010.76.5.732

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*(3), 397-438. https://doi.org10.1080/10705510903008204

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: A comment. *Journal of Marketing Research (pre-1986)*, *18*(000003), 375. Retrieved from https://doi.org/10.1177/002224378101800312

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*(1), 78-102. Retrieved from https://doi.org/10.1207/S15328007SEM0901_5

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*(3), 349-360. Retrieved from https://doi.org/10.1037/0021-9010.85.3.349

Bergh, D. D., & Fairbank, J. F. (2002). Measuring and testing change in strategic management research. *Strategic Management Journal*, *23*(4), 359-365. https://doi.org/10.1002/smj.232

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). Jossey-Bass.

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal–formative indicators: A minority report. *Psychological Methods*, *22*(3), 581-596. Retrieved from https://doi.org/10.1037/met0000056

Boyd, B. K. (1994). Board control and ceo compensation. *Strategic Management Journal*, *15*(5), 335-344. Retrieved from https://doi.org/10.1002/smj.4250150502

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Construct measurement in strategic management research: Illusion or reality? *Strategic Management Journal*, *26*(3), 239-257. Retrieved from https://doi.org/10.1002/smj.444

Bozeman, D. P., & Perrewé, P. L. (2001). The effect of item content overlap on organizational commitment questionnaire–turnover cognitions relationships. *Journal of Applied Psychology*, *86*(1), 161-173. Retrieved from https://doi.org/10.1037/0021-9010.86.1.161

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design – for market research, political polls, and social and health questionnaires* (First edition ed.). Jossey-Bass.

Brady, D. L., Brown, D. J., & Liang, L. H. (2017). Moving beyond assumptions of deviance: The reconceptualization and measurement of workplace gossip. *Journal of Applied Psychology*, *102*(1), 1-25. Retrieved from https://doi.org/10.1037/apl0000164

Brannick, M. T., & Spector, P. E. (1990). Estimation problems in the block-diagonal model of the multitrait-multimethod matrix. *Applied Psychological Measurement*, *14*(4), 325-339. Retrieved from https://doi.org/10.1177/014662169001400401

Brayfield, A. H., & Rothe, H. F. (1951). An index of job satisfaction. *Journal of Applied Psychology*, *35*(5), 307-311. Retrieved from https://doi.org/10.1037/h0055617

Bromiley, P., Rau, D., & Zhang, Y. (2017). Is R&D risky? *Strategic Management Journal*, *38*(4), 876-891. Retrieved from https://doi.org/10.1002/smj.2520

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36-52. Retrieved from https://doi.org/10.1037/a0030641

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford 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

Campbell, J. P., Gasser, M. B., & Oswald, F. (1996). The substantive nature of job performance variability. In K. Murphy (ed.), Individual differences and behavior in organizations (1st ed, pp. 258–299).

Cardinal, L. B., Sitkin, S. B., & Long, C. P. (2010). A configurational theory of control. *Organizational Control*, *51*, 79, 85-100.

Carpenter, N. C., Son, J., Harris, T. B., Alexander, A. L., & Horner, M. T. (2016). Don't forget the items: Item-level meta-analytic and substantive validity techniques for reexamining scale validation. *Organizational Research Methods*, *19*(4), 616-650. https://doi.org/10.1177/1094428116639132

Cattell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika*, *12*, 197-220. https://doi.org/10.1007/BF02289253

Certo, S. T., Busenbark, J. R., Kalm, M., & LePine, J. A. (2020). Divided we fall: How ratios undermine research in strategic management. *Organizational Research Methods*, *23*(2), 211-237. https://doi.org/10.1177/1094428118773455

Chan, D. (1998a). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indivator latent growth modeling (MLGM). *Organizational Research Methods*, *2*(4), 421-483 Retreived from 10.1177/109442819814004.

