Routledge
Taylor & Francis Group

Annals of the International Communication Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rica20

Evolving considerations and empirical approaches to construct validity in communication science

Nicholas David Bowman & Alan K. Goodboy

To cite this article: Nicholas David Bowman & Alan K. Goodboy (2020) Evolving considerations and empirical approaches to construct validity in communication science, Annals of the International Communication Association, 44:3, 219-234, DOI: 10.1080/23808985.2020.1792791

To link to this article: https://doi.org/10.1080/23808985.2020.1792791

ANNALS OF THE INTERNATIONAL COMMUNICATION ASSOCIATION
2020, VOL. 44, NO. 3, 219–234
https://doi.org/10.1080/23808985.2020.1792791

Routledge
Taylor & Francis Group

REVIEW ARTICLE

Evolving considerations and empirical approaches to construct validity in communication science

Nicholas David Bowman [a] and Alan K. Goodboy[b]

[a]College of Media and Communication, Texas Tech University, Lubbock, TX, USA; [b]Department of Communication, West Virginia University, USA

ABSTRACT

For social sciences such as communication studies, in which key variables are often indirectly measured using myriad operationalizations, issues of construct validity are critical to the veracity of claims made from empirical data. This manuscript considers the ways we can improve how we demonstrate construct validity when using survey items. Some of these improvements are associated with our discussions of the sources of construct validity, and others are associated with the statistical analyses (and mistakes in those analyses) that we engage to demonstrate construct validity. The end goal of this manuscript is to review and reveal best practices for establishing construct validity as an incremental step in improving the quality of communication science.

KEYWORDS

Unified view of validity; exploratory factor analysis; confirmatory factor analysis; construct validity

A common refrain in the quantitative social sciences is 'garbage in, garbage out' – a reference to the fact that the claims we can make about the world around us are only as accurate as the data on which those claims are based. At least one source of 'garbage' data in communication sciences stems from a lack of sufficient evidence for construct validity. In the current manuscript, we provide an overview of traditional and contemporary perspectives on construct validity, before shifting to a discussion of common methods (and mistakes) of assessing construct validity in factor analysis techniques. Our end goal is to encourage a careful and principled consideration of construct validity – conceptually and empirically – to protect against the determinantal influence of poor measurement on the veracity of communication science.

Validity, defined

When scholars discuss measurement validity, they are often referring to what Shadish et al. (2002) labeled *construct validity*, or 'inferences about the constructs that research operations represent' (p. 20). To better understand construct validity, it is important to step back and understand what is meant by the term construct. Broadly, constructs are understood as the concepts or variables a researcher is interested in studying. The literature review – or 'front end' – of most manuscripts contains some explication of the conceptual underpinnings of a researcher's constructs of interest.

Constructs are often latent and not directly observed, and thus can be operationalized in myriad ways. For example, one can assess video game players' perceptions of their enjoyment of a recent gaming session by (a) interviewing them about the parts of the game they enjoyed, (b) asking them to list the emotions that they recall feeling during gameplay, (c) completing a short survey of Likert-type questions about their arousal, engagement, and positive affect towards the game,

CONTACT Nicholas David Bowman nick.bowman@ttu.edu College of Media and Communication, Texas Tech University, 3003 15th St, Lubbock, TX 79409, USA

(d) analyzing video of their physical reactions during gameplay, or (e) processing skin conductance data from an apparatus attached to their feet while they were playing, among others. Each of these five measures represents various *operationalizations* (ways of measuring, or making the construct tangibly assessible) the *conceptualization* of enjoyment as an arousing, engaging, and positively emotional experience (see Figure 1). Here, the conceptualization would be informed by prior research (for example, our definition of emotion could be pulled from gaming entertainment research; Oliver et al., 2016) and likely, carefully chosen from a set of similar-but-less focal constructs, as well as unrelated and disparate constructs (also illustrated in Figure 1).

Critical to the current discussion is that each operationalization – each way of measuring the focal construct – presents a bit of a challenge to the researcher, as any given measurement represents one way of assessing the conceptual scheme of a construct but might do so at the expense of others. In Figure 1, measures on the left side might encourage study participants to speak more about their own perceptions of enjoyment as being something they readily recognize as being fun. Moving toward the center of the measurement set, we see a closed-ended item that specifically asks about the extent to which one directly enjoyed their video game (perhaps from a semantic perspective, the most 1:1 assessment of enjoyment as a construct). Shifting to the right side of the graphic, measures move more towards indirect assessments of the focal construct, such as watching players demonstrate behaviors that we might assume indicate positive engagement (and thus, enjoyment) with a video game (expressing joy, being aroused, or leaning forward towards the screen); on the far-right side, psychophysiological indicators of arousal are taken to represent a biological indicator of enjoyment (see Čertický et al., 2019). For the most part, construct validation is not purely a statistical exercise; rather, it is guided by logic and argumentation derived from substantive theory and extant research. Theoretical precepts and assumptions should guide the selection of appropriate measurement for any focal study and thus, the researcher should 'present their case' for construct validity of the chosen measure within their manuscript.

Shadish et al. (2002) discussed construct validity with respect to other types of validity of critical concern for research, including internal and external validity – both concerned broadly with replication, the former with a focus on whether observed findings are non-spurious and the latter with a focus on whether observed findings will generalize. Construct validity and internal validity are

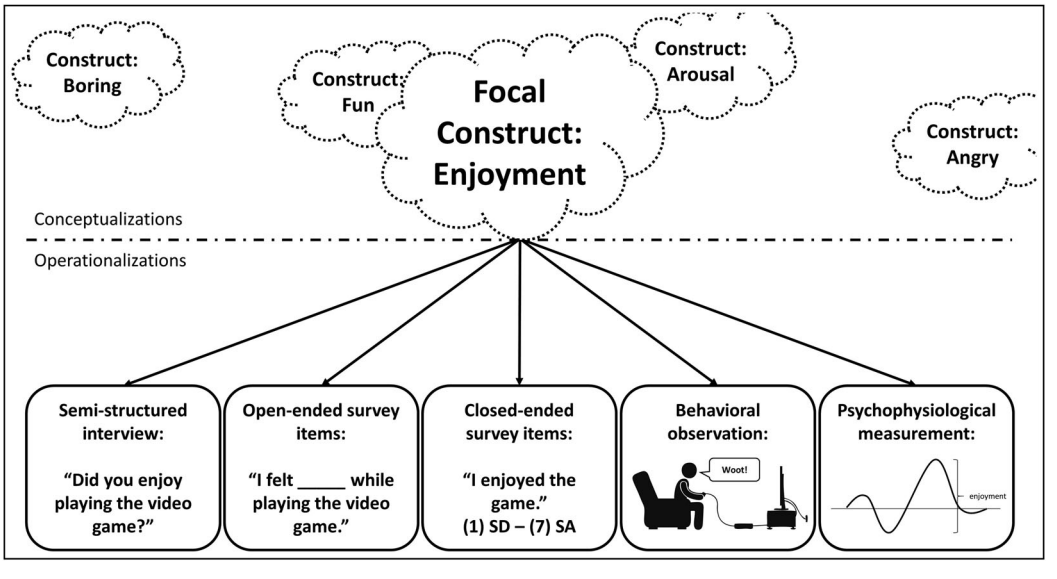

Figure 1. Conceptual relationship between a latent construct (here, enjoyment of a video game) and one of several possible operationalizations (measurements) of that construct.

