

# Enhancing Assessment Literacy in Professional Counseling: A Practical Overview of Factor Analysis



*The Professional Counselor*<sup>™</sup>  
Volume 11, Issue 3, Pages 267–284  
<http://tpcjournal.nbcc.org>  
© 2021 NBCC, Inc. and Affiliates  
doi: 10.15241/mtk.11.3.267

Michael T. Kalkbrenner

Assessment literacy is an essential competency area for professional counselors who administer tests and interpret the results of participants' scores. Using factor analysis to demonstrate internal structure validity of test scores is a key element of assessment literacy. The underuse of psychometrically sound instrumentation in professional counseling is alarming, as a careful review and critique of the internal structure of test scores is vital for ensuring the integrity of clients' results. A professional counselor's utilization of instrumentation without evidence of the internal structure validity of scores can have a number of negative consequences for their clients, including misdiagnoses and inappropriate treatment planning. The extant literature includes a series of articles on the major types and extensions of factor analysis, including exploratory factor analysis, confirmatory factor analysis (CFA), higher-order CFA, and multiple-group CFA. However, reading multiple psychometric articles can be overwhelming for professional counselors who are looking for comparative guidelines to evaluate the validity evidence of scores on instruments before administering them to clients. This article provides an overview for the layperson of the major types and extensions of factor analysis and can serve as reference for professional counselors who work in clinical, research, and educational settings.

**Keywords:** Factor analysis, overview, professional counseling, internal structure, validity

Professional counselors have a duty to ensure the veracity of tests before interpreting the results of clients' scores because clients rely on their counselors to administer and interpret the results of tests that accurately represent their lived experience (American Educational Research Association [AERA] et al., 2014; National Board for Certified Counselors [NBCC], 2016). Internal structure validity of test scores is a key assessment literacy area and involves the extent to which the test items cluster together and represent the intended construct of measurement.

Factor analysis is a method for testing the internal structure of scores on instruments in professional counseling (Kalkbrenner, 2021b; Mvududu & Sink, 2013). The rigor of quantitative research, including psychometrics, has been identified as a weakness of the discipline, and instrumentation with sound psychometric evidence is underutilized by professional counselors (Castillo, 2020; C.-C. Chen et al., 2020; Mvududu & Sink, 2013; Tate et al., 2014). As a result, there is an imperative need for assessment literacy resources in the professional counseling literature, as assessment literacy is a critical competency for professional counselors who work in clinical, research, and educational settings alike.

## Assessment Literacy in Professional Counseling

Assessment literacy is a crucial proficiency area for professional counselors, as counselors in a variety of the specialty areas of the Council for Accreditation of Counseling and Related Educational Programs (2015), such as clinical rehabilitation (5.D.1.g. & 5.D.3.a.), clinical mental health (5.C.1.e. & 5.C.3.a.), and addiction (5.A.1.f. & 5.A.3.a.), select and administer tests to clients and use the results to inform diagnosis and treatment planning, and to evaluate the utility of clinical interventions (Mvududu & Sink,

---

Michael T. Kalkbrenner, PhD, NCC, is an associate professor at New Mexico State University. Correspondence may be addressed to Michael T. Kalkbrenner, Department of Counseling and Educational Psychology, New Mexico State University, Las Cruces, NM 88003, [mkalk001@nmsu.edu](mailto:mkalk001@nmsu.edu).

2013; NBCC, 2016; Neukrug & Fawcett, 2015). The extant literature includes a series of articles on factor analysis, including exploratory factor analysis (EFA; Watson, 2017), confirmatory factor analysis (CFA; Lewis, 2017), higher-order CFA (Credé & Harms, 2015), and multiple-group CFA (Dimitrov, 2010). However, reading several articles on factor analysis is likely to overwhelm professional counselors who are looking for a desk reference and/or comparative guidelines to evaluate the validity evidence of scores on instruments before administering them to clients. To these ends, professional counselors need a single resource (“one-stop shop”) that provides a brief and practical overview of factor analysis. The primary purpose of this manuscript is to provide an overview for the layperson of the major types and extensions of factor analysis that counselors can use as a desk reference.

## Construct Validity and Internal Structure

*Construct validity*, the degree to which a test measures its intended theoretical trait, is a foundation of assessment literacy for demonstrating validity evidence of test scores (Bandalos & Finney, 2019). *Internal structure validity*, more specifically, is an essential aspect of construct validity and assessment literacy. Internal structure validity is vital for determining the extent to which items on a test combine to represent the construct of measurement (Bandalos & Finney, 2019). *Factor analysis* is a key method for testing the internal structure of scores on instruments in professional counseling as well as in social sciences research in general (Bandalos & Finney, 2019; Kalkbrenner, 2021b; Mvududu & Sink, 2013). In the following sections, I will provide a practical overview of the two primary methodologies of factor analysis (EFA and CFA) as well as the two main extensions of CFA (higher-order CFA and multiple-group CFA). These factor analytic techniques are particularly important elements of assessment literacy for professional counselors, as they are among the most common psychometric analyses used to validate scores on psychological screening tools (Kalkbrenner, 2021b). Readers might find it helpful to refer to Figure 1 before reading further to become familiar with some common psychometric terms that are discussed in this article and terms that also tend to appear in the measurement literature.

**Figure 1**

### *Technical and Layperson’s Definitions of Common Psychometric Terms*

Psychometric Term	Technical Definition	Layperson’s Definition
Confirmatory Factor Analysis	Theory-testing strategy for appraising the <i>internal structure</i> of a hypothesized factor solution with a new sample.	A statistical test for evaluating the <i>internal structure</i> of scores on an existing instrument when it is given to a new sample.
Construct Validity	A type of <i>validity</i> evidence involving the extent to which test scores accurately appraise the intended <i>latent variable</i> .	A form of <i>validity</i> for determining if scores on a test collectively measure what the test developers claim it measures.
Dimensionality	See <i>internal structure</i> .	See <i>internal structure</i> .
Exploratory Factor Analysis	Psychometric test for reducing an inter-item correlation matrix and uncovering the <i>internal structure</i> of scores on a new or previously untested instrument.	A psychometric test for uncovering the <i>internal structure</i> of scores on a new test.
Factor(s)	A collection of test items that cluster together based on conceptually related dimensions to form a composite <i>latent variable</i> . Factors are also referred to as “subscales.”	The overall <i>latent variable</i> that is created by a series of test items that are grouped together because they share a common meaning.
Factor Analysis	Multivariate statistical analyses for uncovering and confirming the <i>internal structure</i> of test scores.	A series of tests for evaluating the <i>internal structure</i> of test scores.