Chan, D. (1998b). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234-246. Retrieved from doi.org/10.1037/0021-9010.83.2.234

Chang, C.-H., Ferris, D. L., Johnson, R. E., Rosen, C. C., & Tan, J. A. (2012). Core self-evaluations:A review and evaluation of the literature. *Journal of Management*, *38*(1), 81-128. https://doi.org/10.1177/0149206311419661

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods*, *8*(4), 375-409. Retrevied from https://doi.org//10.1177/1094428105280056

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651-682. https://doi.org/10.1177/1094428116656239

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207-230. https://doi.org/10.1177/1094428114555994

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412-1427. https://doi.org/10.1037/pas0000626

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, *12*(4), 381-398. https://doi.org/10.1037/1082-989x.12.4.381

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386-400. Retrieved from doi.org/10.1037/0021-9010.86.3.386

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. 10.1037/apl0000406.supp (Supplemental)

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147-168. https://doi.org/10.1177/1094428103251541

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational & Organizational Psychology*, *67*(2), 89-100. https://doi.org/10.1111/j.2044-8325.1994.tb00553.x

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2016). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods*, *20*(3), 350-378. https://doi.org/10.1177/1094428116676345

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. (2020). From alpha and omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the journal of applied psychology. *Journal of Applied Psychology*, *105*(12), 1351-1381. https://doi.org/10.1037/apl0000815

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511. https://doi.org/10.1037/pspp0000102

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed). Harper.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

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

Danneels, E. (2016). Survey measures of first- and second-order competences. *Strategic Management Journal*, *37*(10), 2174-2188. https://doi.org/10.1002/smj.2428

DeShon, R. P. (2002). Generalizability theory. In *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 189-220). Jossey-Bass.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2 ed., Vol. 26). Sage Publications.

DeVellis, R. F. (2017). *Scale development : theory and applications* (Fourth edition. ed.). Thousand Oaks, CA: Sage.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662-680. https://doi.org/10.1037/0021-9010.70.4.662

Edwards, J. R. (2001). Multidimensional constructs in organizational behavior research. *Organizational Research Methods*, *4*(2), 144-192. https://doi.org/10.1177/109442810142004

Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow, & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 350-400). Jossey-Bass.

Edwards, J. R. (2003). Construct validation in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: the state of the science* (2nd edition ed.). Erlbaum.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14* (2), 370-388. https://doi.org/10.1177/1094428110378369

Edwards, J. R., & Bagozzi, R. (2000). Relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*(4), 668-689. https://doi.org/10.1177/1094428110380467

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*(3), 509-529. https://doi.org/10.1080/00273170701382864

Finkel, S. E. (1995). *Causal analysis with panel data*. Sage Publications.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*(3), 329-344. https://doi.org/10.1037/h0057198

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291-314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, *18*(3), 382-388. https://doi.org/10.2307/3150980

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research*, *20*(3), 255-271. https://doi.org/10.1207/s15327906mbr2003_2

Grégoire, D. A., Shepherd, D. A., & Schurer Lambert, L. (2010). Measuring opportunity-recognition beliefs: Illustrating and validating an experimental approach. *Organizational Research Methods*, *13*(1), 114-145. https://doi.org/10.1177/1094428109334369

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management*, *45*(6), 2596-2627. https://doi.org/10.1177/0149206319850280

Helm, J. L., Castro-Schilo, L., & Oravecz, Z. (2017). Bayesian Versus maximum likelihood estimation of multitrait–multimethod confirmatory factor models. *Structural Equation Modeling*, *24*(1), 17-30. https://doi.org/10.1080/10705511.2016.1236261

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*(3), 167-184. https://doi.org/10.1037/h0029780

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967-988. https://doi.org/10.1177/014920639502100509

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*(1), 104-121. https://doi.org/10.1177/109442819800100106

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

Hoelter, J. W. (1983). The analysis of covariance structures:Goodness-of-fit indices. *Sociological Methods & Research*, *11*(3), 325-344. https://doi.org/10.1177/0049124183011003003