similar insofar as both are concerned with the presence of confounds, or extraneous variables that either suppress or inflate observed relationships. In most cases, confounds in a study can be removed with careful attention to study design (such as considerations of laboratory procedures or survey language) or statistically controlled for, but it is possible for measurements to introduce confounds into a study. For example, by measuring demographics, researchers might force respondents to choose from pre-set categories for sex and gender, which requires minority populations to choose an 'other' category (a similar argument can be made for race and ethnicity questions, education questions, or any other demographic status). Forcing individuals from minority populations to 'other' themselves might result in a form of resentful demoralization such that they might feel as if their opinions are less valued (Bauer et al., 2017; Fraser, 2018). Shadish et al. (2002) also discussed the association between construct validity and external validity with respect to their shared concern over generalizability. When researchers select a measurement instrument for their study, it is plausible that the measurement itself requires the study to be conducted in such a way that restricts replication. For example, a researcher reliant on interviews might be heavily dependent on the skill of a single researcher that cannot be easily replicated by another researcher (indeed, variability in the researcher-participant relationship could also be seen as a confounding variable; a threat to internal validity as well); research using psychophysiological measures requires the use of measurement devices that are not readily available outside of a carefully controlled research environment and thus, is *de facto* difficult to replicate. To these ends, and a core takeaway of our discussions of the selection of a measurement model, is that *no one measurement is completely valid* and likewise *some measurements are more valid than others, depending on the research context*. Related to this, construct validation is an ongoing process, even when an area of research – or a given measurement instrument – has already compiled ample evidence of its validity.

Moving towards a unified view of validity

Traditional views on validity – such as the Shadish et al. (2002) discussion above – are commonly taught in the communication studies discipline, featured in undergraduate and graduate course curriculum and textbooks. This said, a more contemporary and refined perspective is the unified view of validity offered by the *Standards for Educational and Psychological Testing* (2014) published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). According to this perspective, validity is 'the degree to which evidence and theory support the interpretations of test scores for proposed uses' and 'involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretation' (p. 11). In this view, evidence is collected by researchers for validity purposes, and is best understood by focusing on five sources of validity, rather than distinct types of validity. These sources include: (a) internal structure, (b) associations with other variables, (c) response processes, (d) test content, and (e) consequences of use. Conceptually, these sources of validity are closely aligned with, but more comprehensive than, more traditional views of validity (Shadish et al., 2002). We note that communication studies, as a field, aligns with APA recommendations for published scholarship and thus, we should take more care to consider validity in terms of these five sources of validity.

Internal structure

This source of evidence answers the question: Do the measurement items and components match the structure they are supposed to have? In other words, does the dimensionality of a communication measure hold up as it is theoretically supposed to, whether it be (among other possibilities) unidimensional (e.g. congeneric model), essentially unidimensional (e.g. bifactor model), multidimensional (e.g. correlated factors model), or a second-order factor model? For this validity evidence, communication researchers will rely on exploratory and confirmatory factor analysis (both discussed later in the manuscript) to extract and retain a valid number of factors, interpret the meaning of factors,

and model correlations between factors. Factor loadings are especially meaningful for internal structure evidence and so is the fit of a measurement model. Item response theory (IRT) also provides important evidence for the internal structure of a measure by examining individual item properties presumed to be manifestations of the latent communication variable (see de Ayala, 2009).

Associations with other variables

This source of evidence answers the question: Does the communication measure associate with other measures that it should?

> In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and, as a result, analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence. (AERA et al., 2014, p. 16)

This is what communication researchers might typically be accustomed to in the traditional validity view; that communication measures should correlate with similar constructs (i.e. convergent validity) but relate less closely to different constructs (i.e. discriminant validity), and should relate to relevant criterion (i.e. criterion-related validity) that are distinct from the measure (i.e. predictive, concurrent). Construct validity coefficients can be obtained between theoretically predicted and observed correlations (Westen & Rosenthal, 2003). This validity evidence can be further strengthened when culled from a variety of different operationalizations (such as behavioral observations or psychophysiological measures) which can be analyzed using a multitrait-multimethod (MTMM) matrix (Campbell & Fiske, 1959).

Response processes

This source of evidence answers the question: Do the processes that affect responses align with the processes that should be occurring as participants complete a measurement instrument? Stated differently, 'theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers' (AERA et al., 2014, p. 15). For instance, if a researcher is measuring observer reports of communication competence at work, it is important to determine that respondents are thinking about communication competence ratings instead of using heuristic determinations unrelated to the measure (e.g. whether an employee is well-liked at the office). Validity evidence for response processes is gathered from individual responses typically at the item-level of analysis with targeted subgroups in a population. This evidence can include interviews from participants (e.g. asking participants why they responded to items in the way they did), think aloud procedures (e.g. asking participants to speak their thoughts as they respond to items), tracking response times (e.g. timing how much thought went into responding), and/or evidence from other methods and measures.

Test content

This source of evidence answers the question: Does the content of a measure (e.g. tasks, wording, questions, response formats) sufficiently represent the full scope of the construct? Researchers should ask themselves whether all features of a construct are represented in the measure or have features or dimensions of the construct been left out? Essentially, 'evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores' (AERA et al., 2014, p. 14). A communication researcher can obtain evidence for test content by conducting systematic observations of communication behavior (e.g. observing families communicating social support), seeking expert judgments (e.g. asking a panel of family communication experts to review a measure and make recommendations), or identifying variance that is extraneous to the construct (e.g. determining response biases based on wording).

Consequences of testing

This source of evidence answers the question: Do the consequences of using this measurement instrument correspond with the intended consequences of the measure? Put simply, communication measurements might result in consequences beyond the measurement itself. For instance, a measurement of student communication apprehension might predict which performance-based communication studies courses a student will struggle in. However, it could also create unintended (usually negative) consequences for certain groups of people. For example, perhaps measurement instruments created in the United States are not valid in other countries because they exhibit an ethnocentric bias in other cultures. The aforementioned confounds that can be introduced with closed-option gender questions (see Bauer et al., 2017; Fraser, 2018) are examples of unintended consequences – a lack of recognition of an individual's gender identity in a survey could cause psychological discomfort or reactance. When examining testing consequences, evidence is gathered that the measure creates intended effects beyond the direct scoring of that instrument and minimizes unintended and unwanted consequences from subjecting participants to the instrument.