Psychometric Term	Technical Definition	Layperson's Definition
Factorial Validity	See <i>internal structure</i> .	See <i>internal structure</i> .
Higher-Order Confirmatory Factor Analysis	Extension of <i>confirmatory factor analysis</i> for examining nested models and determining if a second-order (or beyond) <i>latent variable</i> explains the co-variation between the single-order factors.	A type of <i>confirmatory factor analysis</i> for identifying if the relationship between <i>factors</i> is explained by a larger and more general <i>latent variable</i> .
Internal Structure	A method for measuring <i>construct validity</i> that involves the degree to which the relationships among test items and test components conform to the construct that the proposed test score interpretations are based upon.	One way to examine the <i>construct validity</i> of test scores by evaluating how, if at all, and in what ways the test questions group together to form <i>factors</i> .
Latent Variable	Theoretical or abstract traits that are inferred based on the compilation of scores on a series of <i>observed variables</i> .	A variable that cannot be measured directly by one test question; for example, a group of <i>observed variables</i> about temperament might collectively measure the latent variable of personality.
Multiple-Group Confirmatory Factor Analysis	Extension of <i>confirmatory factor analysis</i> for examining the invariance (psychometric equivalence) of instrumentation across subgroups of a larger sample or population.	A type of <i>confirmatory factor analysis</i> for evaluating if the scales/subscales of a test have the same meaning with smaller groups of a larger sample.
Observed Variable	Data that is directly measured, usually by one test item.	Information that can be gathered directly from a single test question; for example, asking a test taker to specify their age.
Psychometrics	The field of study centered on the theory and practice of psychological measurement.	Approaches and strategies for measuring the mental and emotional states of human beings.
Validity	The degree to which the inferences made from test scores accurately reflect the test taker's experiences.	The extent to which a test actually measures what the test developers claim it measures.

Note. *Italicized* terms are defined in this figure.

## Exploratory Factor Analysis

EFA is “exploratory” in that the analysis reveals how, if at all, test items band together to form factors or subscales (Mvududu & Sink, 2013; Watson, 2017). EFA has utility for testing the factor structure (i.e., how the test items group together to form one or more scales) for newly developed or untested instruments. When evaluating the rigor of EFA in an existing psychometric study or conducting an EFA firsthand, counselors should consider sample size, assumption checking, preliminary testing, factor extraction, factor retention, factor rotation, and naming rotated factors (see Figure 2).

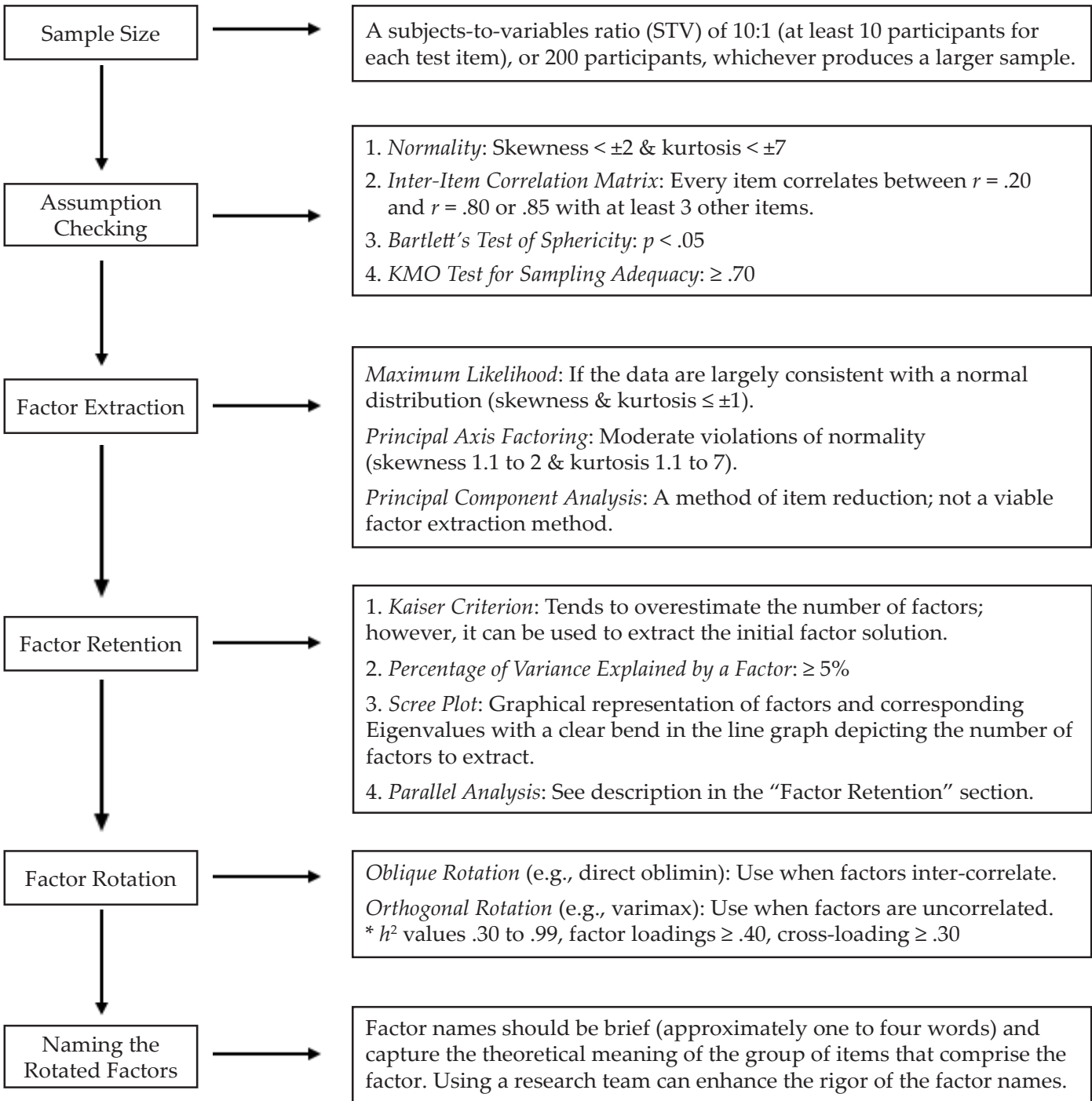
### *EFA: Sample Size, Assumption Checking, and Preliminary Testing*

Researchers should carefully select the minimum sample size for EFA before initiating data collection (Mvududu & Sink, 2013). My 2021 study (Kalkbrenner, 2021b) recommended that the minimal a priori sample size for EFA include either a subjects-to-variables ratio (STV) of 10:1 (at least 10 participants for each test item) or 200 participants, whichever produces a larger sample. EFA tends to be robust to moderate violations of normality; however, results are enriched if data are normally distributed (Mvududu & Sink, 2013). A review of skewness and kurtosis values is one way to test for univariate normality; according to Dimitrov (2012), extreme deviations from normality include skewness values  $> \pm 2$  and kurtosis  $> \pm 7$ ; however, ideally these values are  $\leq \pm 1$  (Mvududu & Sink, 2013). The Shapiro-Wilk and Kolmogorov-Smirnov tests can also be computed to test for normality, with non-significant  $p$ -values indicating that the parametric properties of the data are not statistically different from a normal

distribution (Field, 2018); however, the Shapiro-Wilk and Kolmogorov-Smirnov tests are sensitive to large sample sizes and should be interpreted cautiously. In addition, the data should be tested for linearity (Mvududu & Sink, 2013). Furthermore, extreme univariate and multivariate outliers must be identified and dealt with (i.e., removed, transformed, or winsorized; see Field, 2018) before a researcher can proceed with factor analysis. Univariate outliers can be identified via z-scores ( $> 3.29$ ), box plots, or scatter plots, and multivariate outliers can be discovered by computing Mahalanobis distance (see Field, 2018).