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117-144. https://doi.org/10.1080/03610739208253916

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205-218. https://doi.org/10.1037/1082-989X.12.2.205

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Ironson, G. H., Smith, P. C., Brannick, M. T., Gibson, W. M., & Paul, K. B. (1989). Construction of a job in general scale: A comparison of global, composite, and specific measures. *Journal of Applied Psychology*, *74*(2), 193-200. https://doi.org/10.1037/0021-9010.74.2.193

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A monte carlo investigation. *Structural Equation Modeling*, *8*(2), 205–223. doi:10.1207/S15328007SEM0802_3

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128-141. https://doi.org/10.1207/S15328007SEM1001_6

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(1), 48-76. https://doi.org/10.1080/10705510709336736

Jackson, D. L., Gillaspy, J. A.Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6-23. https://doi.org/10.1037/a0014694

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin*, *81*(12), 1096-1112. https://doi.org/10.1037/h0037511

Jex, S. M., & Bliese, P. D. (1999). Efficacy beliefs as a moderator of the impact of work-related stressors: A multilevel study. *Journal of Applied Psychology*, *84*(3), 349-361. https://doi.org/10.1037/0021-9010.84.3.349

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*(5), 755-768. https://doi.org/10.1037/0021-9010.89.5.755

Kellermanns, F. W., Walter, J., Lechner, C., & Floyd, S. W. (2005). The lack of consensus about strategic consensus: Advancing theory and research. *Journal of Management*, *31*(5), 719-737. https://doi.org/10.1177/0149206305279114

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book.

Ketchen, D. J., Ireland, R. D., & Baker, L. T. (2013). The use of archival proxies in strategic management studies: Castles made of sand? *Organizational Research Methods*, *16*(1), 32-42. https://doi.org/10.1177/1094428112459911

Klein, H. J., Cooper, J. T., Molloy, J. C., & Swanson, J. A. (2014). The assessment of commitment: Advantages of a unidimensional, target-free approach. *Journal of Applied Psychology*, *99*(2), 222-238. https://doi.org/10.1037/a0034751

Klein, H. J., Molloy, J. C., & Brinsfield, C. T. (2012). Reconceptualizing workplace commitment to redress a stretched construct: Revisiting assumptions and removing confounds. *Academy of Management Review*, *37*(1), 130-151. https://doi.org/10.5465/arma.2010.0018

Klein, K. J., & Kozlowski, S. W. J. (2000a). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein, & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations; foundations, extensions, and new directions* (pp. 3-90). Jossey-Bass.

Klein, K. J., & Kozlowski, S. W. J. (Eds.). (2000b). *Multilevel theory, research, and methods in organizations; foundations, extensions, and new directions*. Jossey-Bass.

Kronmal, R. A. (1993). Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society*, *156*(3), 379-392. https://doi.org/10.2307/2983064

Krosnick, J. A., & Presser, S. (2010). Questions and questionanaire design. In J. D. Wright, & P. V. Marsden (Eds.), *Handbook of survey research*. Emerald Group.

Lance, C. E., & Vandenberg, R. J. (2002). Confirmatory factor analysis. In F. Drasgow, & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 221-254). Jossey-Bass.

Lance, C. E., & Vandenberg, R. J. (2009). *Statistical and methodological myths and urban legends : Doctrine, verity and fable in the organizational and social sciences*. Routledge.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist*, *41*(11), 1183-1192. https://doi.org/10.1037/0003-066X.41.11.1183

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*(2), 112-125. https://doi.org/10.1016/j.obhdp.2010.02.003

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815-852. https://doi.org/10.1177/1094428106296642

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285-300. https://doi.org/10.1037/a0033266

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635-694. https://doi.org/10.2466/pr0.1957.3.3.635

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores. Reading*, MA: Addison-Wesley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

MacCallum, R. C., Lee, T., & Browne, M. W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling*, *17*(1), 23-41. https://doi.org/10.1080/10705510903438906