Factor analysis techniques: statistical tests of construct validity

In quantitative approaches to communication studies, a common measurement instrument is the self-report, closed-ended survey. A common way of demonstrating the construct validity of survey instruments is by reporting the results of a factor analysis, either by using exploratory factor analysis (EFA) or confirmatory factor analysis (CFA). When conducting EFA or CFA, there are many methodological decisions a researcher must make that have appreciable consequences for the final model solution and retained factor structure. We review some of these decisions and provide methodological and statistical guidance with implications directed toward construct validity.

Exploratory factor analysis

Ideally, researchers measure communication constructs based on an *a priori* measurement model informed by theory and empirical research bolstered with validity evidence. However, researchers might not always find adequate or existing scales for their research. Extant validity evidence of an existing scale might be suspect, or perhaps a given construct has evolved such that existing scales (or scale items) no longer address the underlying concept (e.g. constructs change over decades of time). Or, a nascent construct might exist only at the theoretical level and no measurement of the construct has been attempted (perhaps a researcher is both explicating and testing a newly developed construct). In these scenarios, researchers are left to develop their own measurements and must establish the validity of their scales.

One popular technique for testing the construct validity of newly-created scales is to use EFA to examine the extent to which a proposed set of scale items are related to underlying latent factor(s) of interest – these factor(s) representing the conceptual scheme of the construct being measured. As explained by Osborne (2014), EFA is a statistical analysis that examines all pairwise relationships between a given set of scale items and uses the strength of those relationships to be explained by latent (unobserved) factors extracted from those items – put simply, EFA tries to group items together based on their shared variance. The logic for this technique is widely attributed to Spearman (1904) and depicted conceptually in Figure 2 (adapted from Matsunaga, 2010), on the left-side of the diagram.

In Figure 2, observed measures are represented by individual items drawn with right-angle edges, to suggest defined or concrete measures. The constructs 'driving' scores on those individual items (the latent factors) are represented by ellipses, drawn without edges to represent their unobserved,

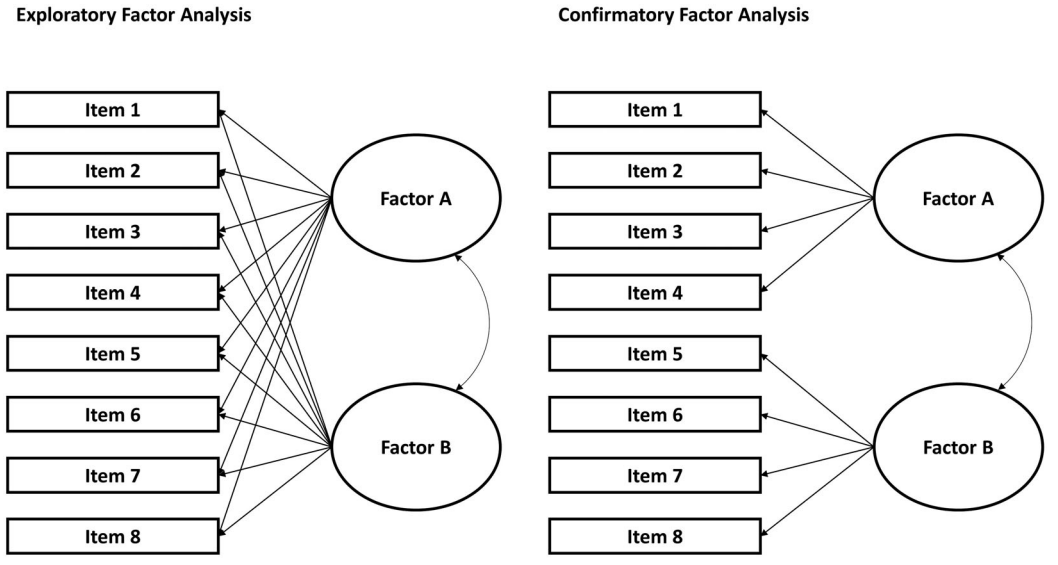

Figure 2. Conceptual differences between EFA and CFA (adapted from Matsunaga, 2010).

undefined, and more ephemeral nature. In an EFA, the researcher (for the most part) has no presumption about how many unobserved factors are included among their items and thus, the analysis assumes *by proxy* that there could be as many factors as there are observed (measured) items (Osborne, 2014). From this assumption, the analysis then helps inform the selection of those factors that explain the most amount of variance in the underling construct(s) being measured (it is rarely the case that all potential factors are retained, as this would suggest the items included in the EFA share nominal variance and thus, do not belong on scales together). Fabrigar et al. (1999) discussed numerous methodological concerns when designing studies that plan to use EFA techniques, two of which are specifically related to the number and nature of items selected for the study. First, they suggested including three to five times as many items as expected factors. Second, they suggested including items for which expected communalities are high and related to this, items with an expected high internal consistency with one another – such determinations likely rely on extended considerations of test content (one of the five sources of validity, discussed earlier). In terms of determining the sample size necessary for a robust EFA, MacCallum et al. (1999) argued such determinations depend on a number of different factors largely unknown to the researcher, such as the empirical associations between the items measured and the factor uncovered – as with most considerations of *a priori* sample size, stronger correlations require smaller samples. Others suggest a focus on sampling adequacy is more important than sample size. Two such tests are the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO, Cerny & Kaiser, 1977) that examines the proportion of variance among items that might be common (rather than unique) variance, and Bartlett's test of sphericity (cf., Tobias & Carlson, 1969), which tests the hypothesis that the correlation matrix among measured items is an identity matrix (i.e. the items are uncorrelated). Both should be considered and reported in EFA to determine if the analysis is appropriate to begin with.

Assuming the study is designed appropriately (or at least, have the necessarily number of items for which to use EFA), there are a constellation of critical decisions that the researcher must make. Here, several guides (including Costello & Osborne, 2005; Fabrigar et al., 1999; Matsunaga, 2010) distill these decisions into a serial list: determining an extraction method, determining a factor rotation method, deciding on factor loading and cross-loading values, and finally choosing which factors to retain.