**Figure 2**

*Flow Chart for Reviewing Exploratory Factor Analysis*





Three preliminary tests are necessary to determine if data are factorable, including (a) an inter-item correlation matrix, (b) the Kaiser–Meyer–Olkin (KMO) test for sampling adequacy, and (c) Bartlett’s test of sphericity (Beavers et al., 2013; Mvududu & Sink, 2013; Watson, 2017). The purpose of computing an inter-item correlation matrix is to identify redundant items (highly correlated) and individual items that do not fit with any of the other items (weakly correlated). An inter-item correlation matrix is factorable if a number of correlation coefficients for each item are between approximately  $r = .20$  and  $r = .80$  or  $.85$  (Mvududu & Sink, 2013; Watson, 2017). Generally, a factor or subscale should be composed of at least three items (Mvududu & Sink, 2013); thus, an item should display intercorrelations between  $r = .20$  and  $r = .80/.85$  with at least three other items. However, inter-item correlations in this range with five to 10+ items are desirable (depending on the total number of items in the inter-item correlation matrix).

Bartlett’s test of sphericity is computed to test if the inter-item correlation matrix is an identity matrix, in which the correlations between the items is zero (Mvududu & Sink, 2013). An identity matrix is completely unfactorable (Mvududu & Sink, 2013); thus, desirable findings are a significant  $p$ -value, indicating that the correlation matrix is significantly different from an identity matrix. Finally, before proceeding with EFA, researchers should compute the KMO test for sampling adequacy, which is a measure of the shared variance among the items in the correlation matrix (Watson, 2017). Kaiser (1974) suggested the following guidelines for interpreting KMO values: “in the .90s – marvelous, in the .80s – meritorious, in the .70s – middling, in the .60s – mediocre, in the .50s – miserable, below .50 – unacceptable” (p. 35).

### **Factor Extraction Methods**

Factor extraction produces a factor solution by dividing up shared variance (also known as common variance) between each test item from its unique variance, or variance that is not shared with any other variables, and error variance, or variation in an item that cannot be accounted for by the factor solution (Mvududu & Sink, 2013). Historically, principal component analysis (PCA) was the dominant factor extraction method used in social sciences research. PCA, however, is now considered a method of data reduction rather than an approach to factor analysis because PCA extracts all of the variance (shared, unique, and error) in the model. Thus, although PCA can reduce the number of items in an inter-item correlation matrix, one cannot be sure if the factor solution is held together by shared variance (a potential theoretical model) or just by random error variance.

More contemporary factor extraction methods that only extract shared variance—for example, principal axis factoring (PAF) and maximum likelihood (ML) estimation methods—are generally recommended for EFA (Mvududu & Sink, 2013). PAF has utility if the data violate the assumption of normality, as PAF is robust to modest violations of normality (Mvududu & Sink, 2013). If, however, data are largely consistent with a normal distribution (skewness and kurtosis values  $\leq \pm 1$ ), researchers should consider using the ML extraction method. ML is advantageous, as it computes the likelihood that the inter-item correlation matrix was acquired from a population in which the extracted factor solution is a derivative of the scores on the items (Watson, 2017).

**Factor Retention.** Once a factor extraction method is deployed, psychometric researchers are tasked with retaining the most parsimonious (simple) factor solution (Watson, 2017), as the purpose of factor analysis is to account for the maximum proportion of variance (ideally, 50%–75%+) in an inter-item correlation matrix while retaining the fewest possible number of items and factors (Mvududu & Sink, 2013). Four of the most commonly used criteria for determining the appropriate number of factors to retain in social sciences research include the (a) Kaiser criterion, (b) percentage of variance among items explained by each factor, (c) scree plot, and (d) parallel analysis (Mvududu

& Sink, 2013; Watson, 2017). Kaiser's criterion is a standard for retaining factors with Eigenvalues (EV)  $\geq 1$ . An EV represents the proportion of variance that is explained by each factor in relation to the total amount of variance in the factor matrix.

The Kaiser criterion tends to overestimate the number of retainable factors; however, this criterion can be used to extract an initial factor solution (i.e., when computing the EFA for the first time). Interpreting the percentage of variance among items explained by each factor is another factor retention criterion based on the notion that a factor must account for a large enough percentage of variance to be considered meaningful (Mvududu & Sink, 2013). Typically, a factor should account for at least 5% of the variance in the total model. A scree plot is a graphical representation or a line graph that depicts the number of factors on the X-axis and the corresponding EVs on the Y-axis (see Figure 6 in Mvududu & Sink, 2013, p. 87, for a sample scree plot). The cutoff for the number of factors to retain is portrayed by a clear bend in the line graph, indicating the point at which additional factors fail to contribute a substantive amount of variance to the total model. Finally, in a parallel analysis, EVs are generated from a random data set based on the number of items and the sample size of the real (sample) data. The factors from the sample data with EVs larger than the EVs from the randomly generated data are retained based on the notion that these factors explain more variance than would be expected by random chance. In some instances, these four criteria will reveal different factor solutions. In such cases, researchers should retain the simplest factor solution that makes both statistical and substantive sense.

**Factor Rotation.** After determining the number of factors to retain, researchers seek to uncover the association between the items and the factors or subscales (i.e., determining which items load on which factors) and strive to find simple structure or items with high factor loadings (close to  $\pm 1$ ) on one factor and low factor loadings (near zero) on the other factors (Watson, 2017). The factors are rotated on vectors to enhance the readability or detection of simple structure (Mvududu & Sink, 2013). Orthogonal rotation methods (e.g., varimax, equamax, and quartimax) are appropriate when a researcher is measuring distinct or uncorrelated constructs of measurement. However, orthogonal rotation methods are rarely appropriate for use in counseling research, as counselors almost exclusively appraise variables that display some degree of inter-correlation (Mvududu & Sink, 2013). Oblique rotation methods (e.g., direct oblimin and promax) are generally more appropriate in counseling research, as they allow factors to inter-correlate by rotating the data on vectors at angles less than  $90^\circ$ . The nature of oblique rotations allows the total variance accounted for by each factor to overlap; thus, the total variance explained in a post-oblique rotated factor solution can be misleading (Bandalos & Finney, 2019). For example, the total variance accounted for in a post-oblique rotated factor solution might add up to more than 100%. To this end, counselors should report the total variance explained by the factor solution before rotation as well as the sum of each factor's squared structure coefficient following an oblique factor rotation.