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504. https://doi.org/10.1037/0033-2909.111.3.490

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, *90*(4), 710-730. https://doi.org/10.1037/0021-9010.90.4.710

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing technologies. *MIS Quarterly*, *35*(2), 293-A295. Retrieved from http://ezproxy.gsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=60461934&site=ehost-live&scope=site.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320-342. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*(3), 257-284. https://doi.org/10.1037/a0032773. 10.1037/a0032773.supp (Supplemental)

Mathieu, J. E., & Farr, J. L. (1991). Further evidence for the discriminant validity of measures of organizational commitment, job involvement, and job satisfaction. *Journal of Applied Psychology*, *76*(1), 127-133. https://doi.org/10.1037/0021-9010.76.1.127

Mayer, J., Roberts, R. D., & Barsade, S. G. (2008). Emerging research in emotional intelligence. *Annual Review of Psychology*, *59*, 507-536. https://doi.org/10.1146/annurev.psych.59.103006.093646

McDonald, R. P., & Ho, R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82. https://doi.org/10.1037/1082-989x.7.1.64

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of computer-aided text analysis. *Journal of Management*, *44*(7), 2909-2933. https://doi.org/10.1177/0149206316657594

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412-433. https://doi.org/10.1037/met0000144

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531-552. https://doi.org/10.1348/0963179042596504

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*(3), 369-403. https://doi.org/10.1177/1094428105283384

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, *89*(3), 575-588. https://doi.org/10.1037/0033-2909.89.3.575

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Miller, C. C., Washburn, N. T., & Glick, W. H. (2013). PERSPECTIVE—The myth of firm performance. *Organization Science*, *24*(3), 948-964. https://doi.org/10.1287/orsc.1120.0762

Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, *14*(2), 224-247. https://doi.org/10.1016/0001-8791(79)90072-1

Muthen, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376-398. https://doi.org/10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

Newman, D. A., Harrison, D. A., Carpenter, N. C., & Rariden, S. M. (2016). Construct mixology: Forming new management constructs by combining old ones. *Academy of Management Annals*, *10*(1), 943-995. https://doi.org/10.1080/19416520.2016.1161965

Newman, D. A., Joseph, D. L., & Hulin, C. L. (2010). Job attitudes and employee engagement: Considering the attitude "A-factor". In *Handbook of employee engagement: Perspectives, issues, research and practice* (pp. 43-61). Edward Elgar Publishing.

Newman, D. A., & Sin, H.-P. (2020). Within-group agreement (rWG): Two theoretical parameters and their estimators. *Organizational Research Methods*, *23*(1), 30-64. https://doi.org/10.1177/1094428118809504

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574-583. https://doi.org/10.1037/h0040291

Nunnally, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill Inc.

O'Boyle, E. H.Jr., & Williams, L. J. (2011). Decomposing model fit: Measurement vs. Theory in organizational research using latent variables. *Journal of Applied Psychology*, *96*(1), 1-12. https://doi.org/10.1037/a0020539

O'Neill, O. A., & Rothbard, N. (2017). Is love all you need? The effects of emotional culture, suppression, and work–family conflict on firefighter risk-taking and health. *Academy of Management Journal*, *60*(1), 78-108. https://doi.org/10.5465/amj.2014.0952

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660-679. https://doi.org/10.1037/0021-9010.81.6.660

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, *78*(4), 569-582. https://doi.org/10.1037/0021-9010.78.4.569

Pandey, S., & Pandey, S. K. (2019). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods*, *22*(3), 765-797. https://doi.org/10.1177/1094428117745648

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Lawrence Erlbaum Associates, Publishers.

Petrenko, O. V., Aime, F., Ridge, J., & Hill, A. (2016). Corporate social responsibility or CEO narcissism? CSR motivations and organizational performance. *Strategic Management Journal*, *37*(2), 262-279. https://doi.org/10.1002/smj.2348

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, *19*(2), 159-203. https://doi.org/10.1177/1094428115624965

Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods*, *12*(4), 762. https://doi.org/10.1177/1094428109332834

Rönkkö, M., & Cho, E. (2020). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *0*(0), 1094428120968614. https://doi.org/10.1177/1094428120968614

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*, 1-37.