Extraction method

The first analytical decision is to determine which procedure will be used to determine the fit between the final extracted factor structure and the data from which this model is being extracted. Here, two factor extraction methods are made available in most statistical packages, and the choice to use which one is primarily a function of multivariate normality: the extent to which any linear combination of those variables is normally distributed. Fabrigar et al. (1999) argued that for data with multivariate normality, maximum likelihood (ML) is preferred as it allows for an expanded set of indices for model fit, as well as allows for statistical significance testing of both (a) factor loadings (the strength of association between a given item and a given factor; discussed later) and (b) correlations between latent factors. Such extraction methods also provide two key fit indices. One is the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993), which is an estimate of the discrepancy between the factors extracted (the measurement model) and the observed data (lower values are indicative of less discrepancy and thus, better fit). The other is the expected cross-validation index (ECVI; Browne & Cudeck, 1989), which compares how well a focal model fits compared to other models (smaller values correspond with stronger evidence of cross-validation, although ECVI is only useful when comparing between factor solutions with a different number of factors) (Park et al., 2002).

When multivariate normality cannot be assumed, principal axis factoring (PAF) is preferred, although the technique does not allow for statistical significance tests for factor loadings or latent factor correlations. Here, de Winter and Dodou (2012) presented evidence suggesting principal axis factoring is more robust for extracting weak factors (factors with items that have weak communalities) and might be preferred in scenarios in which a researcher has fewer items per factor; maximum likelihood was more robust when factors loadings (within a given factor) were increasingly unequal. This said, Fabrigar et al. (1999) were careful to note that 'when the common factor model holds reasonably well in the population and severe violations of distributional assumptions are not present, solutions provided by these methods are usually very similar.' (p. 277). For EFA, either extraction method (ML or PAF) will suffice, but neither should be confused with principal components analysis (PCA). As explained by Park et al. (2002), PCA are EFA are conceptually different tests, both aimed at reducing model complexity and thus, easily confused. For purposes of the current discussion, PCA is an inappropriate and invalid method for testing the construct validity of a given measure for (at least) one core reason – it does not allow for measurement error. As a result, PCA assumes scores are perfectly reliable, which makes the analysis unsuitable for assessing the factor structure of multi-item scales reflecting a latent construct (Park et al., 2002).

Factor rotation method

Associated with the discussion of factor extraction is that of factor rotations. A factor rotation can be difficult to understand, but as explained by Park et al. (2002), 'for interpretability of factor loadings, factors are rotated in *n* dimensional space in a way to produce simple structures' (p. 566). Osborne (2014) and Brown (2009) both explained that the term rotation refers to rotating the dimensions of a plot of item scores such that the axes of each dimension (each factor assigned its own axis) are centered around the clustered items measuring that factor (See Figure 3).

In EFA, there are two broad types of rotations, the differences of which are self-defined in their names: orthogonal and oblique. For purpose of testing the construct validity of measurement models that might have multiple factors, Park et al. (2002) explained broadly that there is rarely a reason to use orthogonal rotations when creating measures in social sciences (Varimax rotation chief among them), as such rotations force an assumption of non-correlated factors, which is unlikely to be met when examining multiple factors from the same scale. Conversely, oblique rotations allow underlying factors to be correlated (which is expected in multidimensional scales) and thus, are preferred; common oblique rotations include promax and direct oblimin rotations. An emerging oblique rotation method proposed by Yates (1987) is the geomin rotation (which is the default in Mplus), so

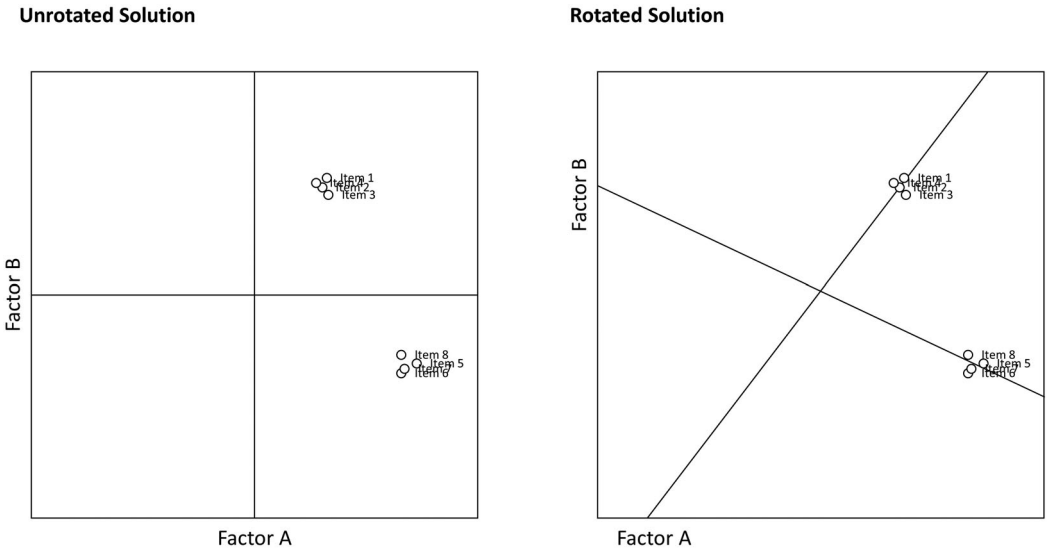

Figure 3. Conceptual distinction between an unrotated and rotated factor solutions.

named because it uses the geometric means of squared factor pattern coefficients (other oblique methods instead focus on squared reference structure elements). Asparouhov and Muthén (2009) demonstrated that geomin rotations were more accurate than other oblique rotations at estimating factor cross-loadings (discussed later) and factor correlations. In communication science, promax, direct oblimin, or geomin rotations should be used for EFA when there is more than one dimension/factor (i.e. the measure is multidimensional).

Retaining factors, factor loadings, and cross-loadings

The end goal of EFA is to uncover the latent factors from a specified set of items and as such, the 'heavy lifting' of any EFA is to interpret the results of factor extraction and rotation decisions above to determine (1) how many factors can be reasonably extracted from observed data, and (2) which items are most strongly associated with which factors; for both (1) and (2), the researcher must also decide which factors and which items to disregard.

Focusing on factor retention in EFA, the analysis will first return as many factors as there are items included in the analysis. Park et al. (2002) discussed two dominant methods of determining how many factors to retain: the *eigenvalues greater than 1* method (Kasier rule; K1 criterion) and the *scree test* (both discussed in Zwick & Velicer, 1982). Eigenvalues are proportions of the variance in items explained by a factor; as such, eigenvalues greater than 1 suggest some level of gain for using a factor rather than an individual item (see Kaiser, 1960). Overall, the logic is that an Eigenvalue greater than 1 explains more variance than a single item. For example, if an EFA returned no eigenvalue greater than 1, then the best 'factor solution' would be to retain each individual item as its own measure (i.e. the distinct items explain more variance than any grouping of items into factors). Some scholars such as Fabrigar et al. (1999) are also critical of the 'greater than 1' criteria, because it can result in scenarios by which factors that explain .01 more variance than individual items (eigenvalue = 1.01) are retained, yet factors explaining .01 less variance than individual items (eigenvalue = .99) are rejected. Since eigenvalues are affected by sampling error, these determinations based on eigenvalues that are slightly below or above one are problematic, which is why some scholars prefer parallel analysis to determine Eigenvalue cutoffs instead.