Following factor rotation, researchers examine a number of factor retention criteria to determine the items that load on each factor (Watson, 2017). Commonality values ( $h^2$ ) represent the proportion of variance that the extracted factor solution explains for each item. Items with  $h^2$  values that range between .30 and .99 should be retained, as they share an adequate amount of shared variance with the other items and factors (Watson, 2017). Items with small  $h^2$  values ( $< .30$ ) should be considered for removal. However, commonality values should not be too high ( $\geq 1$ ), as this suggests one's sample size was insufficient or too many factors were extracted (Watson, 2017). Items with problematic  $h^2$  values should be removed one at a time, and the EFA should be re-computed after each removal because these values will fluctuate

following each deletion. Oblique factor rotation methods produce two matrices, including the *pattern matrix*, which displays the relationship between the items and a factor while controlling for the items' association with the other factors, and the *structure matrix*, which depicts the correlation between the items and all of the factors (Mvududu & Sink, 2013). Researchers should examine both the pattern and the structure matrices and interpret the one that displays the clearest evidence of simple structure with the least evidence of cross-loadings.

Items should display a factor loading of at least  $\geq .40$  ( $\geq .50$  is desirable) to mark a factor. Items that fail to meet a minimum factor loading of  $\geq .40$  should be deleted. Cross-loading is evident when an item displays factor loadings  $\geq .30$  to  $.35$  on two or more factors (Beavers et al., 2013; Mvududu & Sink, 2013; Watson, 2017). Researchers may elect to assign a variable to one factor if that item's loading is  $.10$  higher than the next highest loading. Items that cross-load might also be deleted. Once again, items should be deleted one at a time and the EFA should be re-computed after each removal.

### ***Naming the Rotated Factors***

The final step in EFA is naming the rotated factors; factor names should be brief (approximately one to four words) and capture the theoretical meaning of the group of items that comprise the factor (Mvududu & Sink, 2013). This is a subjective process, and the literature is lacking consistent guidelines for the process of naming factors. A research team can be incorporated into the process of naming their factors. Test developers can separately name each factor and then meet with their research team to discuss and eventually come to an agreement about the most appropriate name for each factor.

### **Confirmatory Factor Analysis**

CFA is an application of structural equation modeling for testing the extent to which a hypothesized factor solution (e.g., the factor solution that emerged in the EFA or another existing factor solution) demonstrates an adequate fit with a different sample (Kahn, 2006; Lewis, 2017). When validating scores on a new test, investigators should compute both EFA and CFA with two different samples from the same population, as the emergent internal structure in EFA can vary substantially. Researchers can collect two sequential samples or they may elect to collect one large sample and divide it into two smaller samples, one for EFA and the second for CFA.

Evaluating model fit in CFA is a complex task that is typically determined by examining the collective implications of multiple goodness-of-fit (GOF) indices, which include absolute, incremental, and parsimonious (Lewis, 2017). *Absolute fit indices* evaluate the extent to which the hypothesized model or the dimensionality of the existing measure fits with the data collected from a new sample. *Incremental fit indices* compare the improvement in fit between the hypothesized model and a null model (also referred to as an independence model) in which there is no correlation between observed variables. *Parsimonious fit indices* take the model's complexity into account by testing the extent to which model fit is improved by estimating fewer pathways (i.e., creating a more parsimonious or simple model). Psychometric researchers generally report a combination of absolute, incremental, and parsimonious fit indices to demonstrate acceptable model fit (Mvududu & Sink, 2013). Table 1 includes tentative guidelines for interpreting model fit based on the synthesized recommendations of leading psychometric researchers from a comprehensive search of the measurement literature (Byrne, 2016; Dimitrov, 2012; Fabrigar et al., 1999; Hooper et al., 2008; Hu & Bentler, 1999; Kahn, 2006; Lewis, 2017; Mvududu & Sink, 2013; Schreiber et al., 2006; Worthington & Whittaker, 2006).

**Table 1***Fit Indices and Tentative Thresholds for Evaluating Model Fit*

Absolute Fit Indices				
<i>Index</i>	<i>Abbreviation</i>	<i>Strong Fit</i>	<i>Acceptable Fit</i>	<i>Poor Fit</i>
Chi-square	$\chi^2$ or CMIN	$p$ -value > .05 or $\chi^2$ to $df \leq 1$	$\chi^2$ to $df \leq 2$ or 3	$\chi^2$ to $df > 3$
Standardized root mean square residual	SRMR	< .05	.06 to .08	> .08
Root mean square error of approximation	RMSEA	< .05, report confidence interval	.06 to .08 (.081 to .10 can denote a somewhat acceptable fit)	> .10
Goodness-of-fit index & Adjusted goodness-of-fit index	GFI/AGFI	$\geq .97$	$\geq .95$ ( $\geq .90$ to .94 can denote a somewhat acceptable fit)	< .90
Incremental Fit Indices				
<i>Index</i>	<i>Abbreviation</i>	<i>Strong Fit</i>	<i>Acceptable Fit</i>	<i>Poor Fit</i>
Comparative fit index	CFI	$\geq .97$	.95 to .90	< .90
Normed fit index	NFI	$\geq .97$	.95 to .90	< .90
Incremental fit index	IFI	$\geq .97$	.95 to .90*	< .90
Tucker–Lewis index	TLI	$\geq .97$	.95 to .90	< .90
Parsimonious Fit Indices				
<i>Index</i>	<i>Abbreviation</i>	<i>Strong Fit</i>	<i>Acceptable Fit</i>	<i>Poor Fit</i>
Parsimony-adjusted goodness-of-fit index	PGFI	Parsimony-adjusted indices range from 0 to 1 and have utility for making comparisons between different models. Values closer to 1 indicate a stronger fit.		
Parsimony-adjusted normed fit index	PNFI			

*Note.* The fit indices and benchmarks to estimate the degree of model fit in this table are offered as tentative guidelines for scores on attitudinal measures based on the synthesized recommendations of numerous psychometric researchers (see citations in the “Confirmatory Factor Analysis” section of this article). The list of fit indices in this table are not all-inclusive (i.e., not all of them are typically reported). There is no universal approach for determining which fit indices to investigate nor are there any absolute thresholds for determining the degree of model fit. No single fit index is sufficient for determining model fit. Researchers are tasked with selecting and interpreting fit indices holistically (i.e., collectively), in ways that make both statistical and substantive sense based on their construct of measurement and goals of the study. \*.90 to .94 can denote an acceptable model fit for incremental fit indices; however, the majority of values should be  $\geq .95$ .

**Model Respecification**

The results of a CFA might reveal a poor or unacceptable model fit (see Table 1), indicating that the dimensionality of the hypothesized model that emerged from the EFA was not replicated or confirmed with a second sample (Mvududu & Sink, 2013). CFA is a rigorous model-fitting procedure and poor model fit in a CFA might indicate that the EFA-derived factor solution is insufficient for appraising the construct of measurement. CFA, however, is a more stringent test of structural validity than EFA,



and psychometric researchers sometimes refer to the modification indices (also referred to as Lagrange multiplier statistics), which denote the expected decrease in the  $X^2$  value (i.e., degree of improvement in model fit) if the parameter is freely estimated (Dimitrov, 2012). In these instances, correlating the error terms between items or removing problematic items will improve model fit; however, when considering model respecification, psychometric researchers should proceed cautiously, if at all, as a strong theoretical justification is necessary to defend model respecification (Byrne, 2016; Lewis, 2017; Schreiber et al., 2006). Researchers should also be clear that model respecification causes the CFA to become an EFA because they are investigating the dimensionality of a different or modified model rather than confirming the structure of an existing, hypothesized model.