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, *36*(3), 577-600. https://doi.org/10.1111/j.1744-6570.1983.tb02236.x

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206-224. https://doi.org/10.1037/1082-989X.8.2.206

Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, *17*(3), 343-358. https://doi.org/10.1207/s15327906mbr1703_3

Schriesheim, C. A., Castro, S. L., & Cogliser, C. C. (1999). Leader–member exchange (LMX) research: A comprehensive review of theory, measurement, and data-analytic practices. *The Leadership Quarterly*, *10*(1), 63–113. https://doi.org/10.1016/S1048-9843(99)80009-5

Schriesheim, C. A., Hinkin, T. R., & Podsakoff, P. M. (1991). Can ipsative and single-item measures produce erroneous results in field studies of French and Raven's (1959) five bases of power? An empirical investigation. *Journal of Applied Psychology*, *76*(1), 106-114. https://doi.org/10.1037/0021-9010.76.1.106

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*(2), 385-417. https://doi.org/10.1177/014920639301900208

Schwab, D. P. (1980). Construct validity in organizational behavior. *Research in Organizational Behavior*, *2*, 03-43.

Schwab, D. P. (2005). *Research methods for organizational studies* (second ed.). Lawrence Erlbaum Associates.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods*, *19*(1), 80-110. https://doi.org/10.1177/1094428115598239

Shipp, A. J., Edwards, J. R., & Lambert, L. S. (2009). Conceptualization and measurement of temporal focus: The subjective experience of the past, present, and future. *Organizational Behavior and Human Decision Processes*, *110*(1), 1-22. https://doi.org/10.1016/j.obhdp.2009.05.001

Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science*, *1*(3), 375-388. https://doi.org/10.1177/2515245918775707

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct measures. *Organizational Research Methods*, *17*(1), 77-106. https://doi.org/10.1177/1094428113517008

Tepper, B. J., Dimotakis, N., Lambert, L. S., Koopman, J., Matta, F. K., Park, H. M., & Goo, W. (2018). Examining follower responses to transformational leadership from a dynamic, person–environment fit perspective. *Academy of Management Journal*, *61*(4), 1343-1368. https://doi.org/10.5465/amj.2014.0163

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford University Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic—transformational leadership research: Back to the drawing board? *Academy of Management Annals*, *7*(1). https://doi.org/10.1080/19416520.2013.759433

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), 4-69.

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

Williams, L. J., & McGonagle, A. (2016). Four research designs and a comprehensive analysis strategy for investigating common method variance with self-report measures using latent variables. *Journal of Business & Psychology*, *31*(3), 339-359. https://doi.org/10.1007/s10869-015-9422-9

Williams, L. J., Vandenberg, R. J., & Edwards, J. R. (2009). Structural equation modeling in management research: A guide for improved analysis. *The Academy of Management Annals*, *3*, 543-604. Retrieved from http://www.informaworld.com/10.1080/19416520903065683.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Wiseman, R. (2009). On the use and misuse of ratios in strategic management research. In *Research methods in strategy and management* (Vol. Vol. 5, pp. 75-110). Emerald Group Publishing Limited.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multi-trait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134-161. https://doi.org/10.1177/1094428111408616

Wolfson, M. A., Tannenbaum, S. I., Mathieu, J. E., & Maynard, M. T. (2018). A cross-level investigation of informal field-based learning and performance improvements. *Journal of Applied Psychology*, *103*(1), 14-36. https://doi.org/10.1037/apl0000267

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569-590. https://doi.org/10.1177/1094428119836486

Zickar, M. J. (2020). Measurement development and evaluation. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1), 213-232. https://doi.org/10.1146/annurev-orgpsych-012119-044957