Scree tests are also based on eigenvalues, although they attempt to correct Fabrigar et al.'s (1999) critique by focusing less on the absolute value of eigenvalues and instead focus on the relative loss of

explained variance for each additional factor – given that EFAs report the strongest factors first (those that explain the most amount of shared variance), there is less variance to explain with each successive factor and thus, scree tests can demonstrate this loss (often done graphically, in a *scree plot*). However, as discussed by Park et al. (2002), 'problems can arise when there are multiple breaks or no discernable break points [in the scree plot]' (p. 565). Notably, scree tests can be bolstered when paired with parallel analysis (Horn, 1965), a process by which researchers generate an additional dataset with the same number of measured items but created using randomly generated data. After repeated EFAs on both datasets (as many as 500–1000 computations; aided with the use of Monte Carlo simulations; Watkins, 2008), the average factor eigenvalues are compared and factors are retained if their averaged eigenvalues are larger when extracted from the empirical dataset than those same eigenvalues generated from the random dataset. Henson and Roberts (2006) reported that scree tests were most accurate among numerous factor retention tests but tend to overextract factors. They found parallel analysis was the most accurate procedure, although Fabrigar et al. (1999) suggested the use of scree tests and parallel analyses in tandem. As mentioned previously in the discussion of factor extraction (factor extractions are often coupled with factor retention decisions, as the extraction methods determine which factors are retained), maximum likelihood estimation can also be used, which has the benefit of fit indices that can be analyzed more objectively.

Once a researcher has determined the number of factors to retain, they must then determine which factors are to be retained on those items – a decision informed by analyzing the pattern of factor loadings. In EFA with oblique rotation, two sets of factor loadings are provided in two sets of matrices: a pattern matrix (pattern coefficients are standardized beta-weights that represent the unique relationship between one item and its corresponding factor, controlling for all other factors) and a structure matrix (structure coefficients are correlations between an item and a specified factor); both are interpreted and reported in EFA. Factor loadings are calculated for each item and for each potential factor, and hopefully sort themselves such that items load strongly on one factor (primary loadings) and weakly on other factors (cross-loadings). In communication, McCroskey and Young (1979) suggested retaining items with primary loadings larger than .60, and removing items with cross-loadings larger than .40; such decisions are for the researcher to make (and should be made *a priori*), noting that accepting weaker primary loadings and/or stronger cross-loadings introduces error into the resultant measurement model.

Confirmatory factor analysis

Like EFA, it has become common for communication researchers to use confirmatory factor analysis (CFA) to evaluate the construct validity of communication measures. After the factor structure of a construct has been tentatively verified with EFA, CFA is used to specify and test a hypothesized measurement model where a latent variable or latent variables (i.e. factor or factors) explain the variation and covariation of measurement items. While an EFA is used to *generate* hypotheses about factors and items, a CFA is used to *test* hypotheses about the association between latent factors and items. In other words, the researcher believes the reason for item scores is explained by a latent construct or correlated but unique latent constructs. CFA must also be guided by theoretical and substantive considerations derived from the empirical literature and should not solely rely on statistical guidance (i.e. not allow mathematics to drive (re)specifications that best fit a covariance or correlation matrix).

Since accumulating validity evidence is an ongoing process, it is important to routinely examine the construct validity of communication measurements using CFA – as noted by Levine (2005), 'research results are no more valid than the measures used to collect the data' (p. 335). To bolster validity claims, researchers should routinely report CFA results when using established measures. To this end, CFA provides important information about the construct validity of a communication measure. For instance, in a multidimensional measure (i.e. two or more latent variables), large factor correlations (e.g. .80+) may indicate poor discriminant validity and that the correlated

constructs are not distinct (i.e. these are redundant factors). Or individual items may be poorly explained by a latent variable, indicated by low factor loadings (e.g. below .40) whereas other items may possess strong loadings (e.g. .70+) and are sufficiently explained by the latent construct. CFA factor loadings may reveal problematic items when an item loads on multiple factors, or that the item was chosen to load on the incorrect factor, or that an item does not load on any factor at all (see Brown, 2015). Moreover, CFA may also uncover methods effects in a measure, which commonly surface with negatively worded (reverse coded) items; these methods effects can be handled by correlating errors. A comprehensive discussion of the possibilities of CFA for measurement validation can be found in Brown (2015). In the following, we have made some observations from CFAs published in communication journals, and from these observations, we hope to improve validity claims being made in the discipline.

Fit is not just global fit

When reporting the results of CFA, communication researchers typically include global fit indices for their measurement model, which provide information about how well the hypothesized measurement model fits the observed data (i.e. sample variances and covariances) on average. Global fit assessment should be evaluated based on four fit indices (Goodboy & Kline, 2017; Kline, 2016): the model χ^2 with degrees of freedom and significance test, standardized root mean squared residual (SRMR), Steiger-Lind root mean square error of approximation (RMSEA) with accompanying 90% confidence interval, and Bentler comparative fit index (CFI).

For the model χ^2, the null hypothesis is that the observed data are an exact fit to the specified measurement model. Quite often, the χ^2 will indicate the model deviates significantly from exact fit (i.e. $p < .05$) so the exact fit hypothesis is rejected. In turn, the significant deviation might lead researchers towards a more hopeful interpretation of the χ^2/df ratio, or the normed chi-square, from a faulty premise that normed chi-square ratios lower than an arbitrary cutoff (e.g. 3.0 or 2.0, falsely attributed to Hu and Bentler, 1999), allow for a measurement model to be retained. This widespread practice can be traced back to Wheaton et al. (1977) who mention χ^2/df briefly 'to get a rough indication of fit per degree of freedom' for their particular sample size. Somehow, the normed chi-square has invaded the communication discipline in reports of CFA, even after Wheaton (1987) himself explained 'it is unfortunate that this measure has gained such widespread use when even in the 1977 article it was not taken as the primary criterion for fit' (p. 127), and he advised 'I do not advocate use of the χ^2/df ratio at this stage' and 'there is no reason to replace the χ^2 test with this ratio' (p. 128). Nonetheless, the practice of reporting it for model fit remains widespread, but it does not actually provide compelling evidence for the validity of a measurement model. Put bluntly, there is no reason to report χ^2/df as evidence of model fit because it has no statistical or logical foundation and no established cutoff (Brown, 2015; Goodboy & Kline, 2017; Kline, 2016); the model χ^2 itself will suffice along with the other global fit indices that follow.