### *Higher-Order CFA*

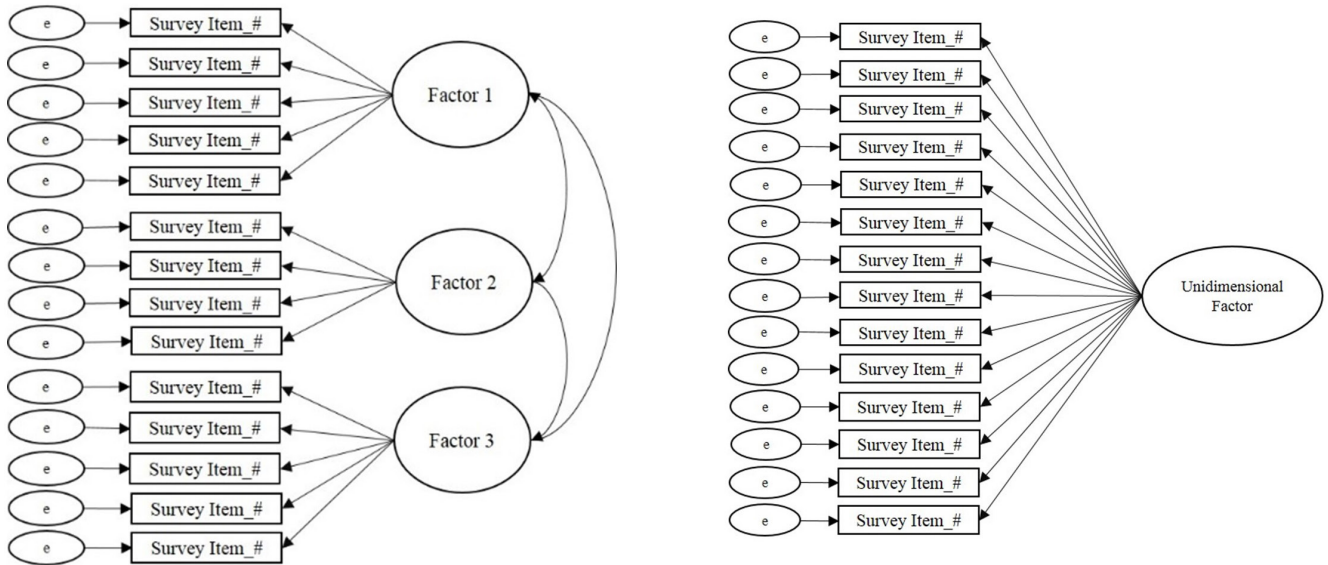
Higher-order CFA is an extension of CFA that allows researchers to test nested models and determine if a second-order latent variable (factor) explains the associations between the factors in a single-order CFA (Credé & Harms, 2015). Similar to single-order CFA (see Figure 3, Model 1) in which the test items cluster together to form the factors or subscales, higher-order CFA reveals if the factors are related to one another strongly enough to suggest the presence of a global factor (see Figure 3, Model 3). Suppose, for example, the test developer of a scale for measuring dimensions of the therapeutic alliance confirmed the three following subscales via single-order CFA (see Figure 3, Model 1): Empathy, Unconditional Positive Regard, and Congruence. Computing a higher-order CFA would reveal if a higher-order construct, which the research team might name Therapeutic Climate, is present in the data. In other words, higher-order CFA reveals if Empathy, Unconditional Positive Regard, and Congruence, collectively, comprise the second-order factor of Therapeutic Climate.

Determining if a higher-order factor explains the co-variation (association) between single-order factors is a complex undertaking. Thus, researchers should consider a number of criteria when deciding if their data are appropriate for higher-order CFA (Credé & Harms, 2015). First, moderate-to-strong associations (co-variance) should exist between first-order factors. Second, the unidimensional factor solution (see Figure 3, Model 2) should display a poor model fit (see Table 1) with the data. Third, theoretical support should exist for the presence of a higher-order factor. Referring to the example in the previous paragraph, person-centered therapy provides a theory-based explanation for the presence of a second-order or global factor (Therapeutic Climate) based on the integration of the single-order factors (Empathy, Unconditional Positive Regard, and Congruence). In other words, the presence of a second-order factor suggests that Therapeutic Climate explains the strong association between Empathy, Unconditional Positive Regard, and Congruence.

Finally, the single-order factors should display strong factor loadings (approximately  $\geq .70$ ) on the higher-order factor. However, there is not an absolute consensus among psychometric researchers regarding the criteria for higher-order CFA and the criteria summarized in this section are not a dualistic decision rule for retaining or rejecting a higher-order model. Thus, researchers are tasked with presenting that their data meet a number of criteria to justify the presence of a higher-order factor. If the results of a higher-order CFA reveal an acceptable model fit (see Table 1), researchers should directly compare (e.g., chi-squared test of difference) the single-order and higher-order models to determine if one model demonstrates a superior fit with the data at a statistically significant level.

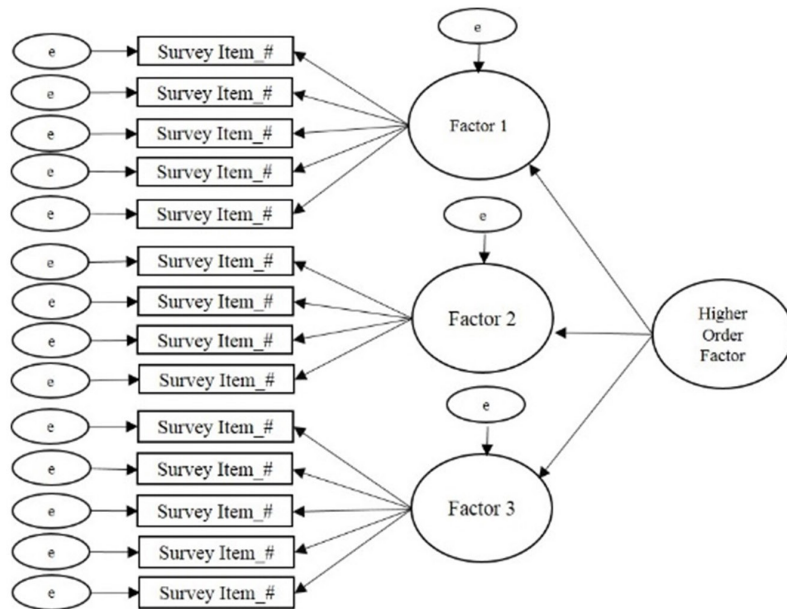
**Figure 3**

*Single-Order, Unidimensional, and Higher-Order Factor Solutions*



Model 1: Single-Order Factor Solution

Model 2: Unidimensional Factor Solution



Model 3: Higher-Order Factor Solution

### **Multiple-Group Confirmatory Factor Analysis**

Multiple-group confirmatory factor analysis (MCFA) is an extension of CFA for testing the factorial invariance (psychometric equivalence) of a scale across subgroups of a sample or population (C.-C. Chen et al., 2020; Dimitrov, 2010). In other words, MCFA has utility for testing the extent to which a particular construct has the same meaning across different groups of a larger sample or population. Suppose, for example, the developer of the Therapeutic Climate scale (see example in the previous section) validated scores on their scale with undergraduate college students. Invariance testing has potential to provide further support for the internal structure validity of the scale by testing whether Empathy, Unconditional Positive Regard, and Congruence have the same meaning across different subgroups of undergraduate college students (e.g., between different gender identities, ethnic identities, age groups, and other subgroups of the larger sample).