Following a significant model χ^2, the standardized root mean squared residual (SRMR) should be consulted next as an approximate fit index. The SRMR provides the average correlation residual and should always accompany the model χ^2. The measurement model yields approximate fit with an SRMR $\leq .08$ (accompanied by small residuals, discussed later) and is interpreted as 'if all correlations are equally misfitted, the model is approximately well fitting if the estimated and observed correlations are less than .08' (Asparouhov & Muthén, 2018, p. 2), Next, the RMSEA is reported, along with its 90% confidence interval, as a close fit hypothesis (.05-.08 might be a reasonable value, depending on model parameters). Finally, the CFI is reported as the relative fit improvement over a baseline independence model (where all model correlations are zero). A CFI $\geq .95$, which is recommended, is interpreted as 95% improvement over the baseline model, which is really just the worst possible measurement model that could exist. Each of these four estimators is critical for reporting the global fit of a CFA.

However, there is a preoccupation with global fit indices as the most critical for retaining a measurement model (Marsh, Hau, & Wen, 2004) and as result, scholars often ignore and omit local

fit from their CFA reports (Goodboy & Kline, 2017). Global fit provides information about the model on average, but it does not tell the researcher where there are specific problems in the measurement model. These problems are only discovered in analysis of local fit, which is evaluated by examining the residuals in the model.

Residuals are at the variable/item level and give the difference between the observed versus predicted values for all pairs of observed variables (i.e. difference between the covariance from the data versus the model-implied covariance in the CFA). They are calculated in four metrics: covariance, standardized, normalized, and correlation residuals. Standardized residuals are interpreted as *z* scores and have a significance test for misfit, often with the 'standard cutoff' of +/−2.58 as indicating significant local misfit ($p < .001$; Vieira, 2011, p. 61). In practice, eyeballing standardized residuals greater than |3| might suffice (a standardized residual of +/−1.96 would be significant at the $p < .05$ level but in large models, there are many residuals which are prone to type I error). Normalized residuals are more conservative local fit tests and are always smaller than standardized residuals; they share a similar interpretation. Correlation residuals are appealing because they are most easily interpreted in a familiar metric, with absolute values greater than .10 highlighting a specification error that could be addressed (Kline, 2016). Specification errors in CFA should be investigated based on sign of the residual indicating local misfit (positive residuals indicate underestimating and negative residuals overestimating the covariance between two measurement items), and can aid researchers in identifying poor fit (such as shared wording of items or reverse coded/negatively worded items; Bandalos, 2018). To accompany global fit, local fit must be reported and interpreted in any CFA, yet it continues to be overwhelmingly ignored in published communication articles.

Rampant model respecification

As stated earlier, it is inevitable that a communication researcher will use CFA and uncover poor global fit, especially when focused on the strict model χ^2. In order to achieve 'fit,' the researcher might then be tempted to revise and respecify the poor fitting model until it fits (sometimes guided by modification indices provided in the CFA outputs). Such modifications include deleting scale items *post-hoc*, adding error correlations between items (including items on the same primary factor, or between items on different factors; often guided by modification indices), or deleting problematic items with low factor loadings. Often, these respecifications are done without any regard to substantive interpretations derived from communication theory or published scholarship – that is, without considering the *a priori* model that drives the CFA in the first place. Rather, the respecifications are done simply to achieve statistical fit for the CFA. Such practices are inappropriate, as they *de facto* shift the research from engaging in confirmatory (hypothesis-testing) to exploratory (hypothesis-generating) research. Instead, the researcher should more critically consider why an *a priori* measurement model is not replicating in a given data collection – likely when the measure is modified to suit the researcher's needs in a specific study. These modifications include changing the wording of original scale items to fit a new context or specific referent, only including some of the items from a full measure, creating or adding new items to an existing measure without validating them, and changing the response scale, among other modifications (Heggestad et al., 2019). CFA is used to confirm communication theory, and theories are not typically modified post-hoc to work only in the researcher's sample.

Critically, established measures that are modified to achieve model fit exploit sample-specific variance. As a result, when communication researchers take established measures that have past validity evidence (i.e. have been subjected to CFA and provide strong factor loadings along with global and local fit in multiple samples) and modify the measure solely based statistical improvements in their sample (e.g. dropping items, correlating errors, etc.), the analysis becomes exploratory rather than confirmatory and capitalizes on sample specific variation. Major modifications to an existing measure (e.g. deleting items from a scale in a CFA to obtain model fit) should not be trusted as the true representation of the construct being measured, unless researchers then replicate the respecified factor structure of their scale in a new and independent sample. We cannot trust respecified

CFAs without theoretical justifications for any modifications to the measurement model, and without clear evidence from multiple samples and/or without theoretical or substantive justifications for alterations to the measurement, a poor fitting CFA should be rejected. The end goal of CFA should not be to discover a model that could fit one's data (any model can be 'forced' to fit observed data, with enough re-specifications) but rather, to test whether an *a priori* model is a fit for a dataset. Understand that *a priori* models are hypotheses – the predictions being that the to-be collected data should fit the specified model. When CFA results indicate poor fit and when there are no apparent and rational explanations for that poor fit, the most appropriate decision for the researcher is to accept the results as evidence that the *a priori* model is flawed; that is, to reject the hypothesized model. Robust models should be able to account for observed data without modifications.

Maximum likelihood estimation and multivariate normality

In structural equation modeling (SEM) software programs, normal theory maximum likelihood (ML) estimation is usually the default estimator, so most often, communication scholars use ML for a CFA with approximately interval quality data (at minimum, items use at least a 5-point response format). Rarely, however, do communication scholars acknowledge or test the ML's distributional assumption of multivariate normality for endogenous (dependent) variables in their continuous data (i.e. there is not substantial skewness or kurtosis for the items). As ML is a large sample estimator, given a large enough sample, minor skewness and kurtosis will not be an issue. However, moderate to severe skewness and kurtosis of the items will affect the results by increasing the model χ^2 and RMSEA while decreasing the CFI, which would lead to worse fit and potential rejection of the model (Brown, 2015). Moreover, with normality violations, standard errors tend to be lower for parameter estimates which inflates the number of significant parameters (i.e. type I error increases; Kline, 2012). A general rule of thumb is that univariate skewness and kurtosis values for items should be less than the absolute value of 2.0 using ML estimation (Bandalos, 2018). From a series of CFA simulations, Curran et al. (1996) recommended that 'obtained univariate values approaching at least 2.0 and 7.0 for skewness and kurtoses are suspect' (p. 26). But even minor deviations from multivariate normality can affect the measurement model.

One solution proposed is the use of a robust ML estimator, which can account for nonnormality by adjusting the global fit indices with a scaling correction factor and providing robust standard errors (Lei & Wu, 2012). Two of these robust ML estimators are MLM with the Satorra-Bentler scaled chi-square (S-Bχ^2; Satorra & Bentler, 1994) or MLR with the Yuan-Bentler residual-based chi-square (Y-Bχ^2, also known as T_2^* test statistic; Yuan & Bentler, 2000). Both of these robust ML will correct the model χ^2 and standard errors for nonnormal data, but the MLM estimator with the S-Bχ^2 cannot handle missing data, so a CFA with missing data will require the MLR estimator with the Y-Bχ^2 (see Byrne, 2012 for the capabilities of software programs that provide these estimators).

Robust ML estimators provide the same model fit as conventional ML when data are multivariate normal (Curran et al., 1996), but because many communication variables will not be normally distributed (e.g. relationship uncertainty in marriage, violence and aggression, video game and internet addiction, etc.), robust ML should be the default estimator for researchers reporting CFA. It is quite possible that many CFAs estimated by ML have yielded poor model fit in previous communication scholarship, not because the model was 'wrong,' but because the data were not multivariate normal. Therefore, for construct validity purposes, the discipline needs to embrace robust ML estimation because in many cases, the CFA will yield equal, if not better model fit depending on the degree of nonnormality.

The 'Split-Half' method of replicating EFA results using CFA on the same sample

Both EFA and CFA share a critical relationship with respect to establishing and testing measurement validity. As explained to this point in the manuscript, EFA is useful for exploring and distilling nascent measurement models from a list of candidate items (particularly useful with emerging or evolving

constructs), and CFA is useful for confirming a proposed or suggested *a priori* measurement model (with extreme caution suggested for any *post hoc* respecifications). As a cross-sectional and sample-specific technique by design, claims of measurement validity based on EFA alone are quite suspect, as it is plausible that they might not replicate in new scenarios – new sampling frames (shifting target populations), administrations (such as using a paper-and-pencil compared to a computer-assisted survey), or referent objects (using a scale to measure reactions to different types of stimulus) can all result in slight differences in both shared variance accounted for by factors, as well as unique or error variance attributed to different items in a measurement model. As such, preliminary EFA results in one sample require CFA results from another sample to provide robust evidence of construct validity.

Given the hypothesis-generating nature of EFA and the hypothesis-testing nature of CFA, as well as the requirement of the latter to (literally) confirm the former, many statistical analysis guides (cf. Bandalos, 2018; Kline, 2016) accept the use of a 'split-half' sample method in order to provide researchers with two data analysis sets from a single data collection. The basic argument here is that if one has a random sample (say, $N = 1000$) from a population of interest, then not only does the full dataset provide a robust estimate of data patterns in the population, but that a random sampling of that subset of that data (say, two $n = 500$ split samples classified using a fair coin) would also be a fair representation of the sample. To the extent that one sample is not made too small for meaningful statistical analysis (for a discussion of sample size requirements for factor analysis, see Mundfrom et al., 2009), splitting a larger sample into two smaller samples could provide the researcher with an original sample for conducting EFA, and then another equally representative sample for conducting CFA.

The argument above represents a scenario in which one has knowledge of and access to both a clearly defined population as well as a truly random sample of all members of that population. However, we know from both intuition and from analysis of research trends that this assumption rarely plays out. Erba et al. (2018) analyzed over 1000 published studies of mass communication in six leading communication journals and found a large majority to rely on nonprobability sampling (and over half relying on student samples); Sarstedt, Bengart, Shaltoni, and Lehmann (2018) found similar patterns in advertising research. Similar critiques could likely be found and/or levied against most areas of communication and related research. From these observations, the assumptions for which the split-half technique rests can be challenged, as well as the validity of such a critique. Here, Flora and Flake (2017) reminded us that

> it is not logical to obtain a good-fitting factor structure using EFA and then seek to confirm that structure using CFA with the same dataset; doing so capitalizes on sample-specific, chance relationships and in no way verifies the EFA findings. (p. 85)

Related to this are concerns with using the split-half method on data collected using nonprobability sampling, such as convenience samples common in communication research. If one randomly splits a larger convenient sample into two smaller samples, the same systematic biases and sample-specific variation will likely be present in both samples – there is little surprise that the factor structure from one half of the data will be reflected in the factor structure from the other half of the data. Indeed, a lack of fit demonstrated by the CFA might be more indicative of a failed randomization (or perhaps non-normally distributed measurement items) than a rejection of the EFA. Here, we also acknowledge that the authors of this manuscript have different approaches with respect to the veracity of the 'split-half' method: Bowman is overall critical of the method given the arguments above, while Goodboy recognizes that with sample size restrictions, splitting a sample for EFA/CFA is not ideal but is still defensible since the participant data do not overlap in the analyses. That said, both of us agree (as do numerous others) that the most robust method of establishing construct validity of a scale with only EFA evidence is to (a) collect additional data for a new and independent sample and (b) use CFA techniques to confirm or disconfirm prior EFA results on the data from this sample.

Exploratory structural equation modeling as an emerging validity testing technique

Although EFA and CFA are important statistical techniques for item-level analyses of latent variables, they both have different limitations in terms of how the measurement model is analyzed (see Kline, 2013 for a comparison of these differences). CFA uses a restricted model requiring no item cross loadings on factors which leads researchers to specify many zero loadings in a multidimensional measurement model. In that regard, CFA is clearly a punishing analysis, especially for communication measures where zero cross-loadings is not tenable. This can lead to misspecifications in the model when cross-loadings are actually nonzero, which might yield poor global and local fit. This poor model fit can encourage researchers to respecify the CFA measurement model in ways that capitalize on sample specific variation by consulting sets of modifications indices. Recognizing this, Asparouhov and Muthén (2009) argued that 'although technically appealing, CFA requires strong measurement science that is often not available in practice' and 'the use of CFA measurement modeling in SEM has disadvantages and these are likely to have contributed to poor applications of SEM where the believability and replicability of the final model is in doubt' (p. 398).