**Levels of Invariance.** Factorial invariance can be tested in a number of different ways and includes the following primary levels or aspects: (a) configural invariance, (b) measurement (metric, scalar, and strict) invariance, and (c) structural invariance (Dimitrov, 2010, 2012). Configural invariance (also referred to as pattern invariance) serves as the baseline mode (typically the best fitting model with the data), which is used as the point of comparison when testing for metric, scalar, and structural invariance. In layperson's terms, configural invariance is a test of whether the scales are approximately similar across groups.

Measurement invariance includes testing for metric and scalar invariance. Metric invariance is a test of whether each test item makes an approximately equal contribution (i.e., approximately equal factor loadings) to the latent variable (composite scale score). In layperson's terms, metric invariance evaluates if the scale reasonably captures the same construct. Scalar invariance adds a layer of rigor to metric invariance by testing if the differences between the average scores on the items are attributed to differences in the latent variable means. In layperson's terms, scalar invariance indicates that if the scores change over time, they change in the same way.

Strict invariance is the most stringent level of measurement invariance testing and tests if the sum total of the items' unique variance (item variation that is not in common with the factor) is comparable to the error variance across groups. In layperson's terms, the presence of strict invariance demonstrates that score differences between groups are exclusively due to differences in the common latent variables. Strict invariance, however, is typically not examined in social sciences research because the latent factors are not composed of residuals. Thus, residuals are negligible when evaluating mean differences in latent scores (Putnick & Bornstein, 2016).

Finally, structural invariance is a test of whether the latent factor variances are equivalent to the factor covariances (Dimitrov, 2010, 2012). Structural invariance tests the null hypothesis that there are no statistically significant differences between the unconstrained and constrained models (i.e., determines if the unconstrained model is equivalent to the constrained model). Establishing structural invariance indicates that when the structural pathways are allowed to vary across the two groups, they naturally produce equal results, which supports the notion that the structure of the model is invariant across both groups. In layperson's terms, the presence of structural invariance indicates that the pathways (directionality) between variables behave in the same way across both groups. It is necessary to establish configural and metric invariance prior to testing for structural invariance.

**Sample Size and Criteria for Evaluating Invariance.** Researchers should check their sample size before computing invariance testing, as small samples (approximately < 200) can overestimate model fit

(Dimitrov, 2010). Similar to single-order CFA, no absolute sample size guidelines exist in the literature for invariance testing. Generally, a minimum sample of at least 200 participants per group is recommended for invariance testing (although < 200 to 300+ is advantageous). Referring back to the Therapeutic Climate scale example (see the previous section), investigators would need a minimum sample of 400 if they were seeking to test the invariance of the scale by generational status (200 first generation + 200 non-first generation = 400). The minimum sample size would increase as more levels are added. For example, a minimum sample of 600 would be recommended if investigators quantified generational status on three levels (200 first generation + 200 second generation + 200 third generation and beyond = 600).

Factorial invariance is investigated through a computation of the change in model fit at each level of invariance testing (F. F. Chen, 2007). Historically, the Satorra and Bentler chi-square difference test was the sole criteria for testing factorial invariance, with a non-significant  $p$ -value indicating factorial invariance (Putnick & Bornstein, 2016). The chi-square difference test is still commonly reported by contemporary psychometric researchers; however, it is rarely used as the sole criteria for determining invariance, as the test is sensitive to large samples. The combined recommendations of F. F. Chen (2007) and Putnick and Bornstein (2016) include the following thresholds for investigating invariance:  $\leq \Delta 0.010$  in CFI,  $\leq \Delta 0.015$  in RMSEA, and  $\leq \Delta 0.030$  in SRMR for metric invariance or  $\leq \Delta 0.015$  in SRMR for scalar invariance. In a simulation study, Kang et al. (2016) found that McDonald's NCI (MNCI) outperformed the CFI in terms of stability. Kang et al. (2016) recommend  $< \Delta 0.007$  in MNCI for the 5th percentile and  $\leq \Delta 0.007$  in MNCI for the 1st percentile as cutoff values for measurement quality. Strong measurement invariance is achieved when both metric and scalar invariance are met, and weak invariance is accomplished when only metric invariance is present (Dimitrov, 2010).

## Exemplar Review of a Psychometric Study

The following section will include a review of an exemplar psychometric study based on the recommendations for EFA (see Figure 2) and CFA (see Table 1) that are provided in this manuscript. In 2020, I collaborated with Ryan Flinn on the development and validation of scores on the Mental Distress Response Scale (MDRS) for appraising how college students are likely to respond when encountering a peer in mental distress (Kalkbrenner & Flinn, 2020). A total of 13 items were entered into an EFA. Following the steps for EFA (see Figure 1), the sample size ( $N = 569$ ) exceeded the guidelines for sample size that I published in my 2021 article (Kalkbrenner, 2021b), including an STV of 10:1 or 200 participants, whichever produces a larger sample. Flinn and I (2020) ensured that our 2020 study's data were consistent with a normal distribution (skewness & kurtosis values  $\leq \pm 1$ ) and computed preliminary assumption checking, including inter-item correlation matrix, KMO (.73), and Bartlett's test of sphericity ( $p < .001$ ).

An ML factor extraction method was employed, as the data were largely consistent (skewness & kurtosis values  $\leq \pm 1$ ) with a normal distribution. We used the three most rigorous factor retention criteria—percentage of variance accounted for, scree test, and parallel analysis—to extract a two-factor solution. An oblique factor rotation method (direct oblimin) was employed, as the two factors were correlated. We referred to the recommended factor retention criteria, including  $h^2$  values .30 to .99, factor loadings  $\geq .40$ , and cross-loading  $\geq .30$ , to eliminate one item with low commonalities and two cross-loading items. Using a research team, we named the first factor Diminish/Avoid, as each item that marked this factor reflected a dismissive or evasive response to encountering a peer in mental distress. The second factor was named Approach/Encourage because each item that marked this factor included a response to a peer in mental distress that was active and likely to help connect their peer to mental health support services.



Our next step was to compute a CFA by administering the MDRS to a second sample of undergraduate college students to confirm the two-dimensional factor solution that emerged in the EFA. The sample size ( $N = 247$ ) was sufficient for CFA ( $STV > 10:1$  and  $> 200$  participants). The MDRS items were entered into a CFA and the following GOF indices emerged:  $CMIN = \chi^2 (34) = 61.34$ ,  $p = .003$ ,  $CMIN/DF = 1.80$ ,  $CFI = .96$ ,  $IFI = .96$ ,  $RMSEA = .06$ ,  $90\% CI [0.03, 0.08]$ , and  $SRMR = .04$ . A comparison between our GOF indices from the 2020 study with the thresholds for evaluating model fit in Table 1 reveal an acceptable-to-strong fit between the MDRS model and the data. Collectively, our 2020 procedures for EFA and CFA were consistent with the recommendations in this manuscript.

## Implications for the Profession

### Implications for Counseling Practitioners

Assessment literacy is a vital component of professional counseling practice, as counselors who practice in a variety of specialty areas select and administer tests to clients and use the results to inform diagnosis and treatment planning (C.-C. Chen et al., 2020; Mvududu & Sink, 2013; NBCC, 2016; Neukrug & Fawcett, 2015). It is important to note that test results alone should not be used to make diagnoses, as tests are not inherently valid (Kalkbrenner, 2021b). In fact, the authors of the *Diagnostic and Statistical Manual of Mental Disorders* stated that “scores from standardized measures and interview sources must be interpreted using clinical judgment” (American Psychiatric Association, 2013, p. 37). Professional counselors can use test results to inform their diagnoses; however, diagnostic decision making should ultimately come down to a counselor’s clinical judgment.

Counseling practitioners can refer to this manuscript as a reference for evaluating the internal structure validity of scores on a test to help determine the extent to which, if any at all, the test in question is appropriate for use with clients. When evaluating the rigor of an EFA for example, professional counselors can refer to this manuscript to evaluate the extent to which test developers followed the appropriate procedures (e.g., preliminary assumption checking, factor extraction, retention, and rotation [see Figure 2]). Professional counselors are encouraged to pay particular attention to the factor extraction method that the test developers employed, as PCA is sometimes used in lieu of more appropriate methods (e.g., PAF/ML). Relatedly, professional counselors should be vigilant when evaluating the factor rotation method employed by test developers because oblique rotation methods are typically more appropriate than orthogonal (e.g., varimax) for counseling tests.

CFA is one of the most commonly used tests of the internal structure validity of scores on psychological assessments (Kalkbrenner, 2021b). Professional counselors can compare the CFA fit indices in a test manual or journal article to the benchmarks in Table 1 and come to their own conclusion about the internal structure validity of scores on a test before using it with clients. Relatedly, the layperson’s definitions of common psychometric terms in Figure 1 might have utility for increasing professional counselors’ assessment literacy by helping them decipher some of the psychometric jargon that commonly appears in psychometric studies and test manuals.

### Implications for Counselor Education

Assessment literacy begins in one’s counselor education program and it is imperative that counselor educators teach their students to be proficient in recognizing and evaluating internal structure validity evidence of test scores. Teaching internal structure validity evidence can be an especially challenging pursuit because counseling students tend to fear learning about psychometrics and statistics (Castillo, 2020; Steele & Rawls, 2015), which can contribute to their reticence and uncertainty when encountering

psychometric research. This reticence can lead one to read the methodology section of a psychometric study briefly, if at all. Counselor educators might suggest the present article as a resource for students taking classes in research methods and assessment as well as for students who are completing their practicum, internship, or dissertation who are evaluating the rigor of existing measures for use with clients or research participants.

Counselor educators should urge their students not to skip over the methodology section of a psychometric study. When selecting instrumentation for use with clients or research participants, counseling students and professionals should begin by reviewing the methodology sections of journal articles and test manuals to ensure that test developers employed rigorous and empirically supported procedures for test development and score validation. Professional counselors and their students can compare the empirical steps and guidelines for structural validation of scores that are presented in this manuscript with the information in test manuals and journal articles of existing instrumentation to evaluate its internal structure. Counselor educators who teach classes in assessment or psychometrics might integrate an instrument evaluation assignment into the course in which students select a psychological instrument and critique its psychometric properties. Another way that counselor educators who teach classes in current issues, research methods, assessment, or ethics can facilitate their students' assessment literacy development is by creating an assignment that requires students to interview a psychometric researcher. Students can find psychometric researchers by reviewing the editorial board members and authors of articles published in the two peer-reviewed journals of the Association for Assessment and Research in Counseling, *Measurement and Evaluation in Counseling and Development* and *Counseling Outcome Research and Evaluation*. Students might increase their interest and understanding about the necessity of assessment literacy by talking to researchers who are passionate about psychometrics.

## **Assessment Literacy: Additional Considerations**

Internal structure validity of scores is a crucial component of assessment literacy for evaluating the construct validity of test scores (Bandalos & Finney, 2019). Assessment literacy, however, is a vast construct and professional counselors should consider a number of additional aspects of test worthiness when evaluating the potential utility of instrumentation for use with clients. Reviewing these additional considerations is beyond the scope of this manuscript; however, readers can refer to the following features of assessment literacy and corresponding resources: reliability (Kalkbrenner, 2021a), practicality (Neukrug & Fawcett, 2015), steps in the instrument development process (Kalkbrenner, 2021b), and convergent and divergent validity evidence of scores (Swank & Mullen, 2017). Moreover, the discussion of internal structure validity evidence of scores in this manuscript is based on Classical Test Theory (CTT), which tends to be an appropriate platform for attitudinal measures. However, Item Response Theory (see Amarnani, 2009) is an alternative to CTT with particular utility for achievement and aptitude testing.

## **Cross-Cultural Considerations in Assessment Literacy**

Professional counselors have an ethical obligation to consider the cross-cultural fairness of a test before use with clients, as the validity of test scores are culturally dependent (American Counseling Association [ACA], 2014; Kane, 2010; Neukrug & Fawcett, 2015; Swanepoel & Kruger, 2011). Cross-cultural fairness (also known as test fairness) in testing and assessment "refers to the comparability of score meanings across individuals, groups or settings" (Swanepoel & Kruger, 2011, p. 10). There exists some overlap between internal structure validity and cross-cultural fairness; however, some distinct differences exist as well.

Using CFA to confirm the factor structure of an established test with participants from a different culture is one way to investigate the cross-cultural fairness of scores. Suppose, for example, an investigator found acceptable internal structure validity evidence (see Table 1) for scores on an anxiety inventory that was normed in America with participants in Eastern Europe who identify with a collectivist cultural background. Such findings would suggest that the dimensionality of the anxiety inventory extends to the sample of Eastern European participants. However, internal structure validity testing alone might not be sufficient for testing the cross-cultural fairness of scores, as factor analysis does not test for content validity. In other words, although the CFA confirmed the dimensionality of an American model with a sample of Eastern European participants, the analysis did not take potential qualitative differences about the construct of measurement (anxiety severity) into account. It is possible (and perhaps likely) that the lived experience of anxiety differs between those living in two different cultures. Accordingly, a systems-level approach to test development and score validation can have utility for enhancing the cross-cultural fairness of scores (Swanepoel & Kruger, 2011).