In response, Asparouhov and Muthén (2009) offered exploratory structural equation modeling (ESEM) as an alternative to traditional CFA models, featuring a less restrictive measurement model that allows for freely estimated cross-loadings and factor loading rotation as is done in EFA measurement models (often, using geomin rotations), but allowing for parameter estimates offered in CFA including standard errors, fit statistics, model fit comparisons, and tests of invariance, among other advantageous possibilities (Morin et al., 2013). Thus, it is a flexible alternative to the traditional EFA and CFA procedures that communication scholars typically use and it can be better suited for validity testing purposes, especially when measurement instruments have many factors and items (which assume zero item cross-loadings) or when theory requires a more flexible factor model. This is important for discriminant validity purposes, because factor correlations can be inflated when cross-loadings that are not zero are fixed to zero as in traditional CFA (Morin et al., 2013). Although it is not commonplace yet, communication scholars should welcome the flexibilities of ESEM for validity testing as has been done in other social sciences. By doing so, the validity of measurement models may be enhanced by drawing upon the advantages of EFA but within the general structural equation modeling framework – two examples in communication research can be found with Bowman et al. (2018) and Banks et al. (2019).

Conclusions and final recommendations

Assessing construct validity is a critically important task for any quantitative science, as research claims are only as robust as the data they were based on. Poor measurement misrepresents the social phenomena being studied, and a lack of construct validity could lend support to inaccurate theories and distort accurate theories. The discussion of construct validity is far more complex than one manuscript, but we offer two core recommendations aimed at helping all of us ensure our field uses measures with strong construct validity. First, we urge a focus on the five sources of validity when designing self-report survey measures, including an in-depth accounting of those sources whenever possible (such as in nascent scale validation papers). Second, we urge more careful attention to the foundational mechanics of popular (and emerging) factor analysis techniques that represent the intentions of those statistical analyses, rather than the desired factor-analytic outcomes of the researchers. Adopting both recommendations in practice and in pedagogy will ensure that our field is one that produces useful data into the complicated study of human communication.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Nicholas David Bowman http://orcid.org/0000-0001-5594-9713

References

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

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus. http://www.statmodel.com/download/SRMR2.pdf.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Banks, J., Bowman, N. D., Lin, J.-H., Pietschmann, D., & Wasserman, J. (2019). The common Player-Avatar Interaction scale (cPAX): Expansion and cross-language validation. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2019.03.003

Bauer, G. R., Braimoh, J., Scheim, A. I., Dharma, C., & Dalby, A. R. (2017). Transgender-inclusive measures of sex/gender for population surveys: Mixed-methods evaluation and recommendations. *PLoS One*, *12*(5), https://doi.org/10.1371/journal.pone.0178043

Bowman, N. D., Wasserman, J., & Banks, J. (2018). Development of the video game Demand scale. In N. D. Bowman (Ed.), *Video games: A medium that demands our attention* (pp. 208–233). Routledge.

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing and Evaluation SIG Newsletter*, *13*(3), 20–25.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, *24*(4), 445–455. https://doi.org/10.1207/s15327906mbr2404_4

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). SAGE.

Byrne, B. M. (2012). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). 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

Cerny, B. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, *12*(1), 43–47. https://doi.org/10.1207/s15327906mbr1201_3

Čertický, M., Čertický, M., Sinčák, R., Magyar, G., Vaščák, J., & Cavallo, F. (2019). Psychophysiological indicators for modeling user experience in interactive digital entertainment. *Sensors*, *19*(5), 989. https://doi.org/10.3390/s19050989

Costello, A. B., & Osborne, J. W. (2005). Best practice in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), https://pareonline.net/getvn.asp?v=10%26n=7

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. https://doi.org/10.1080/02664763.211.610445

Erba, J., Ternes, B., Bobkowski, P., Logan, T., & Liu, Y. (2018). Sampling methods and sample populations in quantitative mass communication research studies: A 15-year census of six journals. *Communication Research Reports*, *35*(1), 42–47. https://doi.org/10.1080/08824096.2017.1362632

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

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue Canadienne des Sciences du Comportement*, *49*(2), 78–88. https://doi.org/10.1037/cbs0000069

Fraser, G. (2018). Evaluating inclusive gender identity measures for use in quantitative psychological research. *Psychology & Sexuality*, *9*(4), 343–357. doi:10.1080/19419899.2018.1497693

Goodboy, A. K., & Kline, R. B. (2017). Statistical and practical concerns with published communication research featuring structural equation modeling. *Communication Research Reports*, *34*(1), 68–77. https://doi.org/10.1080/08824096.2016.1214121

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

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393–416. https://doi.org/10.1177/0013164405282485

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

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

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). Guilford Press.

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher, & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171–207). Routledge.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed). Guilford Press.

Lei, P., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–179). Guilford Press.

Levine, T. R. (2005). Confirmatory factor analysis and scale validation in communication research. *Communication Research Reports*, *22*(4), 335–338. https://doi.org/10.1080/00036810500317730

MacCallum, R. C., Wideman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Bulletin*, *4*, 84–99. https://doi.org/10.1037/1082-989X.4.1.84

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 over generalizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341. doi:10.1207/s15328007sem1103_2

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1), 97–110. https://www.redalyc.org/pdf/2990/299023509007.pdf. https://doi.org/10.21500/20112084.854

McCroskey, J. C., & Young, T. J. (1979). The use and abuse of factor analysis in communication research. *Human Communication Research*, *5*(4), 375–382. https://doi.org/10.1111/j.1468-2958.1979.tb00651.x

Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 267–308). Information Age Publishing.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2009). Minimum sample size recommendations for conducting factor analysis. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Oliver, M. B., Bowman, N. D., Woolley, J. K., Rogers, R., Sherrick, B. I., & Chung, M.-Y. (2016). Video games as meaningful entertainment experiences. *Psychology of Popular Media Culture*, *5*(4), 390–405. https://doi.org/10.1037/ppm0000066

Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. Self-published.

Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principle components analysis in communication research. *Human Communication Research*, *28*(4), 562–577. https://doi.org/10.1111/j.1468-2958.2002.tb00824.x

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

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

Spearman, C. (1904). General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Tobias, S., & Carlson, J. E. (1969). Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate Behavioral Research*, *4*(3), 375–377. https://doi.org/10.1207/s15327906mbr0403_8

Vieira, A. L. (2011). *Interactive LISREL in practice: Getting started with a SIMPLIS approach*. Springer.

Watkins, M. (2008). Monte Carlo for PCA parallel analysis (Version 2.3). www.softpedia.com/get/Others/HomeEducation/Monte-Carlo-PCA-for-ParallelAnalysis.shtml.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608–618. https://doi.org/10.1037/0022-3514.84.3.608

Wheaton, B. (1987). Assessment of fit in overidentified models with latent variables. *Sociological Methods & Research*, *16*(1), 118–154. https://doi.org/10.1177/0049124187016001005

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology* (pp. 84–136). Josey-Bass.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. State University of New York Press.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 167–202. https://doi.org/10.1111/0081-1750.00078

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*(2), 253–269. doi:10.1207/s15327906mbr1702_5