### ***A Systems-Level Approach to Test Development and Score Validation***

Swanepoel and Kruger (2011) outlined a systemic approach to test development that involves circularity, which includes incorporating qualitative inquiry into the test development process, as qualitative inquiry has utility for uncovering the nuances of participants' lived experiences that quantitative data fail to capture. For example, an exploratory-sequential mixed-methods design in which qualitative findings are used to guide the quantitative analyses is a particularly good fit with systemic approaches to test development and score validation. Referring to the example in the previous section, test developers might conduct qualitative interviews to develop a grounded theory of anxiety severity in the context of the collectivist culture. The grounded theory findings could then be used as the theoretical framework (see Kalkbrenner, 2021b) for a psychometric study aimed at testing the generalizability of the qualitative findings. Thus, in addition to evaluating the rigor of factor analytic results, professional counselors should also review the cultural context in which test items were developed before administering a test to clients.

Language adaptations of instrumentation are another relevant cross-cultural fairness consideration in counseling research and practice. Word-for-word translations alone are insufficient for capturing cross-cultural fairness of instrumentation, as culture extends beyond just language (Lenz et al., 2017; Swanepoel & Kruger, 2011). Pure word-for-word translations can also cause semantic errors. For example, feeling "fed up" might translate to feeling angry in one language and to feeling full after a meal in another language. Accordingly, professional counselors should ensure that a translated instrument was subjected to rigorous procedures for maintaining cross-cultural fairness. Reviewing such procedures is beyond the scope of this manuscript; however, Lenz et al. (2017) outlined a 6-step process for language translation and cross-cultural adaptation of instruments.

## **Conclusion**

Gaining a deeper understanding of the major approaches to factor analysis for demonstrating internal structure validity in counseling research has potential to increase assessment literacy among professional counselors who work in a variety of specialty areas. It should be noted that the thresholds for interpreting the strength of internal structure validity coefficients that are provided throughout this manuscript should be used as tentative guidelines, not unconditional standards. Ultimately, internal structure validity is a function of test scores and the construct of measurement. The stakes or consequences of test results should be considered when making final decisions about the strength of validity coefficients. As professional counselors increase their familiarity with factor

analysis, they will most likely become more cognizant of the strengths and limitations of counseling-related tests to determine their utility for use with clients. The practical overview of factor analysis presented in this manuscript can serve as a one-stop shop or resource that professional counselors can refer to as a reference for selecting tests with validated scores for use with clients, a primer for teaching courses, and a resource for conducting their own research.

### ***Conflict of Interest and Funding Disclosure***

The author reported no conflict of interest or funding contributions for the development of this manuscript.

---

## **References**

- Amarnani, R. (2009). Two theories, one theta: A gentle introduction to item response theory as an alternative to classical test theory. *The International Journal of Educational and Psychological Assessment*, 3, 104–109.
- American Counseling Association. (2014). *ACA code of ethics*. <https://www.counseling.org/resources/aca-code-of-ethics.pdf>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. <https://www.apa.org/science/programs/testing/standards>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Bandalos, D. L., & Finney, S. J. (2019). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd ed., pp. 98–122). Routledge.
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation*, 18(5/6), 1–13. <https://doi.org/10.7275/qv2q-rk76>
- Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.
- Castillo, J. H. (2020). Teaching counseling students the science of research. In M. O. Adekson (Ed.), *Beginning your counseling career: Graduate preparation and beyond* (pp. 122–130). Routledge.
- Chen, C.-C., Lau, J. M., Richardson, G. B., & Dai, C.-L. (2020). Measurement invariance testing in counseling. *Journal of Professional Counseling: Practice, Theory & Research*, 47(2), 89–104. <https://doi.org/10.1080/15566382.2020.1795806>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Council for Accreditation of Counseling and Related Educational Programs. (2015). *2016 CACREP standards*. <http://www.cacrep.org/wp-content/uploads/2017/08/2016-Standards-with-citations.pdf>
- Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: A critical review and development of reporting recommendations. *Journal of Organizational Behavior*, 36(6), 845–872. <https://doi.org/10.1002/job.2008>
- Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43(2), 121–149. <https://doi.org/10.1177/0748175610373459>
- Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. American Counseling Association.



- 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>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, 6(1), 53–60.
- Hu, L., & 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>
- Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychologist*, 34(5), 684–718. <https://doi.org/10.1177/0011000006286347>
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. <https://doi.org/10.1007/BF02291575>
- Kalkbrenner, M. T. (2021a). Alpha, omega, and *H* internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*. Advance online publication. <https://doi.org/10.1080/21501378.2021.1940118>
- Kalkbrenner, M. T. (2021b). A practical guide to instrument development and score validation in the social sciences: The MEASURE Approach. *Practical Assessment, Research, and Evaluation*, 26, Article 1. <https://scholarworks.umass.edu/pare/vol26/iss1/1>
- Kalkbrenner, M. T., & Flinn, R. E. (2020). The Mental Distress Response Scale and promoting peer-to-peer mental health support: Implications for college counselors and student affairs officials. *Journal of College Student Development*, 61(2), 246–251. <https://doi.org/10.1353/csd.2020.0021>
- Kane, M. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182. <https://doi.org/10.1177/0265532209349467>
- Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, 76(4), 533–561. <https://doi.org/10.1177/0013164415603764>
- Lenz, A. S., Gómez Soler, I., Dell’Aquila, J., & Uribe, P. M. (2017). Translation and cross-cultural adaptation of assessments for use in counseling research. *Measurement and Evaluation in Counseling and Development*, 50(4), 224–231. <https://doi.org/10.1080/07481756.2017.1320947>
- Lewis, T. F. (2017). Evidence regarding the internal structure: Confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, 50(4), 239–247. <https://doi.org/10.1080/07481756.2017.1336929>
- Mvududu, N. H., & Sink, C. A. (2013). Factor analysis in counseling research and practice. *Counseling Outcome Research and Evaluation*, 4(2), 75–98. <https://doi.org/10.1177/2150137813494766>
- National Board for Certified Counselors. (2016). *NBCC code of ethics*. <https://www.nbcc.org/Assets/Ethics/NBCCCodeofEthics.pdf>
- Neukrug, E. S., & Fawcett, R. C. (2015). *Essentials of testing and assessment: A practical guide for counselors, social workers, and psychologists* (3rd ed.). Cengage.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Steele, J. M., & Rawls, G. J. (2015). Quantitative research attitudes and research training perceptions among master’s-level students. *Counselor Education and Supervision*, 54(2), 134–146. <https://doi.org/10.1002/ceas.12010>
- Swanepoel, I., & Kruger, C. (2011). Revisiting validity in cross-cultural psychometric-test development: A systems-informed shift towards qualitative research designs. *South African Journal of Psychiatry*, 17(1), 10–15. <https://doi.org/10.4102/sajpsychiatry.v17i1.250>

- Swank, J. M., & Mullen, P. R. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and Evaluation in Counseling and Development*, 50(4), 270–274.  
<https://doi.org/10.1080/07481756.2017.1339562>
- Tate, K. A., Bloom, M. L., Tassara, M. H., & Caperton, W. (2014). Counselor competence, performance assessment, and program evaluation: Using psychometric instruments. *Measurement and Evaluation in Counseling and Development*, 47(4), 291–306. <https://doi.org/10.1177/0748175614538063>
- Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development*, 50(4), 232–238.  
<https://doi.org/10.1080/07481756.2017.1336931>
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838.  
<https://doi.org/10.1177/0011000006288127>

